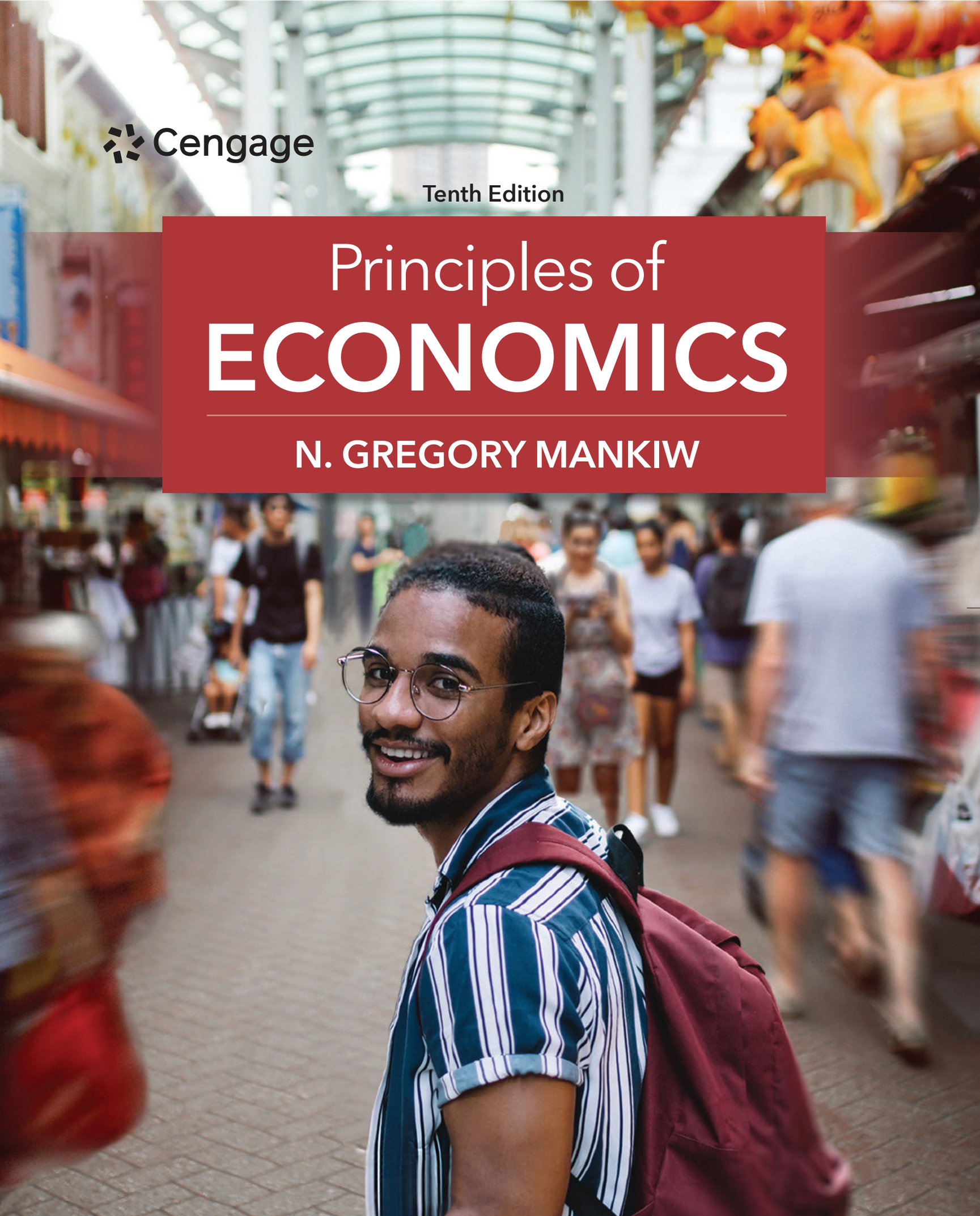


Cengage

Tenth Edition

Principles of ECONOMICS

N. GREGORY MANKIW

Principles of Economics: a Guided Tour

Introduction

1 Ten Principles of Economics — The study of economics is guided by a few big ideas.

2 Thinking Like an Economist — Economists view the world as both scientists and policymakers.

3 Interdependence and the Gains from Trade — The theory of comparative advantage explains how people benefit from economic interdependence.

How Markets Work

4 The Market Forces of Supply and Demand

5 Elasticity and Its Application

How does the economy coordinate interdependent economic actors? Through the market forces of supply and demand.

6 Supply, Demand, and Government Policies — The tools of supply and demand are put to work to examine the effects of various government policies.

Markets and Welfare

7 Consumers, Producers, and the Efficiency of Markets

8 Application: The Costs of Taxation

9 Application: International Trade

Why is the equilibrium of supply and demand desirable for society as a whole? The concepts of consumer and producer surplus explain the efficiency of markets, the costs of taxation, and the benefits of international trade.

The Economics of the Public Sector

10 Externalities

11 Public Goods and Common Resources

12 The Economics of Healthcare

Market outcomes are not always efficient, and governments can sometimes remedy market failure.

13 The Design of the Tax System — To fund programs, governments raise revenue through their tax systems, which are designed with an eye toward balancing efficiency and equity.

Firm Behavior and the Organization of Industry

14 The Costs of Production

15 Firms in Competitive Markets

The theory of the firm sheds light on the decisions that lie behind supply in competitive markets.

16 Monopoly

17 Monopolistic Competition

18 Oligopoly

Firms with market power can cause market outcomes to be inefficient.

The Economics of Labor Markets

19 The Markets for the Factors of Production
20 Earnings and Discrimination
21 Income Inequality and Poverty

These chapters examine the special features of labor markets, in which most people earn most of their income.

Topics for Further Study

22 The Theory of Consumer Choice
23 Frontiers of Microeconomics

Additional topics in microeconomics include household decision making, asymmetric information, political economy, and behavioral economics.

The Data of Macroeconomics

24 Measuring a Nation's Income
25 Measuring the Cost of Living

The overall quantity of production and the overall price level are used to monitor developments in the economy as a whole.

The Real Economy in the Long Run

26 Production and Growth
27 Saving, Investment, and the Financial System
28 The Basic Tools of Finance
29 Unemployment

These chapters describe the forces that in the long run determine key real variables, including GDP growth, saving, investment, real interest rates, and unemployment.

Money and Prices in the Long Run

30 The Monetary System
31 Money Growth and Inflation

The monetary system is crucial in determining the long-run behavior of the price level, the inflation rate, and other nominal variables.

The Macroeconomics of Open Economies

32 Open-Economy Macroeconomics: Basic Concepts

A nation's economic interactions with other nations are described by its trade balance, net foreign investment, and exchange rate.

33 A Macroeconomic Theory of the Open Economy

A long-run model of the open economy explains the determinants of the trade balance, the real exchange rate, and other real variables.

Short-Run Economic Fluctuations

34 Aggregate Demand and Aggregate Supply
35 The Influence of Monetary and Fiscal Policy on Aggregate Demand
36 The Short-Run Trade-off between Inflation and Unemployment

The model of aggregate demand and aggregate supply explains short-run economic fluctuations, the short-run effects of monetary and fiscal policy, and the short-run linkage between real and nominal variables.

Final Thoughts

37 Six Debates over Macroeconomic Policy

A capstone chapter presents both sides of six major debates over economic policy.

38 Appendix: How Economists Use Data

The analysis of data to test theories and estimate parameters is central to the science of economics.

Suggestions for Summer Reading

If you enjoyed the economics course that you just finished, you might like to read more about economic issues in the following books.

Abhijit V. Banerjee and Esther Duflo

Good Economics for Hard Times

(New York: PublicAffairs, 2019)

Two prominent economists—winners of the Nobel prize in 2019—offer their ideas about how to build a better world.

Yoram Bauman and Grady Klein

The Cartoon Introduction to Economics

(New York: Hill and Wang, 2010)

Basic economic principles, with humor.

Bryan Caplan

The Myth of the Rational Voter: Why Democracies Choose Bad Policies

(Princeton, NJ: Princeton University Press, 2008)

An economist asks why elected leaders often fail to follow the policies that economists recommend.

Kimberly Clausing

Open: The Progressive Case for Free Trade, Immigration, and Global Capital

(Cambridge, MA: Harvard University Press, 2019)

An economist explains why Americans benefit from interacting with the rest of the world.

Avinash K. Dixit and Barry J. Nalebuff

The Art of Strategy: A Game Theorist's Guide to Success in Business and Life

(New York: Norton, 2008)

This introduction to game theory discusses how all people—from arrested criminals to corporate executives—should, and do, make strategic decisions.

Mihir Desai

The Wisdom of Finance: Discovering Humanity in the World of Risk and Return

(Boston: Houghton Mifflin Harcourt, 2017)

A charming look at how the insights of finance inform our lives.

William Easterly

The Tyranny of Experts: Economists, Dictators, and the Forgotten Rights of the Poor

(New York: Basic Books, 2013)

A former World Bank economist examines the many attempts to help the world's poorest nations and why these attempts have often failed.

Milton Friedman

Capitalism and Freedom

(Chicago: University of Chicago Press, 1962)

One of the most important economists of the 20th century argues that society should rely less on the government and more on the free market.

Robert L. Heilbroner

The Worldly Philosophers

(New York: Touchstone, 1953, revised 1999)

A classic introduction to the lives, times, and ideas of the great economic thinkers, including Adam Smith, David Ricardo, and John Maynard Keynes.

Steven E. Landsburg

The Armchair Economist: Economics and Everyday Life

(New York: Free Press, 2012)

Why does popcorn cost so much at movie theaters? Steven Landsburg discusses this and other puzzles of economic life.




Steven D. Levitt and Stephen J. Dubner

Freakonomics: A Rogue Economist Explores the Hidden Side of Everything

(New York: Morrow, 2005)

Economic principles and clever data analysis applied to a wide range of offbeat topics, including drug dealing, online dating, and sumo wrestling.

Roger Lowenstein

America's Bank: The Epic Struggle to Create the Federal Reserve

(New York: Penguin Press, 2015)

A history of the founding of one of the most important policymaking institutions in the United States.

Annie Lowrey

Give People Money: How a Universal Basic Income Would End Poverty, Revolutionize Work, and Remake the World

(New York: Crown, 2018)

The case for a substantial rethinking of the social safety net.

Burton G. Malkiel

A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing

(New York: Norton, 2019)

This introduction to stocks, bonds, and financial economics is not a "get rich quick" book, but it might help you get rich slowly.

Deirdre McCloskey and Art Carden

Leave Me Alone and I'll Make You Rich: How the Bourgeois Deal Enriched the World

(Chicago: University of Chicago Press, 2020)

An overview of economic history that asks why most modern societies have, over the past two centuries, escaped the grinding poverty that previously characterized most of human existence.

John McMillan

Reinventing the Bazaar: A Natural History of Markets

(New York: Norton, 2002)

A deep and nuanced, yet still very readable, analysis of how society can make the best use of market mechanisms.

Branko Milanovic

Capitalism, Alone: The Future of the System that Rules the World

(Cambridge, MA: Harvard University Press, 2019)

A look at how capitalism manifests itself in different ways in different countries.

Sendhil Mullainathan and Eldar Shafir

Scarcity: Why Having Too Little Means So Much

(New York: Times Books, 2013)

An economist and psychologist team up to examine the causes and consequences of our limited cognitive abilities.

Sylvia Nasar

Grand Pursuit: The Story of Economic Genius

(New York: Simon and Schuster, 2011)

A sweeping narrative that tells the story of economic discovery.

William D. Nordhaus

The Spirit of Green: The Economics of Collisions and Contagions in a Crowded World

(Princeton, NJ: Princeton University Press, 2021)

The 2018 Nobel laureate in economics examines how to best address critical externalities, such as the carbon emissions that lead to global climate change.

Roger W. Spencer and David A. Macpherson

Lives of the Laureates

(Cambridge, MA: MIT Press, 2014)

Twenty-three winners of the Nobel Prize in Economics offer autobiographical essays about their lives and work.

Tenth Edition

Principles of ECONOMICS

N. GREGORY MANKIW
HARVARD UNIVERSITY

Cengage

Australia • Brazil • Mexico • Singapore • United Kingdom • United States

Cengage

***Principles of Economics,* 10e**
N. Gregory Mankiw

SVP, Higher Education Product Management: Erin Joyner

VP, Product Management, Learning Experiences: Thais Alencar

Product Director: Joe Sabatino

Senior Product Manager: Chris Rader

Product Assistant: Autumn Lala-Sonora

Learning Designers: Natasha Allen, Sarah Keeling

Senior Content Managers: Colleen A. Farmer, Allison K. Janneck, Anita Verma

Digital Project Managers: Judy Kwan, Dan Swanson

In-House Subject Matter Experts: Shannon Aucoin, Eugenia Belova, Alex Lewis

Executive Product Marketing Manager: John Carey

IP Analyst: Ashley Maynard

IP Project Manager: Anjali Kambli

Production Service: MPS Limited

Designer: Erin Griffin

Cover Image Sources: lechatnoir/Getty Images, GaudiLab/Shutterstock.com

Interior Image Sources: Maremagnum/Getty Images, lechatnoir/Getty Images, GaudiLab/Shutterstock.com, Roman Samborskyi/Shutterstock.com, Arno Van Engeland/EyeEm/Getty Images, blackCAT/Getty Images, gowithstock/Shutterstock.com, Shift Drive/Shutterstock.com, Erdark/Getty Images, AzmanJaka/Getty Images, Carlos Flores/EyeEm/Getty Images

Library of Congress Control Number: 2022908687

ISBN: 978-0-357-72271-8

Cengage
200 Pier 4 Boulevard
Boston, MA 02210
USA

Printed in the United States of America
Print Number: 01 Print Year: 2023

To Catherine, Nicholas, and Peter,
my other contributions to the next generation

About the Author

JORDI CABRÉ

N. Gregory Mankiw is the Robert M. Beren Professor of Economics at Harvard University. As a student, he studied economics at Princeton University and MIT. As a teacher, he has taught macroeconomics, microeconomics, statistics, and principles of economics. He even spent one summer long ago as a sailing instructor on Long Beach Island.

Professor Mankiw is a prolific writer and regular participant in academic and policy debates. His work has been published in scholarly journals such as the *American Economic Review, Journal of Political Economy,* and *Quarterly Journal of Economics* and in more popular forums, such as the *New York Times* and *The Wall Street Journal*. He is also the author of the best-selling intermediate-level textbook *Macroeconomics* (Worth Publishers).

In addition to his teaching, research, and writing, Professor Mankiw has been a research associate of the National Bureau of Economic Research, a member of the Brookings Panel on Economic Activity, an adviser to the Congressional Budget Office and the Federal Reserve Banks of Boston and New York, a trustee of the Urban Institute and the Economic Club of New York, and a member of the ETS test development committee for the Advanced Placement exam in economics. From 2003 to 2005, he served as chairman of the President's Council of Economic Advisers.

Preface: To the Instructor

During my 20-year career as a student, the course that excited me most was the two-semester sequence on the principles of economics that I took during my first year in college. It is no exaggeration to say that it changed my life.

I had grown up in a family that often discussed politics over the dinner table. The pros and cons of various solutions to society's problems generated fervent debate. But in school, I had been drawn to the sciences. While politics seemed vague, rambling, and subjective, science was analytic, systematic, and objective. Political debate continued without end, but scientific research made progress.

My freshman course on the principles of economics opened my eyes to a new way of thinking. Economics combines the virtues of politics and science. It is, truly, a social science. Its subject matter is society—how people choose to lead their lives and how they interact with one another—but it approaches the subject with the dispassion of a science. By bringing the methods of science to the questions of politics, economics aims to make progress on the challenges that all societies face.

I wrote this book with the hope that I could convey some of the excitement about economics that I felt as a student in my first economics course. Economics is a subject in which a little knowledge goes a long way. (The same cannot be said, for instance, of the study of physics or the Chinese language.) Economists have a unique worldview, much of which can be taught in one or two semesters. My goal in this book is to transmit this way of thinking to the widest possible audience and to convince readers that it illuminates much about their lives and the world around them.

I believe that everyone should study the fundamental ideas that economics has to offer. One purpose of general education is to teach people about the world and thereby make them better citizens. The study of economics, as much as any discipline, serves this goal. Writing an economics textbook is, therefore, a great honor and a great responsibility. It is one way that economists can help promote better government and a more prosperous future. As the great economist Paul Samuelson put it, "I don't care who writes a nation's laws, or crafts its advanced treaties, if I can write its economics textbooks."

What's New in the Tenth Edition?

Economics aims to understand the world in which we live. Most chapters of this book include Case Studies that illustrate how the principles of economics can be applied. In the News boxes offer excerpts from newspapers, magazines, and online news sources to show how economic ideas shed light on current issues facing society. After students finish their first course in economics, they should think about news reports from a new perspective and with greater insight. To keep the study of economics fresh and relevant for each new cohort of students, I update each edition to keep pace with the ever-changing world.

The new applications in this tenth edition are too numerous to list in their entirety, but here is a sample of the topics covered (and the chapters in which they appear):

- Shortages during the coronavirus pandemic renewed the debate over whether it is fair for businesses to increase prices during a crisis. (Chapter 4)
- The future of the ride-share market hinges on the elasticities of supply and demand. (Chapter 5)
- The minimum wage remains a contentious topic. (Chapter 6)
- A carbon tax is a versatile tool to combat global climate change. (Chapter 10)
- Putting a price on road use gets renewed attention as the United States embarks on building new infrastructure. (Chapter 11)
- The pandemic of 2020 taught some lessons about why it's hard to cut wasteful medical spending. (Chapter 12)
- The value-added tax might be a policy for the United States to consider. (Chapter 13)
- The Biden administration looked to expand the scope of antitrust policy. (Chapter 16)
- Amazon found itself in the crosshairs of antitrust enforcers. (Chapter 18)
- Immigration policy creates winners and losers in the labor market. (Chapter 19)
- The forgone schooling during the coronavirus pandemic might have long-lasting effects on earnings. (Chapter 20)
- New research takes a lifetime perspective on measuring inequality. (Chapter 21)
- Robust expansions of the social safety net reduced poverty during the coronavirus pandemic. (Chapter 21)
- People are not good at rationally responding to small-probability events. (Chapter 23)
- Research has shed light on how the aftermath of the slave trade affects modern Africa. (Chapter 26)
- The four-decade decline in real interest rates is puzzling. (Chapter 27)
- Women are generally better investors than men. (Chapter 28)
- New research has examined the use of efficiency wages. (Chapter 29)
- The recession caused by the coronavirus pandemic was unusual in several ways. (Chapter 34)
- The monetary and fiscal response during the pandemic caused some economists to worry about a resurgence in inflation. (Chapter 36)
- The Federal Reserve is asked to expand its set of economic goals. (Chapter 37)

This edition also includes two new chapters. Chapter 12 examines the economics of healthcare. As this sector's share of the economy has increased, its distinctive features, problems, and policy challenges have become more important for students to understand. Chapter 38 is an optional appendix chapter that discusses how economists use data. In recent years, economic research has grown increasingly empirical, and some instructors want to introduce students to the statistical methods that economists use. Instructors who teach this chapter can move it earlier in the course.

As always, I have carefully gone through every chapter to refine the book's coverage and pedagogy. There are numerous changes, large and small, to ensure that the book is clear, accurate, and up-to-date.

All the changes that I made, and the many others that I considered, were evaluated in light of the benefits of brevity. Like most things studied in economics, a

student's time is a scarce resource. I always keep in mind a dictum from the novelist Robertson Davies: "One of the most important things about writing is to boil it down and not bore the hell out of everybody."

How Is This Book Organized?

This book is organized to make economics as student-friendly as possible. What follows is a whirlwind tour, which will, I hope, give instructors some sense of how the pieces fit together.

Introductory Material

Chapter 1, "Ten Principles of Economics," introduces students to the economist's view of the world. It previews the big ideas that recur in economics, such as opportunity cost, marginal decision making, the role of incentives, the gains from trade, and the efficiency of market allocations. Throughout the book, I refer regularly to the **Ten Principles of Economics** in Chapter 1 to remind students that these ideas are the foundation for all economics.

Chapter 2, "Thinking Like an Economist," examines how economists approach their subject. It discusses the role of assumptions in developing a theory and introduces the concept of an economic model. It also explores the role of economists in making policy. This chapter's appendix offers a brief refresher course on how graphs are used as well as how they can be abused.

Chapter 3, "Interdependence and the Gains from Trade," presents the theory of comparative advantage. This theory explains why individuals trade with their neighbors and why nations trade with other nations. Much of economics is about how market forces coordinate many individual production and consumption decisions. As a starting point for this analysis, students see in this chapter why specialization, interdependence, and trade can benefit everyone.

The Fundamental Tools of Supply and Demand

The next three chapters introduce the basic tools of supply and demand. Chapter 4, "The Market Forces of Supply and Demand," develops the supply curve, the demand curve, and the notion of market equilibrium. Chapter 5, "Elasticity and Its Application," introduces the concept of elasticity and uses it to analyze events in three different markets. Chapter 6, "Supply, Demand, and Government Policies," uses these tools to examine price controls, such as rent-control and minimum-wage laws, and tax incidence.

Chapter 7, "Consumers, Producers, and the Efficiency of Markets," extends the analysis of supply and demand using the concepts of consumer surplus and producer surplus. It begins by developing the link between consumers' willingness to pay and the demand curve and the link between producers' costs of production and the supply curve. It then shows that the market equilibrium maximizes the sum of the producer and consumer surplus. Thus, students learn early about the efficiency of market allocations.

The next two chapters apply the concepts of producer and consumer surplus to policy questions. Chapter 8, "Application: The Costs of Taxation," shows why taxation results in deadweight losses and what determines the size of those losses. Chapter 9, "Application: International Trade," considers who wins and who loses from international trade and presents the debate over protectionist trade policies.

More Microeconomics

Having examined why market allocations are often desirable, the book then considers how the government can sometimes improve on them. Chapter 10, "Externalities," explains how external effects such as pollution can render market outcomes inefficient and discusses the possible public and private solutions to those inefficiencies. Chapter 11, "Public Goods and Common Resources," considers the problems that arise when goods, such as national defense, have no market price. Chapter 12, "The Economics of Healthcare," examines the distinctive features, problems, and policy challenges of an increasingly important sector of the economy. Chapter 13, "The Design of the Tax System," describes how the government raises the revenue necessary to pay for public goods. It presents some institutional background about the U.S. tax system and then discusses how the goals of efficiency and equity come into play when designing a tax system.

The next five chapters examine firm behavior and industrial organization. Chapter 14, "The Costs of Production," discusses what to include in a firm's costs, and it introduces cost curves. Chapter 15, "Firms in Competitive Markets," analyzes the behavior of price-taking firms and derives the market supply curve. Chapter 16, "Monopoly," discusses the behavior of a firm that is the sole seller in its market. It examines the inefficiency of monopoly pricing, the possible policy responses, and the attempts by monopolies to price discriminate. Chapter 17, "Monopolistic Competition," looks at behavior in a market in which many sellers offer similar but differentiated products. It also discusses the debate over the effects of advertising. Chapter 18, "Oligopoly," covers markets in which there are only a few sellers, using the prisoners' dilemma as the model for examining strategic interaction.

The next three chapters present issues related to labor markets. Chapter 19, "The Markets for the Factors of Production," emphasizes the link between factor prices and marginal productivity. Chapter 20, "Earnings and Discrimination," discusses the determinants of equilibrium wages, including compensating differentials, human capital, and discrimination. Chapter 21, "Income Inequality and Poverty," examines the degree of inequality in U.S. society, alternative views about the government's role in changing the distribution of income, and various policies aimed at helping members of society experiencing poverty.

The next two chapters present optional material. Chapter 22, "The Theory of Consumer Choice," analyzes individual decision making using budget constraints and indifference curves. Chapter 23, "Frontiers of Microeconomics," introduces the topics of asymmetric information, political economy, and behavioral economics. Some instructors may skip all or some of this material, but these chapters are useful in motivating and preparing students for future courses in microeconomics. Instructors who cover these topics may assign these chapters earlier than they are presented in the book, and I have written them to facilitate this flexibility.

Macroeconomics

My overall approach to teaching macroeconomics is to examine the economy in the long run (when prices are flexible) before examining the economy in the short run (when prices are sticky). I believe that this organization simplifies learning macroeconomics for several reasons. First, the classical assumption of price flexibility is more closely linked to the basic lessons of supply and demand, which students have already mastered. Second, the classical dichotomy allows the study of the long run to be broken up into several easily digested pieces. Third, because the business cycle represents a

transitory deviation from the economy's long-run growth path, studying the transitory deviations is more natural after the long-run equilibrium is understood. Fourth, the macroeconomic theory of the long run is less controversial among economists than is the macroeconomic theory of the short run. For these reasons, most upper-level courses in macroeconomics now follow this long-run-before-short-run approach; my goal is to offer introductory students the same advantage.

I start the coverage of macroeconomics with issues of measurement. Chapter 24, "Measuring a Nation's Income," discusses the meaning of gross domestic product and related statistics from the national income accounts. Chapter 25, "Measuring the Cost of Living," examines the measurement and use of the consumer price index.

The next four chapters describe the behavior of the real economy in the long run. Chapter 26, "Production and Growth," examines the determinants of the large variation in living standards over time and across countries. Chapter 27, "Saving, Investment, and the Financial System," discusses the types of financial institutions in our economy and examines their role in allocating resources. Chapter 28, "The Basic Tools of Finance," introduces present value, risk management, and asset pricing. Chapter 29, "Unemployment," considers the long-run determinants of the unemployment rate, including job search, minimum-wage laws, the market power of unions, and efficiency wages.

Having described the long-run behavior of the real economy, the book then turns to the long-run behavior of money and prices. Chapter 30, "The Monetary System," introduces the economist's concept of money and the role of the central bank in controlling the quantity of money. Chapter 31, "Money Growth and Inflation," develops the classical theory of inflation and discusses the costs that inflation imposes on a society.

The next two chapters present the macroeconomics of open economies, maintaining the long-run assumptions of price flexibility and full employment. Chapter 32, "Open-Economy Macroeconomics: Basic Concepts," explains the relationship among saving, investment, and the trade balance, the distinction between the nominal and real exchange rate, and the theory of purchasing-power parity. Chapter 33, "A Macroeconomic Theory of the Open Economy," presents a classical model of the international flow of goods and capital. The model sheds light on various issues, including the link between budget deficits and trade deficits and the macroeconomic effects of trade policies. Because instructors differ in their emphasis on this material, these chapters are written so they can be used in different ways. Some may choose to cover Chapter 32 but not Chapter 33, others may skip both chapters, and still others may choose to defer the analysis of open-economy macroeconomics until the end of their courses.

After developing the long-run theory of the economy in Chapters 26 through 33, the book turns to explaining short-run fluctuations around the long-run trend. Chapter 34, "Aggregate Demand and Aggregate Supply," begins with some facts about the business cycle and then introduces the model of aggregate demand and aggregate supply. Chapter 35, "The Influence of Monetary and Fiscal Policy on Aggregate Demand," explains how policymakers can use the tools at their disposal to shift the aggregate-demand curve. Chapter 36, "The Short-Run Trade-Off between Inflation and Unemployment," explains why policymakers who control aggregate demand face a trade-off between inflation and unemployment. It examines why this trade-off exists in the short run, why it shifts over time, and why it does not exist in the long run.

The discussion of macroeconomics concludes with Chapter 37, "Six Debates over Macroeconomic Policy." This capstone chapter considers six controversial issues facing policymakers: the proper degree of policy activism in response to the

business cycle, the relative efficacy of government spending hikes and tax cuts to fight recessions, the choice between rules and discretion in the conduct of monetary policy, the desirability of reaching zero inflation, the importance of balancing the government's budget, and the need for tax reform to encourage saving. For each issue, the chapter presents both sides of the debate and encourages students to make their own judgments.

The final chapter in the book is Chapter 38, "Appendix: How Economists Use Data." This chapter introduces students to the statistical methods that economists use to test and apply their theories. Instructors who teach it may choose to move it earlier in the course.

Learning Tools

The purpose of this book is to help students learn the fundamental lessons of economics and to show how they can apply these lessons to their lives and the world in which they live. Toward that end, I have used various learning tools that recur throughout the book.

Case Studies

Economic theory is useful and interesting only if it can be applied to understanding actual events and policies. This book, therefore, contains numerous case studies that apply the theory that has just been developed.

In the News Boxes

One benefit that students gain from studying economics is a new perspective and greater understanding of news from around the world. To highlight this benefit, I have included excerpts from many newspaper and magazine articles, some of which are opinion columns written by prominent economists. These articles, together with my brief introductions, show how basic economic theory can be applied. Most of these boxes are new to this edition. Each news article ends with "Questions to Discuss," which can be used to start a dialogue in the classroom.

FYI Boxes

These boxes provide additional material "for your information." Some of them offer a glimpse into the history of economic thought. Others clarify technical issues. Still others discuss supplementary topics that instructors might choose to either discuss or skip in their lectures.

Ask the Experts Boxes

This feature summarizes results from the IGM Economic Experts Panel, an ongoing survey of several dozen prominent economists. Every few weeks, these experts are offered a statement and then asked whether they agree with it, disagree with it, or are uncertain about it. The survey results appear in the chapters near the coverage of the relevant topic. They give students a sense of when economists are united, when they are divided, and when they just don't know what to think.

Definitions of Key Concepts

When key concepts are introduced in the chapter, they are presented in **blue** typeface. In addition, their definitions are placed in the margins. This treatment should aid students in learning and reviewing the material.

Quick Quizzes

After each major section in a chapter, students are offered a brief multiple-choice Quick Quiz to check their comprehension of what they have just learned. If students cannot readily answer these quizzes, they should stop and review the material before continuing. The answers to all Quick Quizzes are available at the end of each chapter.

Chapter in a Nutshell

Each chapter concludes with a brief summary that reminds students of the most important lessons they have learned. Later in their study, it offers an efficient way to review for exams.

List of Key Concepts

A list of key concepts at the end of each chapter offers students a way to test their understanding of the new terms that have been introduced. Page references are included, so students can review the terms they do not understand.

Questions for Review

Located at the end of each chapter, questions for review cover the chapter's primary lessons. Students can use these questions to check their comprehension and prepare for exams.

Problems and Applications

Each chapter also contains a variety of problems and applications that ask students to apply the material they have learned. Some instructors may use these questions for homework assignments. Others may use them as a starting point for classroom discussions.

Alternative Versions of the Book

The book you are now holding is one of five versions of this text that are available for introducing students to economics. Cengage and I offer this menu of books because instructors differ in how much time they have and what topics they choose to cover. Here is a brief description of each:

- *Principles of Economics*. This complete version of the book contains all 38 chapters. It is designed for two-semester introductory courses that cover both microeconomics and macroeconomics.
- *Principles of Microeconomics.* This version contains 24 chapters and is designed for one-semester courses in introductory microeconomics.
- *Principles of Macroeconomics*. This version contains 24 chapters and is designed for one-semester courses in introductory macroeconomics. It contains a full development of the theory of supply and demand.
- *Brief Principles of Macroeconomics.* This shortened macro version of 19 chapters contains only one chapter on the basics of supply and demand. It is designed for instructors who want to jump to the core topics of macroeconomics more quickly.
- *Essentials of Economics.* This version of the book contains 24 chapters. It is designed for one-semester survey courses that cover the basics of both microeconomics and macroeconomics.

Table 1 shows which chapters are included in each book. Instructors who want more information about these alternative versions should contact their local Cengage representative.

Table 1

The Five Versions of This Book

	Principles of Economics	Principles of Microeconomics	Principles of Macroeconomics	Brief Principles of Macroeconomics	Essentials of Economics
Ten Principles of Economics	1	1	1	1	1
Thinking Like an Economist	2	2	2	2	2
Interdependence and the Gains from Trade	3	3	3	3	3
The Market Forces of Supply and Demand	4	4	4	4	4
Elasticity and Its Application	5	5	5		5
Supply, Demand, and Government Policies	6	6	6		6
Consumers, Producers, and the Efficiency of Markets	7	7	7		7
Application: The Costs of Taxation	8	8	8		8
Application: International Trade	9	9	9		9
Externalities	10	10			10
Public Goods and Common Resources	11	11			11
The Economics of Healthcare	12	12			
The Design of the Tax System	13	13			
The Costs of Production	14	14			12
Firms in Competitive Markets	15	15			13
Monopoly	16	16			14
Monopolistic Competition	17	17			
Oligopoly	18	18			
The Markets for the Factors of Production	19	19			
Earnings and Discrimination	20	20			
Income Inequality and Poverty	21	21			
The Theory of Consumer Choice	22	22			
Frontiers of Microeconomics	23	23			
Measuring a Nation's Income	24		10	5	15
Measuring the Cost of Living	25		11	6	16
Production and Growth	26		12	7	17
Saving, Investment, and the Financial System	27		13	8	18
The Basic Tools of Finance	28		14	9	19
Unemployment	29		15	10	20
The Monetary System	30		16	11	21
Money Growth and Inflation	31		17	12	22
Open-Economy Macroeconomics: Basic Concepts	32		18	13	
A Macroeconomic Theory of the Open Economy	33		19	14	
Aggregate Demand and Aggregate Supply	34		20	15	23
The Influence of Monetary and Fiscal Policy on Aggregate Demand	35		21	16	24
The Short-Run Trade-off between Inflation and Unemployment	36		22	17	
Six Debates over Macroeconomic Policy	37		23	18	
Appendix: How Economists Use Data	38	24	24	19	

Supplements

Cengage offers various supplements for instructors and students who use this book. These resources make teaching the principles of economics easy for the instructor and learning them easy for the student. David R. Hakes of the University of Northern Iowa, a dedicated teacher and economist, supervised the development of the supplements for this edition. A complete list of available supplements follows this Preface.

Optional Online Chapter on the Keynesian Cross

I have written a brief chapter on the Keynesian Cross (sometimes called the income-expenditure model) that complements the material on aggregate demand and aggregate supply. Instructors who want to teach this model can add this chapter to their students' e-books for no additional cost.

Translations and Adaptations

I am delighted that versions of this book are (or will soon be) available in many of the world's languages. Currently scheduled translations include Azeri, Chinese (in both standard and simplified characters), Croatian, Czech, Dutch, French, Georgian, German, Greek, Indonesian, Italian, Japanese, Korean, Macedonian, Montenegrin, Portuguese, Romanian, Russian, Serbian, and Spanish. In addition, adaptations of the book for Australian, Canadian, European, and New Zealand students are also available. Instructors who would like more information about these books should contact Cengage.

Acknowledgments

In writing this book, I benefited from the input of many talented people. Indeed, the list of people who have contributed to this project is so long, and their contributions so valuable, that it seems an injustice that only a single name appears on the cover.

Let me begin with my colleagues in the economics profession. The many editions of this text and its supplemental materials have benefited enormously from their input. In reviews and surveys, they have offered suggestions, identified challenges, and shared ideas from their own classroom experience. I am indebted to them for the perspectives they have brought to the text. Unfortunately, the list has become too long to thank those who contributed to previous editions, even though students reading the current edition are still benefiting from their insights.

Most important in this process has been David Hakes (University of Northern Iowa). David has served as a reliable sounding board for ideas and a hardworking partner with me in putting together the superb package of supplements.

A special thanks to my friend Jeff Sommer. For many years, Jeff was my editor at the *New York Times*. For this edition, he graciously read through the entire book, offering numerous suggestions for improvement. I am deeply grateful for his input.

The publishing team who worked on the book improved it tremendously. Jane Tufts, developmental editor, provided truly spectacular editing—as she always does. Joe Sabatino, economics Product Director, and Christopher Rader, Senior Product Manager,

did a splendid job of overseeing the many people involved in such a large project. Colleen Farmer, Allison Janneck, and Anita Verma, Senior Content Managers, were crucial in managing the whole project and putting together an excellent team to revise the supplements and, with Pradhiba Kannaiyan, project manager at MPS Limited, had the patience and dedication necessary to turn my manuscript into this book. Erin Griffin, Senior Designer, gave this book its clean, friendly look and designed the wonderful cover. Tiffany Lee, copyeditor, refined my prose, and Vikas Makkar, indexer, prepared a careful and thorough index. John Carey, Executive Marketing Manager, worked long hours getting the word out to potential users of this book. The rest of the Cengage team has, as always, been consistently professional, enthusiastic, and dedicated.

We have a top team of veterans who have worked across multiple editions producing the supplements that accompany this book. Working with those at Cengage, the following have been relentless in making sure that the suite of ancillary materials is unmatched in both quantity and quality. No other text comes close.

PowerPoint: Andreea Chiritescu (Eastern Illinois University)

Test Bank: Shannon Aucoin, Eugenia Belova, and Alex Lewis (in-house Subject Matter Experts)

Instructor manual: David Hakes (University of Northern Iowa)

I am also grateful to Sarah Lao and Nathan Sun, two star undergraduates at Harvard, who helped me check the page proofs for this edition.

As always, I must thank my "in-house" editor Deborah Mankiw. As the first reader of most things I write, she continued to offer just the right mix of criticism and encouragement.

Finally, I should mention my three children, Catherine, Nicholas, and Peter. Their contribution to this book was putting up with a father spending too many hours in his study. The four of us have much in common—not least of which is our love of ice cream (which becomes apparent in Chapter 4).

N. Gregory Mankiw
May 2022

Brief Contents

Contents

Preface: To the Student

"Economics is a study of mankind in the ordinary business of life." So wrote Alfred Marshall, the great 19th-century economist, in his textbook, *Principles of Economics*. Knowledge about the economy has advanced substantially since Marshall's time, but this definition of economics is as true today as it was in 1890, when the first edition of his text was published.

Why should you, as a student in the 21st century, embark on the study of economics? There are three reasons.

The first reason to study economics is that it will help you understand the world in which you live. Many questions about the economy might spark your curiosity. Why are apartments so hard to find in New York City? Why do airlines charge less for a round-trip ticket if the traveler stays over a Saturday night? Why is Scarlett Johansson paid so much to perform in movies? Why are living standards so meager in many African countries? Why do some countries have high rates of inflation while others have stable prices? Why are jobs easy to find in some years and hard to find in others? These are just a few of the questions that a course or two in economics will help you answer.

The second reason to study economics is that it will make you a more astute participant in the economy. Throughout your life, you'll make many economic decisions. As a student, you decide how many years to remain in school. Once you take a job, you'll decide how much of your income to spend, how much to save, and how to invest your savings. Someday you may find yourself running a small business or a large corporation, and you will decide how many workers to hire and what prices to charge for your products. Studying economics will give you a new perspective on how best to make these decisions. The insights explored in the coming chapters will not, by themselves, make you rich, but they may help in that endeavor.

The third reason to study economics is that it will help you understand the potential and limits of economic policy. Economic questions are always on the minds of policymakers in mayors' offices, governors' mansions, and the White House. What are the burdens of alternative forms of taxation? What are the effects of free trade with other countries? What is the best way to protect the environment? How does a government budget deficit affect growth? As a voter, you help choose the policies that guide the allocation of society's resources. An understanding of economics will help you carry out that responsibility. And who knows: Someday, you may end up as one of those policymakers yourself.

The principles of economics can be applied in many of life's situations. Whether the future finds you following the news, running a business, or sitting in the Oval Office, you will be glad that you studied economics.

N. Gregory Mankiw
May 2022

Chapter

1

Ten Principles of Economics

The word **economy** comes from the Greek word **oikonomos**, which means "one who manages a household." At first, the connection between households and economies may seem obscure. But in fact, they have much in common.

No matter how you picture a modern household, its members face endless decisions. Somehow, they must decide which members do which tasks and what each receives in return. Who cooks dinner? Who gets some extra dessert? Who cleans the bathroom? Who gets to drive the car? Whether a household's income is high, low, or somewhere in between, its resources (time, dessert, car mileage) must be allocated among alternative uses.

Like a household, a society faces countless decisions. It must find some way to decide what jobs will be done and who will do them. Society needs people to grow food, make clothing, and design software. Once society has allocated people (as well as land, buildings, and machines) to various jobs, it must distribute the goods and services they produce. It must decide who will eat potatoes and who will eat caviar, who will live in a grand manor and who will live in a fifth-floor walk-up.

ROMAN SAMBORSKYI/SHUTTERSTOCK.COM, ARNO VAN ENGELAND/EYEEM/GETTY IMAGES

scarcity
the limited nature of society's resources

These decisions are important because resources are scarce. **Scarcity** means that society has limited resources and, therefore, cannot produce all the goods and services people want. Just as members of a household cannot always get their desires satisfied, individuals in a society cannot always attain the standard of living to which they might aspire.

economics
the study of how society manages its scarce resources

Economics is the study of how society manages its scarce resources. In most societies, resources are allocated through the combined choices of millions of households and businesses. Economists examine how people make these choices: how much they work, what they buy, how much they save, how they invest their savings, and so on. Economists also study how people interact with one another. For instance, economists examine how buyers and sellers together determine the price at which a good is sold and the quantity that is sold. Finally, economists analyze the forces and trends that affect the overall economy, including the growth in average income, the fraction of the population that cannot find work, and the rate at which prices are rising.

Economics covers a wide range of topics and encompasses many approaches, but it is unified by several central ideas. This chapter discusses **Ten Principles of Economics**. Don't worry if you don't understand them all at first or if you aren't completely convinced that they are sensible or important. These ideas will be explored more fully in later chapters. This introduction to the ten principles will give you a sense of what economics is all about. Consider this chapter a preview of coming attractions.

1-1 How People Make Decisions

There is no mystery about what an economy is. Whether it encompasses Los Angeles, the United States, or the entire planet, an economy is just a group of people dealing with one another as they go about their lives. Because the behavior of an economy reflects the behavior of the individuals within it, the first four principles concern individual decision making.

1-1a Principle 1: People Face Trade-Offs

"There ain't no such thing as a free lunch." Grammar aside, this old saying contains much truth. To get one thing you want, you usually have to give up another thing you want. Making decisions requires trading off one goal for another.

Consider Selena, a student who is deciding how to use her most valuable resource—time. Selena can spend all her time studying economics, all her time studying psychology, or divide her time between the two. For every hour she devotes to one subject, she gives up an hour she could have used studying the other. And for every hour spent studying, she gives up an hour that could have been spent napping, bike riding, playing video games, or working at a job for some extra spending money.

Consider Selena's parents, who are deciding how to use the family income. They can spend it on food, clothing, or Selena's tuition. Or they can save some of their income for retirement or a future family vacation. When they allocate a dollar to one of these goods, they have one less dollar to spend on another.

As a society, people face other trade-offs. One classic trade-off is between "guns and butter." The more a society spends on the military, the less it can spend on consumer goods. Another critical trade-off is between a clean environment and the level of income. Laws that require firms to reduce pollution may raise the cost of

producing goods and services. Because of these higher costs, the firms are likely to earn smaller profits, pay lower wages, charge higher prices, or do some combination of these three things. While pollution regulations yield a cleaner environment and the improved health that comes with it, they may reduce the incomes of the regulated firms' owners, workers, and customers.

Another societal trade-off is between efficiency and equality. **Efficiency** means that society is getting the greatest benefits from its scarce resources. **Equality** means that those benefits are distributed uniformly among society's members. In other words, efficiency refers to the size of the economic pie, while equality refers to how evenly the pie is sliced.

efficiency
the property of society getting the most it can from its scarce resources

equality
the property of distributing economic prosperity uniformly among the members of society

These two goals can conflict. Consider, for instance, government policies aimed at reducing inequality. Some of these policies, such as welfare or unemployment insurance, help the members of society most in need. Others, such as the personal income tax, require the financially successful to contribute more than others to support the government. These policies increase equality but may decrease efficiency. When the government redistributes income from the rich to the poor, it reduces the reward for hard work for people at all income levels. As a result, people may work less and produce fewer goods and services. In other words, when the government cuts the economic pie into more equal slices, the pie sometimes shrinks.

Recognizing that people face trade-offs does not tell us what decisions are best. A student should not abandon the study of psychology just because doing so would free up time for studying economics. Society should not live with pollution just because environmental regulations might reduce our material standard of living. The government should not neglect the poor just because helping them would distort work incentives. Yet people will make better choices if they understand the options available to them. Our study of economics, therefore, starts by acknowledging life's trade-offs.

1-1b Principle 2: The Cost of Something Is What You Give Up to Get It

Because people face trade-offs, they need to compare the costs and benefits of alternative decisions. In many cases, however, the costs are not as obvious as they might first appear.

Consider the decision to attend college. The main benefits are intellectual enrichment and a lifetime of better job opportunities. But what are the costs? You might be tempted to add up the money spent on tuition, books, room, and board. Yet this total does not truly represent what you give up to spend a year in college.

This calculation has two problems. First, it includes some things that are not really costs of going to college. Even if you quit school, you need a place to sleep and food to eat. Room and board are college costs only to the extent that they exceed the cost of living and eating at home or in your apartment. Second, this calculation ignores the largest cost of going to college—your time. When you listen to lectures, read books, and write papers, you can't spend that time working and earning money. For most students, the earnings they forgo to attend school are the largest cost of their education.

The **opportunity cost** of an item is what you give up to get it. When making decisions, it's smart to take opportunity costs into account, and people often do. College athletes who can earn millions dropping out of school and playing professional sports understand that their opportunity cost of attending college is high. Not surprisingly, they sometimes decide that the benefit of a college education is not worth the cost.

opportunity cost
whatever must be given up to obtain some item

1-1c Principle 3: Rational People Think at the Margin

rational people
people who systematically and purposefully do the best they can to achieve their objectives

Economists often assume that people are rational. **Rational people** systematically and purposefully do the best they can to achieve their goals, given the available opportunities. As you study economics, you will encounter firms that decide how many workers to hire and how much product to make and sell to maximize profits. You will meet people who decide how much to work and what goods and services to buy to achieve the highest possible level of satisfaction. To be sure, human behavior is complex and sometimes deviates from rationality. But the assumption that people do the best they can is, economists have found, a good starting point to explain the decisions that people make.

Rational decision makers know that many issues in life are not black and white but involve shades of gray. At dinnertime, you don't ask yourself, "Should I fast or eat like a pig?" You are more likely to ask, "Should I take that extra spoonful of mashed potatoes?" When exams roll around, your decision is probably not between blowing them off and studying 24 hours a day but whether to spend an extra hour reviewing your notes instead of hanging out with friends. Economists use the term **marginal change** to describe an incremental adjustment to an existing plan of action. Keep in mind that **margin** means "edge," so marginal changes are small adjustments around the edges of what you are doing. Rational people make decisions by comparing **marginal benefits** and **marginal costs**.

marginal change
an incremental adjustment to a plan of action

For example, suppose you are deciding whether to watch a movie tonight. You pay $30 a month for a streaming service that gives you unlimited access to its film library, and you typically watch five movies a month. What cost should you consider when deciding whether to stream another movie? The answer might seem to be $30/5, or $6, the **average** cost of a movie. More relevant for your decision, however, is the **marginal** cost—the extra money that you have to pay if you stream another film. Here, the marginal cost is zero because you pay $30 regardless of how many movies you stream. In other words, at the margin, streaming a movie is free. The only cost of watching a movie tonight is the time it takes away from other activities, such as working at a job or (better yet) reading this textbook.

Thinking at the margin is also useful for business decisions. Consider an airline deciding how much to charge passengers who fly standby. Suppose that flying a 200-seat plane across the United States costs the airline $100,000. The average cost of each seat is $500 ($100,000/200). You might think that the airline should never sell a ticket for less than $500. But imagine that a plane is about to take off with ten empty seats, and Stanley, a standby passenger, is at the gate and willing to pay $300 for a seat. Should the airline sell him the ticket? Yes, it should. If the plane has empty seats, the cost of adding an extra passenger is tiny. The **average** cost of flying a passenger is $500, but the **marginal** cost is merely the cost of the can of soda Stanley will consume and the small bit of jet fuel needed to carry his weight. As long as Stanley pays more than the marginal cost, selling him the ticket is profitable. A rational airline can benefit from thinking at the margin.

VANTAGE_DS/SHUTTERSTOCK.COM

Many movie streaming services set the marginal cost of a movie equal to zero.

Marginal analysis explains some otherwise puzzling phenomena. For example, why is water so cheap while diamonds are so expensive? You might think it should be the other way around: Humans need water to survive, but diamonds merely glitter. Yet people are willing to pay much more for a diamond than for a cup of water. Economists have figured this out. A person's willingness to pay for a good is based on the marginal benefit that an extra unit of the good would yield. The marginal benefit, in turn, depends on how many units a person already has. Water is essential but plentiful, so the marginal benefit of an extra

cup is small. By contrast, no one needs diamonds to survive, but because they are so rare, the marginal benefit of an extra gem is large.

A rational decision maker takes an action if and only if the action's marginal benefit exceeds its marginal cost. This principle explains why people use streaming services as much as they do, why airlines sell tickets below average cost, and why people pay more for diamonds than for water. It can take a while to get used to the logic of marginal thinking, but the study of economics will give you ample opportunity to practice.

1-1d Principle 4: People Respond to Incentives

An **incentive** is something that induces a person to act, such as the prospect of a punishment or reward. People respond to incentives if they make decisions by comparing costs and benefits. Incentives play a central role in economics. One economist went so far as to say that the entire field could be summarized as simply, "People respond to incentives. The rest is commentary."

incentive
something that induces a person to act

Incentives are key to analyzing how markets work. For example, when the price of apples rises, people decide to eat fewer apples. At the same time, apple orchards decide to hire more workers and harvest more apples. In other words, a higher price provides an incentive for buyers to consume less and for sellers to produce more. As we will see, the influence of prices on the behavior of consumers and producers is crucial to how a market economy allocates scarce resources.

Public policymakers need to pay attention to incentives: Many policies change the costs or benefits that people face and, as a result, alter their behavior. A tax on gasoline, for instance, encourages people to drive more fuel-efficient cars and shift to electric ones. That is one reason many people drive electric cars in Norway, where gas taxes are high, and why big SUVs are so popular in the United States, where gas taxes are low. A higher gas tax also encourages people to carpool, take public transportation, ride bikes, and live closer to work.

When policymakers fail to consider incentives, the policies they enact may have unintended consequences. For example, consider auto safety. Today, all cars have seat belts, but this wasn't true 60 years ago. In 1965, Ralph Nader's book *Unsafe at Any Speed* generated much public concern over auto safety. Congress responded with laws requiring seat belts as standard equipment on new cars.

How does a seat belt law affect safety? The direct effect is obvious: When a person wears a seat belt, the likelihood of surviving an auto accident rises. But that's not the end of the story. The law also affects behavior by altering incentives. The relevant behavior here is the speed and care with which drivers operate their cars. Driving slowly and carefully is costly because it uses the driver's time and energy. When deciding how to drive, rational people compare, perhaps unconsciously, the marginal benefit from safer driving with the marginal cost. They drive more slowly and carefully when the benefit of increased safety is high. For example, when road conditions are icy, people drive more attentively and at lower speeds than they do when road conditions are clear.

Consider how a seat belt law alters a driver's cost–benefit calculation. Buckling up makes accidents less costly by reducing the risk of injury or death. It is as if road conditions had improved: When conditions are safer, people drive faster and less carefully. That may be fine for motorists, whose risk of injury in an accident is reduced because of seat belts. But if faster, less careful driving leads to more accidents, the seat belt law adversely affects pedestrians, who are more likely to be in an accident but (unlike drivers) don't benefit from added protection.

This discussion of incentives and seat belts isn't idle speculation. In a classic 1975 study, the economist Sam Peltzman tested the theory and found that auto-safety

laws have had many of these effects. According to Peltzman, these laws give rise not only to fewer deaths per accident but also to more accidents. He concluded that the net result is little change in driver deaths and an increase in pedestrian deaths.

Peltzman's analysis of auto safety is an offbeat and controversial example of the principle that people respond to incentives. When analyzing any policy, it is important to consider not only the direct effects but also the indirect effects that work through incentives. If the policy alters incentives, people may change their behavior.

QuickQuiz

1. Economics is best defined as the study of
   a. how society manages its scarce resources.
   b. how to run a business most profitably.
   c. how to predict inflation, unemployment, and stock prices.
   d. how the government can protect people from unchecked self-interest.
2. Your opportunity cost of going to a movie is
   a. the price of the ticket.
   b. the price of the ticket plus the cost of any soda and popcorn you buy at the theater.
   c. the total cash expenditure needed to go to the movie plus the value of your time.
   d. zero, as long as you enjoy the movie and consider it a worthwhile use of time and money.
3. A marginal change is one that
   a. is not important for public policy.
   b. incrementally alters an existing plan.
   c. makes an outcome inefficient.
   d. does not influence incentives.
4. Because people respond to incentives,
   a. policymakers can alter outcomes by changing punishments or rewards.
   b. policies can have unintended consequences.
   c. society faces a trade-off between efficiency and equality.
   d. All of the above are correct.

Answers are at the end of the chapter.

1-2 How People Interact

The first four principles discussed how individuals make decisions. The next three concern how people interact with one another.

1-2a Principle 5: Trade Can Make Everyone Better Off

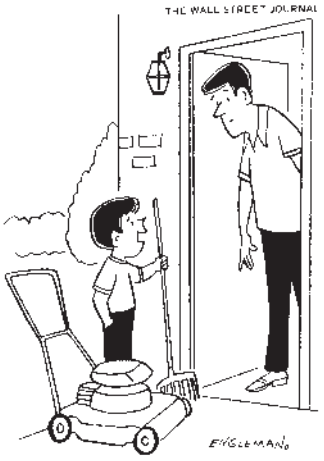
You may have heard on the news that China is the United States' competitor in the world economy. In some ways, this is true. Chinese and U.S. companies compete for customers in the markets for clothing, toys, solar panels, automobile tires, and many other items.

Yet it is easy to be misled when thinking about competition among countries. Trade between the United States and China is not like a sports contest in which one side wins and the other side loses. The opposite is true: Trade between two countries can make each country better off. Even when trade in the world economy is competitive, it can lead to a win–win outcome for the countries involved.

To see why, consider how trade affects a family. When family members look for jobs, they compete against the members of other families who are looking for jobs. Families also compete with one another when they go shopping because each wants to buy the best goods at the lowest prices. In a sense, each family in an economy competes with all other families.

Despite this competition, a family would not be better off isolating itself from other families. If it did, it would need to grow its own food, sew its own clothes, and build its own home. Clearly, a family gains much from being able to trade with others. Trade allows everyone to specialize in the activities they do best, whether it is farming, sewing, or home building. By trading with others, people can buy a greater variety of goods and services at a lower cost.

Like families, countries benefit from trading with one another. Trade allows countries to specialize in what they do best and to enjoy a greater variety of goods and services. The Chinese, as well as the French, Brazilians, and Nigerians, are as much the United States' partners in the world economy as they are its competitors.



FROM THE WALL STREET JOURNAL - PERMISSION, CARTOON FEATURES SYNDICATE

"For $5 a week you can watch baseball without being nagged to cut the grass!"

1-2b Principle 6: Markets Are Usually a Good Way to Organize Economic Activity

The collapse of Communism in the Soviet Union and Eastern Europe in the late 1980s and early 1990s was one of the last century's transformative events. For the most part, countries in the Soviet bloc operated on the premise that government officials were in the best position to allocate the economy's scarce resources. These central planners decided what goods and services were produced, how much was produced, and who produced and consumed them. The theory behind central planning was that the government needed to organize economic activity to ensure the well-being of the country and of like-minded nations.

Most countries that once had centrally planned economies have now shifted toward market economies. In a **market economy**, the decisions of a central planner are replaced by those of millions of firms and households. Firms decide whom to hire and what to make. Households decide where to work and what to buy with their incomes. These firms and households interact in the marketplace, where prices and self-interest guide their decisions.

market economy
an economy that allocates resources through the decentralized decisions of many firms and households as they interact in markets for goods and services

At first glance, the success of market economies may seem puzzling because no one appears to be looking out for the well-being of society as a whole. Competitive markets contain many buyers and sellers of numerous goods and services, all of them interested primarily in their own well-being. Yet despite decentralized decision making and self-interested decision makers, market economies have proven remarkably successful in organizing economic activity to promote prosperity.

In his 1776 book, *An Inquiry into the Nature and Causes of the Wealth of Nations*, Adam Smith made the most famous observation in all of economics: Firms and households in competitive markets act as if they are guided by an "invisible hand" that leads them to desirable outcomes. One of the chief goals of this book is to understand how this invisible hand works its magic.

As you study economics, you will learn that prices are the instrument with which the invisible hand directs economic activity. In a competitive market, sellers look at the price when deciding how much to supply, and buyers look at the price when deciding how much to demand. As a result of their decisions, the price reflects both the sellers' costs of production and the value of the good to the buyers. Smith's great insight was that prices adjust to guide market participants to reach outcomes that, in many cases, maximize the well-being of society as a whole.

Smith's insight has an important corollary: When a government prevents prices from adjusting to supply and demand, it impedes the invisible hand's ability to coordinate the decisions of the firms and households that make up an economy. This corollary explains the adverse effect of most taxes on the allocation of resources:

FYI

Adam Smith and the Invisible Hand

It may be only a coincidence that Adam Smith's great book *The Wealth of Nations* was published in 1776, the exact year in which American revolutionaries signed the Declaration of Independence. But the two documents share a point of view that was prevalent at the time: Individuals are usually best left to their own devices, without the heavy hand of government directing their actions. This philosophy provides the intellectual foundation for the market economy and, more generally, for a free society.

Why do decentralized market economies work reasonably well? Is it because people can be trusted to treat one another with love, kindness, and generosity? Not at all. Here is Adam Smith's description of how people interact in a market economy:

LIBRARY OF CONGRESS PRINTS AND PHOTOGRAPHS DIVISION [LC-USZ62-17407]

Adam Smith

> Man has almost constant occasion for the help of his brethren, and it is in vain for him to expect it from their benevolence only. He will be more likely to prevail if he can interest their self-love in his favour, and show them that it is for their own advantage to do for him what he requires of them. . . . Give me that which I want, and you shall have this which you want, is the meaning of every such offer; and it is in this manner that we obtain from one another the far greater part of those good offices which we stand in need of.
>
> 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. Nobody but a beggar chooses to depend chiefly upon the benevolence of his fellow-citizens. . . .
>
> Every individual . . . neither intends to promote the public interest, nor knows how much he is promoting it. . . . He intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention. Nor is it always the worse for the society that it was no part of it. By pursuing his own interest he frequently promotes that of the society more effectually than when he really intends to promote it.

Smith is saying that participants in the economy are motivated by self-interest and that the "invisible hand" of the marketplace guides them into promoting general economic well-being.

Many of Smith's insights remain at the center of modern economics. The coming chapters will express Smith's conclusions more precisely and analyze more fully the strengths and weaknesses of the market's invisible hand. ■

Taxes distort prices and the decisions of firms and households. It also explains the problems caused by policies that dictate prices, such as rent control. And it explains the economic failure of Communist countries, where prices were set not in the marketplace but by central planners. These planners lacked the overwhelming amount of complex and ever-changing information about producers' costs and consumers' tastes, which, in a market economy, is reflected in prices. Central planners failed because they tried to run the economy with one hand tied behind their backs—the invisible hand of the marketplace.

Case Study

Adam Smith Would Have Loved Uber

You may have never lived in a centrally planned economy, but if you have tried to hail a cab in a major city, you have likely experienced a highly regulated market. In many cities, the local government imposes strict controls in the market for taxis. The rules usually go well beyond the regulation of insurance and safety. For example, the government may limit entry

into the market by approving only a certain number of taxi medallions or permits. It may determine the prices that taxis are allowed to charge. The government uses its police powers—that is, the threat of fines or jail time—to keep unauthorized drivers off the streets and prevent drivers from charging unauthorized prices.

In 2009, however, this highly controlled market was invaded by a disruptive force: Uber, a company that provides a smartphone app to connect passengers and drivers. Because Uber cars do not roam the streets looking for taxi-hailing pedestrians, they are technically not taxis and so are not subject to the same regulations. But they offer a similar service. Indeed, rides from Uber—and from Uber's competitors that have since entered many markets—are often more convenient. On a cold, rainy day, who wants to wait by the side of the road for an empty cab to drive by? It is more pleasant to remain inside, use a smartphone to arrange a ride, and stay warm and dry until the car arrives.

Uber cars often charge less than taxis, but not always. Uber's prices rise significantly when there is a surge in demand, such as during a sudden rainstorm or late on New Year's Eve, when numerous tipsy partygoers are looking for a safe way to get home. By contrast, regulated taxis are typically prevented from surge pricing.

Not everyone is fond of Uber. Drivers of traditional taxis complain that this new competition reduces their income. This is hardly a surprise: Suppliers of goods and services often dislike new competitors. But vigorous competition among producers makes a market work well for consumers.

That is why economists embraced Uber's entry into the market. A 2014 survey of several dozen prominent economists asked whether car services such as Uber increased consumer well-being. Every single economist said "Yes." The economists were also asked whether surge pricing increased consumer well-being. "Yes," said 85 percent of them. Surge pricing makes consumers pay more at times, but because Uber drivers respond to incentives, it also increases the quantity of car services supplied when they are most needed. Surge pricing also helps allocate the services to those consumers who value them most highly and reduces the costs of searching and waiting for a car.

If Adam Smith were alive today, he would surely have a ride-sharing app on his phone. ●

RICHARD LEVINE/ALAMY STOCK PHOTO

Technology can improve this market.

1-2c Principle 7: Governments Can Sometimes Improve Market Outcomes

If the invisible hand is so great, what is left for a government to do in an economy? One purpose of studying economics is to refine your view about the proper role and scope of government policy.

One reason we need government is that the invisible hand can work its magic only if the government enforces the rules and maintains the institutions that are key to a market economy. Most importantly, market economies need institutions to enforce **property rights** so individuals can own and control scarce resources. Farmers won't grow food if they expect their crop to be stolen, restaurants won't serve meals if many customers leave before paying, and film companies won't produce movies if too many people pirate copies. Market participants rely on government-provided police and courts to enforce their rights, and the invisible hand works well only if the legal system does.

property rights the ability of an individual to own and exercise control over scarce resources

Another reason we need government is that the invisible hand, while powerful, is not omnipotent. There are two broad rationales for a government to intervene in the economy and change the allocation of resources that people would choose on

their own: to promote efficiency or to promote equality. That is, policies can aim either to enlarge the economic pie or to change how the pie is sliced.

Consider the goal of efficiency. The invisible hand usually leads markets to allocate resources to maximize the size of the economic pie, but this is not always the case. Economists use the term **market failure** to refer to a situation in which the market does not produce an efficient allocation of resources on its own. One possible cause of market failure is an **externality**, which is the impact of one person's actions on the well-being of a bystander. The classic example of an externality is pollution. When the production of a good pollutes the air and creates health problems for those who live near the factories, the market may fail to take this cost into account. Another possible cause of market failure is **market power**, which refers to the ability of a single person or firm (or a small group of them) to unduly influence market prices. For example, if everyone in town needs water but there is only one well, the owner of the well does not face the rigorous competition with which the invisible hand normally keeps self-interest in check; the well owner may take advantage of this opportunity by restricting the output of water and charging a higher price. In the presence of externalities or market power, well-designed public policy can enhance efficiency.

market failure
a situation in which a market left on its own does not allocate resources efficiently

externality
the impact of one person's actions on the well-being of a bystander

market power
the ability of a single economic actor (or small group of actors) to have a substantial influence on market prices

Now consider the goal of equality. Even when the invisible hand yields efficient outcomes, it can nonetheless leave large disparities in well-being. A market economy rewards people according to their ability to produce things that other people are willing to pay for. The world's best basketball player earns more than the world's best chess player simply because people are willing to pay more to watch basketball than chess. The invisible hand does not ensure that everyone has enough food, decent clothing, and adequate healthcare. This inequality may call for government intervention. In practice, many public policies, such as the income tax and the welfare system, aim to achieve a more equal distribution of well-being.

To say that the government **can** improve market outcomes does not mean that it always **will**. Public policy is made not by angels but by an imperfect political process. Sometimes, policies are designed to reward the politically powerful. Sometimes, they are made by well-intentioned leaders who are ill-informed. As you study economics, you will become a better judge of when a government policy is justifiable because it promotes efficiency or equality and when it is not.

QuickQuiz

5. International trade benefits a nation when
   a. its revenue from selling abroad exceeds its outlays from buying abroad.
   b. its trading partners experience reduced economic well-being.
   c. all nations specialize in doing what they do best.
   d. no domestic jobs are lost because of trade.
6. Adam Smith's "invisible hand" refers to
   a. the subtle and often hidden methods that businesses use to profit at consumers' expense.
   b. the ability of competitive markets to reach desirable outcomes, despite the self-interest of market participants.
   c. the ability of government regulation to benefit consumers, even if the consumers are unaware of the regulations.
   d. the way in which producers or consumers in unregulated markets impose costs on innocent bystanders.
7. Governments may intervene in a market economy in order to
   a. protect property rights.
   b. correct a market failure due to externalities.
   c. achieve a more equal distribution of income.
   d. All of the above are correct.

Answers are at the end of the chapter.

1-3 How the Economy as a Whole Works

We started by discussing how individuals make decisions and then looked at how people interact. All these decisions and interactions together make up "the economy." The last three principles concern the workings of the economy as a whole.

1-3a Principle 8: A Country's Standard of Living Depends on Its Ability to Produce Goods and Services

The differences in living standards around the world are staggering. In 2019, the average American earned about $65,000. In the same year, the average German earned about $56,000, the average Chinese earned about $17,000, and the average Nigerian earned only $5,000. This variation in average income is reflected in measures of quality of life. People in high-income countries have more computers, more cars, better nutrition, better healthcare, and a longer life expectancy than do those in low-income countries.

Changes in living standards over time are also large. In the United States, incomes have historically grown about 2 percent per year (after adjusting for changes in the cost of living). At this rate, the average income doubles every 35 years. Over the past century, the average U.S. income has risen about eightfold.

What explains these large differences across countries and over time? The answer is simple. Almost all variation in living standards is attributable to differences in countries' **productivity**—that is, the amount of goods and services produced by each unit of labor input. In nations where workers can produce a large quantity of goods and services per hour, most people enjoy a high standard of living; in nations where workers are less productive, most people endure a more meager existence. Similarly, the growth rate of a nation's productivity determines the growth rate of its average income.

productivity
the quantity of goods and services produced from each unit of labor input

The relationship between productivity and living standards is simple, but its implications are far-reaching. If productivity is the main determinant of living standards, other explanations must be less important. For example, it might be tempting to credit generous employers or vigorous labor unions for the rising incomes of American workers over the past century. Yet the real hero of American workers is their rising productivity. As another example, some commentators have suggested that increased international competition explains the slowdown in U.S. income growth that began in the mid-1970s. But the real villain was flagging productivity growth in the United States.

The relationship between productivity and living standards has profound implications for public policy. When thinking about how any policy will affect living standards, the key question is how it will affect the economy's ability to produce goods and services. To boost living standards, policymakers need to raise productivity by ensuring that workers are well trained, have the tools they need to produce goods and services, and have access to the best available technology.

1-3b Principle 9: Prices Rise When the Government Prints Too Much Money

In January 1921, a daily newspaper in Germany cost 0.30 marks. Less than two years later, in November 1922, the same newspaper cost 70,000,000 marks. All other prices in the economy rose by similar amounts. This episode is one of history's most spectacular examples of **inflation**, an increase in the overall level of prices in the economy.

inflation
an increase in the overall level of prices in the economy

The United States has never experienced inflation even close to that of Germany in the 1920s, but inflation has at times been a problem. During the 1970s, the overall level of prices more than doubled, and President Gerald Ford called inflation "public enemy number one." By contrast, inflation in the first two decades of the 21st century

"Well it may have been 68 cents when you got in line, but it's 74 cents now!"

ran about 2 percent per year; at this rate, it takes 35 years for prices to double. Because high inflation imposes various costs on society, keeping inflation at a reasonable rate is a goal of economic policymakers around the world.

What causes inflation? In almost all cases of large or persistent inflation, the culprit is growth in the quantity of money. When a government creates large quantities of the nation's money, the value of the money falls. In Germany in the early 1920s, when prices were, on average, tripling every month, the quantity of money was also tripling every month. Although less dramatic, the history of the United States points to a similar conclusion: The high inflation of the 1970s was associated with rapid growth in the quantity of money, and the return of low inflation in the 1980s was associated with slower growth in the quantity of money.

In 2022, as this book was going to press, U.S. inflation was surging. In February of that year, consumer prices were 7.9 percent higher than a year earlier, the highest inflation rate in 40 years. During the economic downturn caused by the coronavirus pandemic in 2020, the government alleviated the hardship with large increases in spending, and the quantity of money in the economy rose significantly. These policies, together with supply disruptions due to the pandemic, contributed to rising inflation. The key question was whether the inflation surge would be transitory, as many government officials believed, or whether it would become embedded in the economy, as occurred in the 1970s. The outcome would depend, in large part, on future monetary policy.

1-3c Principle 10: Society Faces a Short-Run Trade-Off between Inflation and Unemployment

While an increase in the quantity of money primarily raises prices in the long run, the short-run story is more complex. Most economists describe the short-run effects of money growth as follows:

- Increasing the amount of money in the economy stimulates the overall level of spending and thus the demand for goods and services.
- Higher demand will, over time, cause firms to raise their prices, but in the meantime, it encourages them to hire more workers and produce a larger quantity of goods and services.
- More hiring means lower unemployment.

This line of reasoning leads to one final economy-wide trade-off: a short-run trade-off between inflation and unemployment.

Some economists still question these ideas, but most accept that society faces a short-run trade-off between inflation and unemployment. This simply means that, over a period of a year or two, many economic policies push inflation and unemployment in opposite directions. Policymakers face this trade-off regardless of whether inflation and unemployment both start out at high levels (as they did in the early 1980s), at low levels (as they did in the late 2010s), or someplace in between. This short-run trade-off plays a key role in the analysis of the **business cycle**—the irregular and largely unpredictable fluctuations in economic activity, as measured by the production of goods and services or the number of people employed.

business cycle fluctuations in economic activity, such as employment and production

Policymakers can exploit the short-run trade-off between inflation and unemployment using various policy instruments. By changing the amount that the government spends, the amount it taxes, or the amount of money it prints, policymakers can influence the overall demand for goods and services. Changes in demand, in turn, influence the combination of inflation and unemployment that the economy experiences in the short run. Because these instruments of economic policy are so powerful, how policymakers should use them is the subject of continuing debate.

QuickQuiz

8. The main reason that some nations have higher average living standards than others is that
   a. the richer nations have exploited the poorer ones.
   b. the governments of some nations have created more money.
   c. some nations have stronger laws protecting worker rights.
   d. some nations have higher levels of productivity.
9. If a nation has high and persistent inflation, the most likely explanation is
   a. the government creating excessive amounts of money.
   b. unions bargaining for excessively high wages.
   c. the government imposing excessive levels of taxation.
   d. firms using their market power to enforce excessive price hikes.
10. If a government uses the tools of monetary policy to reduce the demand for goods and services, the likely result is _____ inflation and _____ unemployment in the short run.
    a. lower; lower
    b. lower; higher
    c. higher; higher
    d. higher; lower

Answers are at the end of the chapter.

1-4 Conclusion

You now have a taste of what economics is all about. In the coming chapters, we will develop many specific insights about people, markets, and economies. Mastering them will take some effort, but the task is not overwhelming. The field of economics is based on a few big ideas that can be applied in many situations.

Throughout this book, we will refer to the **Ten Principles of Economics** introduced in this chapter and summarized in Table 1. Keep these building blocks in mind. Even the most sophisticated economic analysis is founded on these ten principles.

Table 1

Ten Principles of Economics

How People Make Decisions

1. People face trade-offs.
2. The cost of something is what you give up to get it.
3. Rational people think at the margin.
4. People respond to incentives.

How People Interact

5. Trade can make everyone better off.
6. Markets are usually a good way to organize economic activity.
7. Governments can sometimes improve market outcomes.

How the Economy as a Whole Works

8. A country's standard of living depends on its ability to produce goods and services.
9. Prices rise when the government prints too much money.
10. Society faces a short-run trade-off between inflation and unemployment.

Chapter in a Nutshell

- The fundamental lessons about individual decision making are that people face trade-offs among alternative goals, that the cost of any action is measured in terms of forgone opportunities, that rational people make decisions by comparing marginal costs and marginal benefits, and that people change their behavior in response to the incentives they face.
- The fundamental lessons about economic interactions among people are that trade and interdependence can be mutually beneficial, that markets are usually a good way of coordinating economic activity, and that governments can potentially improve market outcomes by remedying a market failure or by promoting greater economic equality.
- The fundamental lessons about the economy as a whole are that productivity is the ultimate source of living standards, that growth in the quantity of money is the ultimate source of inflation, and that society faces a short-run trade-off between inflation and unemployment.

Key Concepts

scarcity, p. 2
economics, p. 2
efficiency, p. 3
equality, p. 3
opportunity cost, p. 3
rational people, p. 4
marginal change, p. 4
incentive, p. 5
market economy, p. 7
property rights, p. 9
market failure, p. 10
externality, p. 10
market power, p. 10
productivity, p. 11
inflation, p. 11
business cycle, p. 12

Questions for Review

1. Give three examples of important trade-offs that you face in your life.
2. What items would you include to figure out the opportunity cost of a trip to an amusement park?
3. Water is necessary for life. Is the marginal benefit of a glass of water large or small?
4. Why should policymakers think about incentives?
5. Why isn't trade between two countries like a game in which one country wins and the other loses?
6. What does the "invisible hand" of the marketplace do?
7. What are the two main causes of market failure? Give an example of each.
8. Why is productivity important?
9. What is inflation, and what causes it?
10. How are inflation and unemployment related in the short run?

Problems and Applications

1. Describe some of the trade-offs faced by each of the following:
   a. a family deciding whether to buy a car
   b. a member of Congress deciding how much to spend on national parks
   c. a company president deciding whether to open a new factory
   d. a professor deciding how much to prepare for class
   e. a recent college graduate deciding whether to go to graduate school
   f. a single parent with small children deciding whether to take a job
2. You are trying to decide whether to take a vacation. Most of the costs of the vacation (airfare, hotel, and forgone wages) are measured in dollars, but the benefits of the vacation are psychological. How can you compare the benefits to the costs?
3. You were planning to spend Saturday working at your part-time job, but a friend asks you to go skiing. What is the true cost of going skiing? Now suppose you had been planning to spend the day studying at the library. What is the cost of going skiing in this case? Explain.

4. You win $100 in a basketball pool. You have a choice between spending the money now and putting it away for a year in a bank account that pays 5 percent interest. What is the opportunity cost of spending the $100 now?
5. The company that you manage has invested $5 million in developing a new product, but the development is not quite finished. At a recent meeting, your salespeople report that the introduction of competing products has reduced the expected sales of your new product to $3 million. If it would cost $1 million to finish development and make the product, should you go ahead and do so? What is the most that you should pay to complete development?
6. A 1996 bill reforming the federal government's antipoverty programs limited many welfare recipients to only two years of benefits.
   a. How did this change affect the incentives for working?
   b. How might this change represent a trade-off between equality and efficiency?
7. Explain whether each of the following government activities is motivated by a concern about equality or a concern about efficiency. In the case of efficiency, discuss the type of market failure involved.
   a. regulating cable TV prices
   b. providing some low-income people with vouchers that can be used to buy food
   c. prohibiting smoking in public places
   d. breaking up Standard Oil (which once owned 90 percent of all U.S. oil refineries) into several smaller companies
   e. imposing higher personal income tax rates on people with higher incomes
   f. enacting laws against driving while intoxicated
8. Discuss each of the following statements from the standpoints of equality and efficiency.
   a. "Everyone in society should be guaranteed the best healthcare possible."
   b. "When workers are laid off, they should be able to collect unemployment benefits until they find a new job."
9. In what ways is your standard of living different from that of your parents or grandparents when they were your age? Why have these changes occurred?
10. Suppose Americans decide to save more of their incomes. If banks lend this extra saving to businesses that use the funds to build new factories, how might this lead to faster growth in productivity? Who do you suppose benefits from the higher productivity? Is society getting a free lunch?
11. During the Revolutionary War, the American colonies could not raise enough tax revenue to fully fund the war effort. To make up the difference, the colonies decided to print more money. Printing money to cover expenditures is sometimes referred to as an "inflation tax." Who do you think is being "taxed" when more money is printed? Why?

QuickQuiz Answers

1. **a** 2. **c** 3. **b** 4. **d** 5. **c** 6. **b** 7. **d** 8. **d** 9. **a** 10. **b**

Chapter

2

Thinking Like an Economist

The goal of this book is to help you think like an economist. This can be useful in a thousand ways. When you try to make sense of a news story, manage the finances of your household or business, or evaluate a politician's promises about problems ranging from local traffic congestion to global climate change, knowing some economics can help you think more sensibly and systematically. And that kind of thinking will lead you to better outcomes.

In all fields of study, specialists develop their own terminology and methods. Mathematicians talk about axioms, integrals, and vector spaces. Psychologists talk about ego, id, and cognitive dissonance. Lawyers talk about venue, torts, and promissory estoppel. Economists are no different. Supply, demand, elasticity, comparative advantage, consumer surplus, deadweight loss—these terms are part of the economist's language. In the coming chapters, you will encounter many new terms and some familiar words that economists use in specialized ways. At first, the terms and technicalities may seem needlessly arcane, and in daily life, many of them are. But understanding them will give you a new and useful way of thinking about the world in which you live. This book will guide you gently through the thicket.

Before delving into the substance and details of economics, it is helpful to have an overview of how economists look at the world. This chapter discusses the field's methodology. What is distinctive about how economists confront a question? What does it mean to think like an economist?

GAUDILAB/SHUTTERSTOCK.COM, LECHATNOIR/E+/GETTY IMAGES

2-1 The Economist as Scientist

J.B. HANDELSMAN/THE NEW YORKER COLLECTION/THE CARTOON BANK

"I'm a social scientist, Michael. That means I can't explain electricity or anything like that, but if you ever want to know about people, I'm your man."

Economists try to address their subject with a scientist's objectivity. They approach the study of the economy much as a physicist approaches the study of matter and a biologist approaches the study of life: They devise theories, collect data, and then analyze the data to verify or refute their theories.

The claim that economics is a science can seem odd. After all, economists don't work with test tubes or telescopes, and they don't wear white lab coats. Like other social scientists, they study human beings, a subject that everyone knows something about without the need for a university degree. The essence of science, however, is the **scientific method**—the dispassionate development and testing of theories about how the world works. This method of inquiry is as applicable to studying a nation's economy as it is to studying the earth's gravity or a species' evolution. As Albert Einstein put it, "The whole of science is nothing more than the refinement of everyday thinking."

Einstein's comment is as true for economics as it is for physics, but most people are not accustomed to looking at society through a scientific lens. Let's consider some of the ways economists apply the logic of science to examine how an economy works.

2-1a The Scientific Method: Observation, Theory, and More Observation

Isaac Newton, the 17th-century scientist and mathematician, told his biographer that he became intrigued one day when he saw an apple fall from a tree. Why did the apple always fall straight down to the earth? Newton's musing led him to develop a theory of gravity that applies not only to a falling apple but to any two objects in the universe. Subsequent testing of Newton's theory has shown that it works well in many circumstances (but not all, as Einstein would later show). Because Newton's theory has been so successful at explaining what we observe around us, it is still taught in physics courses.

A similar interplay between theory and observation occurs in economics. An economist who lives in a country with rapidly increasing prices may be moved by this observation to develop a theory of inflation. The theory might assert that high inflation arises when the government issues too much money. To test this theory, the economist could collect and analyze data on prices and money from many different countries. If the growth in the quantity of money were unrelated to the rate of price increase, the economist would start to doubt the validity of this theory of inflation. If money growth and inflation were correlated in international data, as in fact they often are, the economist would become more confident in the proposed theory.

Although economists use theory and observation like other scientists, they face an obstacle that makes their task challenging: Conducting experiments is often impractical. Physicists studying gravity can drop objects in their laboratories to test their theories. By contrast, economists studying inflation are not allowed to manipulate a nation's monetary policy simply to generate useful data. Economists, like astronomers and evolutionary biologists, usually make do with whatever data the world gives them.

To substitute for laboratory experiments, economists pay close attention to the natural experiments offered by history. When a war in the Middle East interrupts the supply of crude oil, for instance, oil prices skyrocket around the world. For consumers of oil and oil products, such an event raises the cost of living. For policymakers,

it poses a difficult choice about how best to respond. But for economic scientists, the event provides an opportunity to study the effects of a key natural resource on the world's economies. Throughout this book, we consider many historical episodes. Studying them yields insights into the economy of the past and helps to illustrate and evaluate economic theories of the present.

2-1b The Role of Assumptions

If you ask physicists how long it would take a marble to fall from the top of a ten-story building, they will likely answer by assuming that the marble falls in a vacuum. Yet this assumption is false. The building is surrounded by air, which exerts friction on the falling marble and slows it down. Why provide answers that ignore the complexity of the real world? In this case, physicists will point out that the friction on the marble is so small that its effect is negligible. Assuming the marble falls in a vacuum simplifies the problem without substantially affecting the answer. Physicists understand, however, that for a more precise answer, they would have to revisit their assumptions and perform a more sophisticated analysis.

Economists make assumptions for the same reason: Assumptions can simplify the complex world and make it easier to understand. To study international trade, for example, we might assume that the world consists of only two countries and that each country produces only two goods. We know that this isn't an accurate representation of the real world, which consists of many countries producing thousands of different types of goods. But by assuming the world has only two countries and two goods, we can focus on the essence of the problem. After analyzing international trade in this simplified imaginary world, we are in a better position to understand trade in the more complex world in which we live.

The art in scientific thinking—whether in physics, biology, or economics—is deciding which assumptions to make. Suppose, for instance, that instead of dropping a marble from the top of a building, we were dropping a beach ball of the same weight. Our physicist would realize that the assumption of no friction is wildly inaccurate in this case: Friction exerts a greater force on the beach ball because it is much larger than a marble. Pretending that gravity works in a vacuum is reasonable when studying a falling marble, but it would lead to large errors when studying a falling beach ball.

Similarly, economists use different assumptions to answer different questions. Suppose, for example, that we want to study what happens to the economy when the government changes the number of dollars in circulation. An important piece of this analysis, it turns out, is how prices respond. Many prices in the economy change infrequently: The newsstand prices of magazines, for instance, change only once every few years. Knowing this fact may lead us to make different assumptions for different time horizons. When studying the short-run effects of the policy, we may assume that prices do not change much. We may even make the extreme assumption that all prices are completely fixed. When studying the long-run effects of the policy, however, we may assume that all prices are completely flexible. Just as physicists use different assumptions when studying falling marbles and falling beach balls, economists use different assumptions when studying the short-run and long-run effects of a change in the quantity of money.

2-1c Economic Models

High school biology teachers teach basic anatomy with plastic replicas of the human body. These models have all the major organs—the heart, liver, kidneys, and so

on—and allow teachers to show their students very simply how the important parts of the body fit together. Because these plastic models are stylized and omit many details, no one would mistake one of them for a real person. Despite this lack of realism—indeed, because of it—studying these models is useful for learning how the human body works.

Economists also use models to learn about the world, but unlike plastic manikins, their models mostly consist of diagrams and equations. Like a biology teacher's plastic model, economic models omit many details to allow us to see what is truly important. Just as the biology teacher's model does not include all the body's muscles and blood vessels, an economist's model does not include every feature of the economy or every aspect of human behavior.

As we use models to examine various issues throughout this book, you will see that models are built with assumptions. Just as physicists begin the analysis of a falling marble by assuming away the existence of friction, economists assume away many details of the economy that are irrelevant to the question at hand. All models—in physics, biology, and economics—simplify reality to improve our understanding of it. And all models are subject to revision when the facts warrant it. The key is to find the right model at the right time. As the statistician George Box put it, "All models are wrong, but some are useful."

2-1d Our First Model: The Circular-Flow Diagram

The economy consists of millions of people engaged in many activities—buying, selling, working, hiring, manufacturing, and so on. To understand how the economy works, we must simplify our thinking about all these activities. In other words, we need a model that explains how the economy is organized and how participants in the economy interact with one another.

circular-flow diagram
a visual model of the economy that shows how dollars flow through markets among households and firms

Figure 1 presents a visual model of the economy called the **circular-flow diagram**. In this model, the economy includes only two types of decision makers—firms and households. Firms produce goods and services using inputs, such as labor, land, and capital (buildings and machines). These inputs are called the **factors of production**. Households own the factors of production and consume all the goods and services that the firms produce.

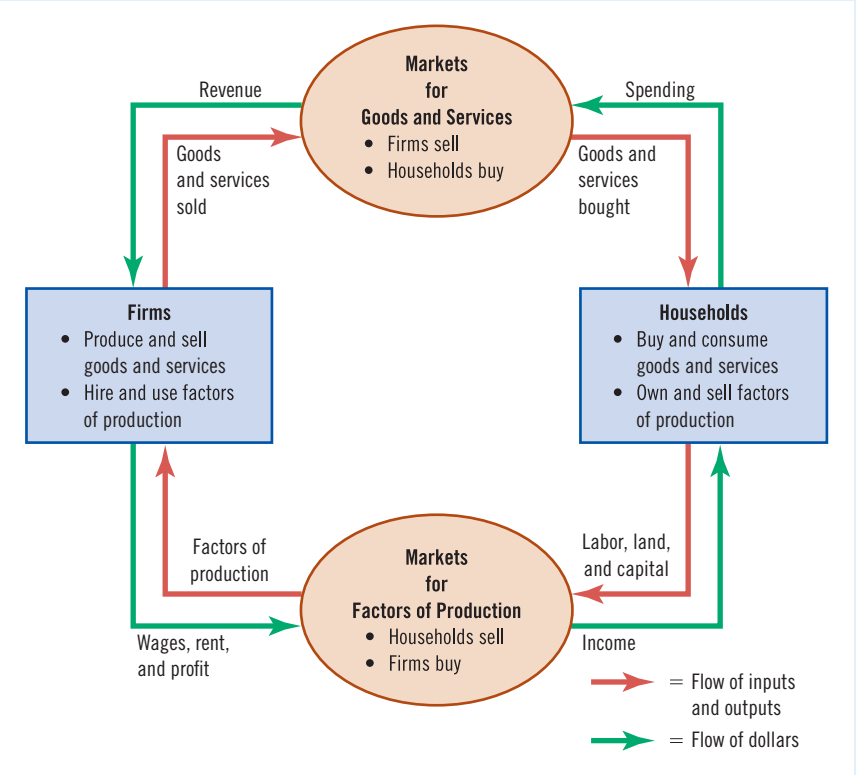
Households and firms interact in two types of markets. In the **markets for goods and services**, households are buyers, and firms are sellers. Specifically, households buy the output of goods and services that firms produce. In the **markets for the factors of production**, households are sellers, and firms are buyers. In these markets, households provide the inputs that firms use to produce goods and services. The circular-flow diagram offers a simple way of organizing all the transactions between households and firms in an economy.

The two loops of the circular-flow diagram are distinct but related. The inner loop represents the flows of inputs and outputs. Households sell the use of their labor, land, and capital to firms in the markets for the factors of production. Firms then use these factors to produce goods and services, which in turn are sold to households in the markets for goods and services. The outer loop of the diagram represents the corresponding flow of dollars. Households spend money to buy goods and services from firms. The firms use some of the revenue from these sales to purchase the factors of production, such as by paying workers' wages. What's left is the profit for the firm owners, who are themselves members of households.

Figure 1

The Circular Flow

This diagram is a schematic representation of the organization of the economy. Decisions are made by households and firms. Households and firms interact in the markets for goods and services (where households are buyers and firms are sellers) and in the markets for the factors of production (where firms are buyers and households are sellers). The outer set of arrows shows the flow of dollars, and the inner set of arrows shows the corresponding flow of inputs and outputs.



Let's take a tour of the circular flow by following a dollar bill as it makes its way from person to person through the economy. Imagine that the dollar begins at a household—say, in your wallet. If you want a cup of coffee, you take the dollar (along with a few of its relatives) to the market for coffee, which is one of the many markets for goods and services. When you buy your favorite drink at your local coffee shop, the dollar moves into the shop's cash register, becoming revenue for the firm. The dollar doesn't stay there for long, however, because the firm spends it on inputs in the markets for the factors of production. The coffee shop might use the dollar to pay rent to its landlord for the space it occupies or to compensate the baristas. Or it might return the dollar as profit to the shop's owner. In any case, the dollar enters the income of some household and, once again, is back in someone's wallet. At that point, the story of the economy's circular flow starts once again.

The circular-flow diagram in Figure 1 is a simple model of the economy. A more complex and realistic circular-flow model would include, for instance, the roles of government and international trade. (A portion of the dollar you gave to the coffee shop might be used to pay taxes or to buy coffee beans from a farmer in Kenya.) Yet these details are not crucial for a basic understanding of how the economy is organized. Because of its simplicity, this circular-flow diagram is useful to keep in mind when thinking about how the pieces of an economy fit together.

2-1e Our Second Model: The Production Possibilities Frontier

Most economic models, unlike the circular-flow diagram, are built using the tools of mathematics. Here, we use one of the simplest of these models, called the production possibilities frontier, to illustrate some basic economic ideas.

Although real economies produce thousands of goods and services, consider an economy that produces only two goods—cars and computers. Together, the car and computer industries use all of this economy's factors of production. The **production possibilities frontier** is a graph that shows the various combinations of output—in this case, cars and computers—the economy can possibly produce given the available factors of production and the production technology that firms use to turn these inputs into output.

production possibilities frontier a graph that shows the combinations of output that the economy can possibly produce with the available factors of production and production technology

Figure 2 shows this economy's production possibilities frontier. If the economy uses all its resources in the car industry, it produces 1,000 cars and no computers. If it uses all its resources in the computer industry, it produces 3,000 computers and no cars. The two endpoints of the production possibilities frontier represent these extreme possibilities.

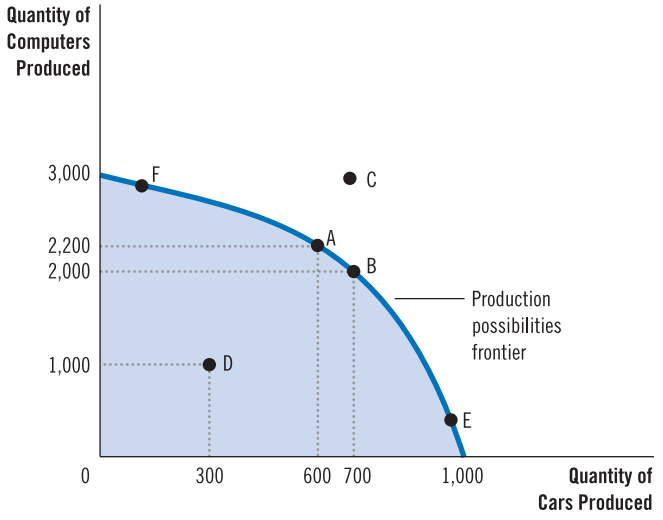
More likely, the economy divides its resources between the two industries, producing some cars and some computers. For example, it can produce 600 cars and 2,200 computers, as shown in the figure by point A. Or, by moving some of the factors of production to the car industry from the computer industry, the economy can produce 700 cars and 2,000 computers, represented by point B.

Because resources are scarce, not every conceivable outcome is feasible. For example, no matter how resources are allocated between the two industries, the economy cannot produce the number of cars and computers represented by point C. With the technology available for making cars and computers, the economy does not have enough of the factors of production to support that level of output. With the resources it has, the economy can produce at any point on or inside the production possibilities frontier, but it cannot produce at points outside the frontier.

Figure 2

The Production Possibilities Frontier

The production possibilities frontier shows the combinations of output—in this case, cars and computers—the economy can produce. Any point on or beneath the curve is a possible output combination in this economy. Points outside the frontier are not feasible given the economy's resources. The slope of the production possibilities frontier measures the opportunity cost of a car in terms of computers. This opportunity cost varies, depending on how much of the two goods the economy is producing.



An outcome is said to be **efficient** if the economy is getting all it can from the scarce resources it has available. Points on (rather than inside) the production possibilities frontier represent efficient levels of production. When the economy is producing at such a point, say, point A, there is no way to produce more of one good without producing less of the other. Point D represents an **inefficient** outcome. For some reason, perhaps widespread unemployment, the economy is producing less than it could from the resources it has available: It is producing only 300 cars and 1,000 computers. If the source of the inefficiency is eliminated, the economy can increase its production of both goods. For example, if the economy moves from point D to point A, its production of cars increases from 300 to 600, and its production of computers increases from 1,000 to 2,200.

One of the **Ten Principles of Economics** in Chapter 1 is that people face trade-offs. The production possibilities frontier shows one trade-off that society faces. Once an economy reaches an efficient point on the frontier, the only way to produce more of one good is to produce less of the other. When the economy moves from point A to point B, for instance, society produces 100 more cars at the expense of producing 200 fewer computers.

This trade-off helps us understand another of the **Ten Principles of Economics**: The cost of something is what you give up to get it. This is called the **opportunity cost**. The production possibilities frontier shows the opportunity cost of one good as measured in terms of the other. When society moves from point A to point B, it gives up 200 computers to get 100 additional cars. That is, at point A, the opportunity cost of 100 cars is 200 computers. Put another way, the opportunity cost of each car is two computers. Notice that the opportunity cost of a car equals the slope of the production possibilities frontier. (Slope is discussed in the graphing appendix to this chapter.)

The opportunity cost of a car in terms of the number of computers is not constant in this economy but depends on how many cars and computers the economy is producing. This is reflected in the shape of the production possibilities frontier. Because the production possibilities frontier in Figure 2 is bowed outward, the opportunity cost of a car is highest when the economy is producing many cars and few computers, such as at point E, where the frontier is steep. When the economy is producing few cars and many computers, such as at point F, the frontier is flatter, and the opportunity cost of a car is lower.

Economists believe that production possibilities frontiers often have this bowed-out shape. When the economy is using most of its resources to make computers, the resources best suited to car production, such as skilled autoworkers, are being used in the computer industry. Because these workers probably aren't very good at making computers, increasing car production by one unit will cause only a slight reduction in the number of computers produced. Thus, at point F, the opportunity cost of a car in terms of computers is small, and the frontier is relatively flat. By contrast, when the economy is using most of its resources to make cars, such as at point E, the resources best suited to making cars are already at work in the car industry. Producing an additional car now requires moving some of the best computer technicians out of the computer industry and turning them into autoworkers. As a result, producing an additional car requires a substantial loss of computer output. The opportunity cost of a car is high, and the frontier is steep.

The production possibilities frontier shows the trade-off between the outputs of different goods at a given time, but the trade-off can change over time. For example, suppose a technological advance in the computer industry raises the number of computers that a worker can produce per week. This advance expands society's set

of opportunities. For any given number of cars, the economy can now make more computers. If the economy does not produce any computers, it can still produce 1,000 cars, so one endpoint of the frontier stays the same. But if the economy devotes some of its resources to the computer industry, it will produce more computers from those resources. As a result, the production possibilities frontier shifts outward, as in Figure 3.

This figure shows what happens when an economy grows. Society can move production from a point on the old frontier to a point on the new one. Which point it chooses depends on its preferences for the two goods. In this example, society moves from point A to point G, producing more computers (2,300 instead of 2,200) and more cars (650 instead of 600).

The production possibilities frontier simplifies a complex economy to highlight some basic but powerful ideas: scarcity, efficiency, trade-offs, opportunity cost, and economic growth. As you study economics, these ideas will recur in various forms. The production possibilities frontier offers one simple way of thinking about them.

2-1f Microeconomics and Macroeconomics

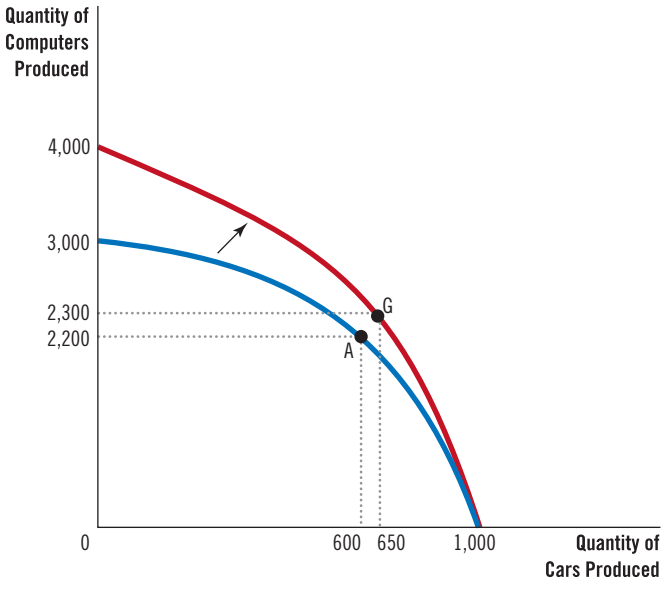
Many subjects are studied on various levels. Consider biology, for example. Molecular biologists study the chemical compounds that make up living things. Cellular biologists study cells, which are made up of chemical compounds and, at the same time, are themselves the building blocks of living organisms. Evolutionary biologists study the diversity of animals and plants and how species gradually change over the centuries.

Economics is also studied on various levels. We can examine the decisions of individual households and firms. We can focus on the interaction of households and firms in markets for specific goods and services. Or we can study the operation of the economy as a whole, encompassing all these activities in all these markets.

Figure 3

A Shift in the Production Possibilities Frontier

A technological advance in the computer industry enables the economy to produce more computers for any given number of cars. As a result, the production possibilities frontier shifts outward. If the economy moves from point A to point G, the production of both cars and computers increases.



Economics is traditionally divided into two broad subfields. **Microeconomics** is the study of how households and firms make decisions and how they interact in specific markets. **Macroeconomics** is the study of the overall economy. A microeconomist might study the effects of rent control on housing in New York City, the impact of foreign competition on the U.S. auto industry, or the effects of education on workers' earnings. A macroeconomist might study the effects of borrowing by the federal government, the changes in the economy's unemployment rate over time, or alternative policies to promote growth in national living standards.

microeconomics
the study of how households and firms make decisions and how they interact in markets

macroeconomics
the study of economy-wide phenomena, including inflation, unemployment, and economic growth

Microeconomics and macroeconomics are closely intertwined. Because changes in the overall economy arise from the decisions of millions of individuals, it is impossible to understand macroeconomic developments without considering the underlying microeconomic decisions. For example, a team of macroeconomists might study the effect of a federal income tax cut on the overall production of goods and services. But to analyze this issue, they must consider how the tax cut affects households' decisions about how much to spend on goods and services.

Despite the inherent link between microeconomics and macroeconomics, the two fields are distinct. Because they address different questions, each field has its own set of models, which are often taught in separate courses.

QuickQuiz

1. An economic model is
   a. a mechanical machine that replicates the functioning of the economy.
   b. a fully detailed, realistic description of the economy.
   c. a simplified representation of some aspect of the economy.
   d. a computer program that predicts the future of the economy.
2. The circular-flow diagram illustrates that, in markets for the factors of production,
   a. households are sellers, and firms are buyers.
   b. households are buyers, and firms are sellers.
   c. households and firms are both buyers.
   d. households and firms are both sellers.
3. A point inside the production possibilities frontier is
   a. efficient but not feasible.
   b. feasible but not efficient.
   c. both efficient and feasible.
   d. neither efficient nor feasible.
4. All of the following topics fall within the study of microeconomics EXCEPT
   a. the impact of cigarette taxes on the smoking behavior of teenagers.
   b. the role of Microsoft's market power in the pricing of software.
   c. the effectiveness of antipoverty programs in reducing homelessness.
   d. the influence of the government's budget deficit on economic growth.

Answers are at the end of the chapter.

2-2 The Economist as Policy Adviser

Often, economists are asked to explain the causes of economic events. Why, for example, is unemployment higher for teenagers than for older workers? That's a factual question, and it may be answered scientifically. But sometimes, economists are asked to recommend policies to improve economic outcomes. What, for instance, should the government do to enhance teenagers' well-being? Answering that requires not only an understanding of what's happening but also value judgments about what ought to be done.

In the News

Why Tech Companies Hire Economists

Many high-tech companies find that expertise in economics is useful in their decision making.

Goodbye, Ivory Tower. Hello, Silicon Valley Candy Store

By Steve Lohr

For eight years, Jack Coles had an economist's dream job at Harvard Business School.

His research focused on the design of efficient markets, an important and growing field that has influenced such things as Treasury bill auctions and decisions on who receives organ transplants. He even got to work with Alvin E. Roth, who won a Nobel in economic science in 2012.

But prestige was not enough to keep Mr. Coles at Harvard. In 2013, he moved to the San Francisco Bay Area. He now works at Airbnb, the online lodging marketplace, one of a number of tech companies luring economists with the promise of big sets of data and big salaries.

Silicon Valley is turning to the dismal science in its never-ending quest to squeeze more money out of old markets and build new ones. In turn, the economists say they are eager to explore the digital world for fresh insights into timeless economic questions of pricing, incentives and behavior.

"It's an absolute candy store for economists," Mr. Coles said. . . .

Businesses have been hiring economists for years. Usually, they are asked to study macroeconomic trends—topics like recessions and currency exchange rates—and help their employers deal with them.

But what the tech economists are doing is different: Instead of thinking about national or global trends, they are studying the data trails of consumer behavior to help digital companies make smart decisions that strengthen their online marketplaces in areas like advertising, movies, music, travel and lodging.

Tech outfits including giants like Amazon, Facebook, Google and Microsoft and up-and-comers like Airbnb and Uber hope that sort of improved efficiency means more profit.

At Netflix, Randall Lewis, an economic research scientist, is finely measuring the effectiveness of advertising. His work also gets at the correlation-or-causation conundrum in economic behavior: What consumer actions occur coincidentally after people see ads, and what actions are most likely caused by the ads?

At Airbnb, Mr. Coles is researching the company's marketplace of hosts and guests

When economists are trying to explain the world, they are scientists. When they are giving guidance on how to improve it, they are policy advisers. Even if you never become a professional economist, you may find yourself using both sides of your economic brain in daily life: analyzing the world as you find it and devising solutions to make things better. Both approaches are indispensable, but it's important to understand how they differ.

2-2a Positive versus Normative Analysis

To clarify the two roles that economists play, let's examine the use of language. Because scientists and policy advisers have different goals, they use language in different ways.

For example, suppose that two people are discussing minimum-wage laws. Here are two statements you might hear:

Prisha: Minimum-wage laws cause unemployment.
Noah: The government should raise the minimum wage.

Ignoring for now whether you agree with these claims, notice that Prisha and Noah differ in what they are trying to do. Prisha is speaking like a scientist: She is describing how the world works. Noah is speaking like a policy adviser: He is talking about how he would like to change the world.

for insights, both to help build the business and to understand behavior. One study focuses on procrastination—a subject of great interest to behavioral economists—by looking at bookings. Are they last-minute? Made weeks or months in advance? Do booking habits change by age, gender or country of origin?

"They are microeconomic experts, heavy on data and computing tools like machine learning and writing algorithms," said Tom Beers, executive director of the National Association for Business Economics.

Understanding how digital markets work is getting a lot of attention now, said Hal Varian, Google's chief economist. But, he said, "I thought it was fascinating years ago."

Mr. Varian, 69, is the godfather of the tech industry's in-house economists. Once a well-known professor at the University of California, Berkeley, Mr. Varian showed up at Google in 2002, part time at first, but soon became an employee. He helped refine Google's AdWords marketplace, where advertisers bid to have their ads shown on search pages….

For the moment, Amazon seems to be the most aggressive recruiter of economists. It even has an Amazon Economists website for soliciting résumés. In a video on the site, Patrick Bajari, the company's chief economist, says the economics team has contributed to decisions that have had "multibillion-dollar impacts" for the company….

A current market-design challenge for Amazon and Microsoft is their big cloud computing services. These digital services, for example, face a peak-load problem, much as electric utilities do.

How do you sell service at times when there is a risk some customers may be bumped off? Run an auction for what customers are willing to pay for interruptible service? Or offer set discounts for different levels of risk? Both Amazon and Microsoft are working on that now.

To answer such questions, economists work in teams with computer scientists and people in business. In tech companies, market design involves not only economics but also engineering and marketing. How hard is a certain approach technically? How easy is it to explain to customers?

"Economics influences rather than determines decisions," said Preston McAfee, Microsoft's chief economist, who previously worked at Google and Yahoo. ■

Questions to Discuss

1. Think of some firms that you often interact with. How might the input of economists improve their businesses?
2. After studying economics in college, what kind of businesses would be the most fun to work for?

Source: Steve Lohr, "Goodbye, Ivory Tower. Hello, Silicon Valley Candy Store," New York Times, September 4, 2016.

In general, statements about the world come in two types. One type, such as Prisha's, is known as positive. That doesn't mean it is necessarily upbeat or optimistic. **Positive statements** are descriptive. They make a claim about how the world **is**. A second type of statement, such as Noah's, is known as normative. **Normative statements** are prescriptive. They make a claim about how the world **ought to be**.

positive statements claims that attempt to describe the world as it is

normative statements claims that attempt to prescribe how the world should be

A key difference between positive and normative statements is how we judge their validity. We can, in principle, confirm or refute positive statements by examining evidence. An economist might try to evaluate Prisha's statement by analyzing data on changes in minimum wages and unemployment over time. Proving causality can be difficult, as we discuss later, but at its root, the issue should be determined by the evidence. By contrast, evaluating normative statements involves values as well as facts. Noah's statement cannot be judged with data alone. Deciding what is good or bad policy is not just a matter of science. It also involves values, and views on ethics, religion, and political philosophy may well come into play.

Positive and normative statements are different but often intertwined. In particular, positive findings about how the world works can easily affect normative judgments about what policies are desirable. Prisha's claim that the minimum wage causes unemployment, if true, might lead her to reject Noah's conclusion that the government should raise the minimum wage. On the other hand, a positive finding that the unemployment effect is small might lead her to accept Noah's policy prescription.

Normative judgments may also influence the positive claims that researchers choose to study. Noah's desire to raise the minimum wage, for example, may lead him to investigate Prisha's claim that it causes unemployment. To be free of bias, he should put aside his normative views and examine the data as objectively as possible. At its best, positive economics proceeds as a science independent of the researcher's personal values or policy agenda.

As you study economics, keep in mind the broad distinction between positive and normative statements because it will help you stay focused on the task at hand. Much of economics is positive: It just tries to explain how the economy works. Yet those who use economics often have normative goals: They want to learn how to improve the economy. When you hear economists making normative statements, you know they are speaking not as scientists but as policy advisers.

2-2b Economists in Washington

President Harry Truman once said that he wanted to find a one-armed economist. When he asked his economists for advice, they always answered, "On the one hand, ... On the other hand, ..."

Truman was right that economists' advice is not always straightforward. This tendency is rooted in one of the **Ten Principles of Economics**: People face trade-offs. Economists are aware that trade-offs are involved in most policy decisions. A policy might increase efficiency at the cost of equality. It might help future generations but hurt the current generation. An economist who says that all policy decisions are easy is an economist not to be trusted.

JAMES STEVENSON/ THE NEW YORKER COLLECTION/THE CARTOON BANK

"Let's switch. I'll make the policy, you implement it, and he'll explain it."

Truman was not the only president who relied on economists' advice. Since 1946, the president of the United States has received guidance from the Council of Economic Advisers, which consists of three members and a staff of a few dozen economists. The council, whose offices are just a few steps from the White House, advises the president and writes the annual *Economic Report of the President*, which discusses recent economic developments and presents the council's analysis of current policy issues. (The author of this textbook was chair of the Council of Economic Advisers from 2003 to 2005.)

The president also receives information and advice from economists in many administrative departments. Economists at the Office of Management and Budget help formulate spending plans and regulatory policies. Economists at the Department of the Treasury help design tax policy. Economists at the Department of Labor analyze data on workers and those looking for work to help formulate labor-market policies. Economists at the Department of Justice help enforce the nation's antitrust laws.

Economists in the federal government are also found outside the executive branch. To obtain independent evaluations of policy proposals, Congress relies on the advice of the Congressional Budget Office, which is staffed by economists. The Federal Reserve, the institution that sets the nation's monetary policy, employs hundreds of economists to analyze developments in the United States and around the world.

The influence of economists on policy goes beyond their role as advisers: Their research and writings can affect policy indirectly. The economist John Maynard Keynes offered this observation:

> The ideas of economists and political philosophers, both when they are right and when they are wrong, are more powerful than is commonly understood.

> Indeed, the world is ruled by little else. Practical men, who believe themselves to be quite exempt from intellectual influences, are usually the slaves of some defunct economist. Madmen in authority, who hear voices in the air, are distilling their frenzy from some academic scribbler of a few years back.

These words were written in 1935, but they remain true today. Indeed, the "academic scribbler" now influencing public policy is often Keynes himself.

2-2c Why Economists' Advice Is Often Not Followed

Economists who advise presidents and other elected leaders know that their recommendations are not always heeded. It is easy to understand why. The process by which economic policy is actually made differs in many ways from the idealized policy process assumed in textbooks.

Throughout this text, whenever we discuss policy, we often focus on one question: What is the best policy for the government to pursue? We act as if policy were set by a benevolent and omnipotent king. After the king figures out the right policy, he has no trouble putting his ideas into action.

In the real world, figuring out the right policy is only part of a leader's job, sometimes the easiest part. Imagine that you're the president. After hearing from your economic advisers what policy they deem best, you turn to others. Your communications advisers will tell you how best to explain the proposed policy to the public, and they will try to anticipate any misunderstandings that might make the challenge more difficult. Your press advisers will tell you how the news media will report on your proposal, what opinions will likely be expressed on editorial pages, and what memes may emerge in social media. Your legislative affairs advisers will tell you how Congress will view the proposal, what amendments members of Congress will suggest, and the likelihood that Congress will enact some version of your proposal. Your political advisers will tell you which groups will organize to support or oppose the proposed policy, how this proposal will affect your standing among different groups in the electorate, and whether it will change support for other policy initiatives. After weighing all this advice, you then decide how to proceed. (And even this picture is idealized; not all recent presidents have proceeded in this systematic fashion.)

Making economic policy in a representative democracy is a messy affair, and there are often good reasons why presidents (and other politicians) do not embrace the policies that economists advocate. The advice of economists is only one ingredient of a complex recipe.

QuickQuiz

5. Which of the following is a positive, rather than a normative, statement?
   a. Law X will reduce national income.
   b. Law X is a good piece of legislation.
   c. Congress ought to pass law X.
   d. The president should veto law X.

6. The following parts of government regularly rely on the advice of economists:
   a. Department of the Treasury.
   b. Office of Management and Budget.
   c. Department of Justice.
   d. All of the above.

Answers are at the end of the chapter.

2-3 Why Economists Disagree

"If all the economists were laid end to end, they would not reach a conclusion." This quip from George Bernard Shaw is revealing. Economists as a group are often criticized for giving conflicting advice to policymakers. President Ronald Reagan once joked that if the game Trivial Pursuit were designed for economists, it would have 100 questions and 3,000 answers.

Why do economists so often appear to give conflicting advice to policymakers? There are two basic reasons:

- Economists may disagree about the validity of alternative positive theories of how the world works.
- Economists may have different values and, therefore, different normative views about what government policy should aim to accomplish.

Let's look more closely at these reasons.

2-3a Differences in Scientific Judgments

Several centuries ago, astronomers debated whether the earth or the sun was at the center of the solar system. More recently, climatologists have debated whether the earth is experiencing global warming and, if so, why. Science is an ongoing search to understand the world around us. It is not surprising that as the search continues, scientists sometimes disagree about the direction in which truth lies.

Economists often disagree for the same reason. Although the field of economics sheds light on much about the world (as you will see throughout this book), there is still much to be learned. Sometimes, economists disagree because they have different hunches about the validity of alternative theories. Sometimes, they disagree because of different judgments about the size of the parameters that measure how economic variables are related.

For example, economists debate whether the government should tax a household's income or its consumption (spending). Advocates of a switch from the current income tax to a consumption tax believe that the change would encourage households to save more because the income that is saved would not be taxed. Higher saving, in turn, would free resources for capital accumulation, leading to more rapid growth in productivity and living standards. Advocates of the current income tax system believe that households would not alter their saving significantly in response to a change in the tax laws. These two groups of economists hold different normative views about the tax system because they have different positive views about how much saving responds to tax incentives.

2-3b Differences in Values

Suppose that Jack and Jill both take the same amount of water from the town well. The town taxes its residents to pay for maintaining the well. Jill has an income of $150,000 and is taxed $15,000, or 10 percent of her income. Jack has an income of $40,000 and is taxed $6,000, or 15 percent of his income.

Is this policy fair? If not, who pays too much, and who pays too little? Does it matter whether Jack's low income is due to a medical disability or to his decision to pursue an acting career? Does it matter whether Jill's high income is due to a large inheritance or to her willingness to work long hours at a dreary job?

These are difficult questions about which people are likely to disagree. If the town hired two experts to study how it should tax its residents to pay for the well, it would not be surprising if they offered conflicting advice.

This simple example shows why economists sometimes disagree about public policy. As we know from our discussion of normative and positive analysis, policies cannot be judged on scientific grounds alone. Sometimes, economists give conflicting advice because they have different values or political philosophies. Perfecting the science of economics will not tell us whether Jack or Jill pays too much.

2-3c Perception versus Reality

Because of differences in scientific judgments and differences in values, some disagreement among economists is inevitable. Yet one should not overstate the amount of disagreement. Economists agree with one another more often than is sometimes understood.

Consider the proposition, "A ceiling on rents reduces the quantity and quality of housing available." When economists were polled about it, 93 percent agreed. Economists believe that rent control—a policy that sets a legal maximum on the amount landlords can charge for their apartments—adversely affects the supply of housing and is a costly way of helping the neediest members of society. Nonetheless, many city governments ignore economists' advice and place ceilings on the rents that landlords may charge their tenants.

Similarly, consider the proposition, "Tariffs and import quotas usually reduce general economic welfare." Once again, 93 percent of economists agreed with the claim. Economists oppose tariffs (taxes on imports) and import quotas (limits on how much of a good can be purchased from abroad) because these policies impede the specialization that raises living standards both at home and abroad. Nonetheless, over the years, presidents and Congress have often chosen to restrict the import of certain goods.

Why do policies such as rent control and trade barriers persist if the experts are united in their opposition? It may be that the realities of the political process stand as immovable obstacles. But it also may be that economists have not yet convinced enough of the public that these policies are undesirable. One purpose of this book is to help you understand the economist's view on these and other subjects and, perhaps, to persuade you that it is the right one.

As you read this book, you will occasionally see small boxes called "Ask the Experts." These are based on the IGM Economic Experts Panel, an ongoing survey of several dozen prominent economists. Every few weeks, these experts are offered a proposition and then asked whether they agree with it, disagree with it, or are uncertain. The results in these boxes will give you a sense of when economists are united, when they are divided, and when they just don't know what to think.

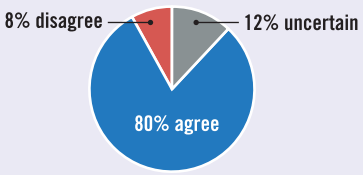
You can see an example here regarding the resale of tickets to entertainment and sporting events. Lawmakers sometimes try to prohibit reselling tickets, or "scalping" as it is sometimes called. The survey results show that many economists side with the scalpers rather than the lawmakers.

Ask the Experts

Ticket Resale

"Laws that limit the resale of tickets for entertainment and sports events make potential audience members for those events worse off on average."

What do economists say?



Source: IGM Economic Experts Panel, April 16, 2012.

QuickQuiz

7. Economists may disagree because they have different
   a. hunches about the validity of alternative theories.
   b. judgments about the size of key parameters.
   c. political philosophies about the goals of public policy.
   d. All of the above.

8. Most economists believe that tariffs are
   a. a good way to promote domestic economic growth.
   b. a poor way to raise general economic well-being.
   c. an often necessary response to foreign competition.
   d. an efficient way for the government to raise revenue.

Answers are at the end of the chapter.

2-4 Let's Get Going

The first two chapters of this book have introduced you to the ideas and methods of economics. We are now ready to get to work. In the next chapter, we start learning in more detail the principles of economic behavior and policy.

As you proceed through this book, you will be asked to draw on many intellectual skills. You might find it helpful to keep in mind some advice from the great economist John Maynard Keynes:

> The study of economics does not seem to require any specialized gifts of an unusually high order. Is it not . . . a very easy subject compared with the higher branches of philosophy or pure science? An easy subject, at which very few excel! The paradox finds its explanation, perhaps, in that the master-economist must possess a rare **combination** of gifts. He must be mathematician, historian, statesman, philosopher—in some degree. He must understand symbols and speak in words. He must contemplate the particular in terms of the general, and touch abstract and concrete in the same flight of thought. He must study the present in the light of the past for the purposes of the future. No part of man's nature or his institutions must lie entirely outside his regard. He must be purposeful and disinterested in a simultaneous mood; as aloof and incorruptible as an artist, yet sometimes as near the earth as a politician.

This is a tall order. But with practice, you will become more and more accustomed to thinking like an economist.

Chapter in a Nutshell

- Economists try to address their subject with a scientist's objectivity. Like all scientists, they make appropriate assumptions and build simplified models to understand the world around them. Two simple economic models are the circular-flow diagram and the production possibilities frontier. The circular-flow diagram shows how households and firms interact in markets for goods and services and in markets for the factors of production. The production possibilities frontier shows how society faces a trade-off between producing different goods.
- The field of economics is divided into two subfields: microeconomics and macroeconomics. Microeconomists study decision making by households and firms and the interactions among households and firms in the marketplace. Macroeconomists study the forces and trends that affect the economy as a whole.

- A positive statement is an assertion about how the world **is**. A normative statement is an assertion about how the world **ought to be**. While positive statements can be judged based on facts and the scientific method, normative statements entail value judgments as well. When economists make normative statements, they are acting more as policy advisers than as scientists.
- Economists who advise policymakers sometimes offer conflicting advice either because of differences in scientific judgments or because of differences in values. At other times, economists are united in the advice they offer, but policymakers may choose to ignore the advice because of the many forces and constraints imposed on them by the political process.

Key Concepts

circular-flow diagram, p. 20
production possibilities frontier, p. 22
microeconomics, p. 25
macroeconomics, p. 25
positive statements, p. 27
normative statements, p. 27

Questions for Review

1. In what ways is economics a science?
2. Why do economists make assumptions?
3. Should an economic model exactly describe reality?
4. Name a way that your family interacts in the markets for the factors of production and a way that it interacts in the markets for goods and services.
5. Name one economic interaction that isn't covered by the simplified circular-flow diagram.
6. Draw and explain a production possibilities frontier for an economy that produces milk and cookies. What happens to this frontier if a disease kills half of the economy's cows?
7. Use a production possibilities frontier to describe the idea of **efficiency**.
8. What are the two subfields of economics? Explain what each subfield studies.
9. What is the difference between a positive and a normative statement? Give an example of each.
10. Why do economists sometimes offer conflicting advice to policymakers?

Problems and Applications

1. Draw a circular-flow diagram. Identify the parts of the model that correspond to the flow of goods and services and the flow of dollars for each of the following activities.
   a. Selena pays a storekeeper $1 for a quart of milk.
   b. Stuart earns $8 per hour working at a fast-food restaurant.
   c. Shanna spends $40 to get a haircut.
   d. Salma earns $20,000 from her 10 percent ownership of Acme Industrial.
2. Imagine a society that produces military goods and consumer goods, which we'll call "guns" and "butter."
   a. Draw a production possibilities frontier for guns and butter. Using the concept of opportunity cost, explain why it most likely has a bowed-out shape.
   b. Show a point on the graph that is impossible for the economy to achieve. Show a point on the graph that is feasible but inefficient.
   c. Imagine that the society has two political parties, called the Hawks (who want a strong military) and the Doves (who want a smaller military). Show a point on your production possibilities frontier that the Hawks might choose and a point that the Doves might choose.
   d. Imagine that an aggressive neighboring country reduces the size of its military. As a result, both the Hawks and the Doves reduce their desired production of guns by the same amount. Which party would get the bigger "peace dividend," measured by the increase in butter production? Explain.
3. The first principle of economics in Chapter 1 is that people face trade-offs. Use a production possibilities frontier to illustrate society's trade-off between two "goods"—a clean environment and the quantity of industrial output. What do you suppose determines the shape and position of the frontier? Show what will happen to the frontier if engineers develop a new way of producing electricity that emits fewer pollutants.

4. An economy consists of three workers: Larry, Moe, and Curly. Each works 10 hours a day and can produce two services: mowing lawns and washing cars. In an hour, Larry can either mow one lawn or wash one car, Moe can either mow one lawn or wash two cars, and Curly can either mow two lawns or wash one car.
   a. Calculate how much of each service is produced in the following scenarios, which we label A, B, C, and D:
      - All three spend all their time mowing lawns. (A)
      - All three spend all their time washing cars. (B)
      - All three spend half their time on each activity. (C)
      - Larry spends half his time on each activity, while Moe only washes cars and Curly only mows lawns. (D)
   b. Graph the production possibilities frontier for this economy. Using your answers to part *a*, identify points A, B, C, and D on your graph.
   c. Explain why the production possibilities frontier has the shape it does.
   d. Are any of the allocations calculated in part *a* inefficient? Explain.
5. Classify each of the following topics as relating to microeconomics or macroeconomics.
   a. a family's decision about how much income to save
   b. the effect of government regulations on auto emissions
   c. the impact of higher national saving on economic growth
   d. a firm's decision about how many workers to hire
   e. the relationship between the inflation rate and changes in the quantity of money
6. Classify each of the following statements as positive or normative. Explain.
   a. Society faces a short-run trade-off between inflation and unemployment.
   b. A reduction in the growth rate of the money supply will reduce the rate of inflation.
   c. The Federal Reserve should reduce the growth rate of the money supply.
   d. Society ought to require welfare recipients to look for jobs.
   e. Lower tax rates encourage more work and more saving.

QuickQuiz Answers

1. **c** 2. **a** 3. **b** 4. **d** 5. **a** 6. **d** 7. **d** 8. **b**

Appendix

Graphing: A Brief Review

Many economic concepts can be expressed with numbers—the price of bananas, the quantity of bananas sold, the cost of growing bananas, and so on. Often, these variables are related to one another: When the price of bananas rises, people buy fewer bananas. One way to express these relationships is with graphs.

Graphs serve two purposes. First, when developing theories, they offer a visual way to express ideas that might be less clear if described with equations or words. Second, when analyzing data, graphs provide a powerful way of finding and interpreting patterns. In either case, graphs provide a lens through which a recognizable forest emerges from a multitude of trees.

Numerical information can be expressed graphically in many ways, just as there are many ways to express a thought in words. A good writer chooses words that will make an argument clear, a description pleasing, or a scene dramatic. An effective economist chooses the type of graph that best suits the purpose at hand.

This appendix discusses how economists use graphs to study the mathematical relationships among variables. It also points out some of the pitfalls that can arise when using graphical methods.

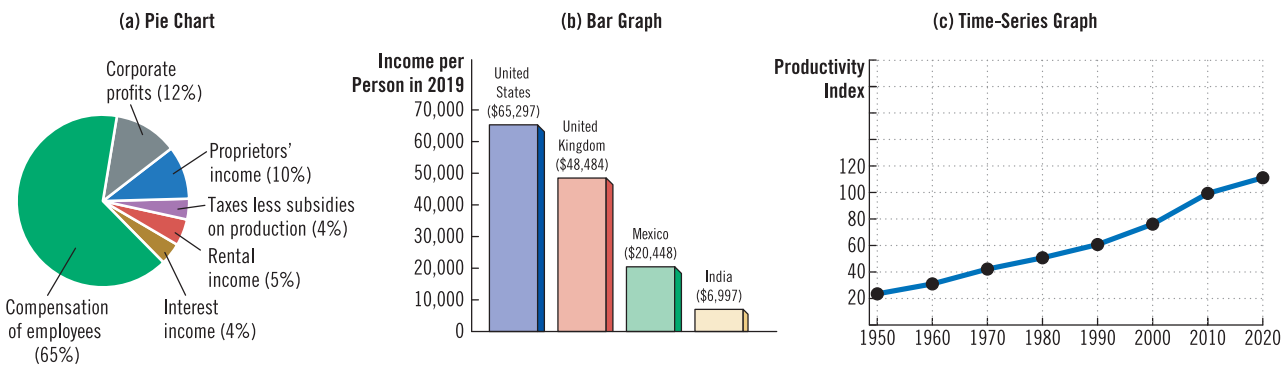
Graphs of a Single Variable

Three common graphs appear in Figure A-1. The **pie chart** in panel (a) shows how total income in the United States is divided among the sources of income, including

Figure A-1

Types of Graphs

The pie chart in panel (a) shows how U.S. national income in 2020 was derived from different sources. The bar graph in panel (b) compares the average income in four countries. The time-series graph in panel (c) shows labor productivity in U.S. businesses over time.



Source: U.S. Department of Commerce, The World Bank.

compensation of employees, corporate profits, and so on. A slice of the pie represents each source's share of the total. The **bar graph** in panel (b) compares income in four countries. The height of each bar represents the average income in each country. The **time-series graph** in panel (c) traces the rising productivity in the U.S. business sector over time. The height of the line shows output per hour in each year. You have probably seen similar graphs in news reports.

Graphs of Two Variables: The Coordinate System

The three graphs in Figure A-1 are useful, but they are limited in how much they can tell us. These graphs display information about only a single variable. If economists are looking at the relationships between variables, they may want to display two variables on a single graph. The **coordinate system** makes this possible.

Suppose you want to examine the relationship between study time and grade point average. For each student in a class, you could record a pair of numbers: study hours per week and grade point average. These numbers could then be placed in parentheses as an **ordered pair** and appear as a single point on the graph. Albert E., for instance, is represented by the ordered pair (25 hours/week, 3.5 GPA), while his "what-me-worry?" classmate Alfred E. is represented by the ordered pair (5 hours/week, 2.0 GPA).

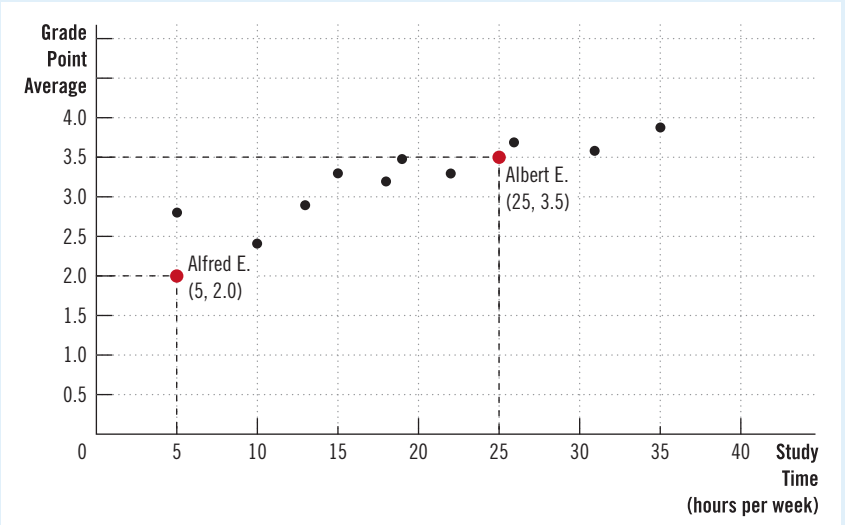
We can graph these ordered pairs on a two-dimensional grid. The first number in each ordered pair, called the **x-coordinate**, tells us the horizontal location of the point. The second number, called the **y-coordinate**, tells us the vertical location. The point with both an x-coordinate and a y-coordinate of zero is called the **origin**. The two coordinates in the ordered pair tell us where the point is located in relation to the origin: x units to the right of the origin and y units above it.

Figure A-2 graphs grade point average against study time for Albert E., Alfred E., and their classmates. This type of graph is called a **scatter plot** because it plots scattered points. Looking at the graph, notice that points farther to the right (indicating more study time) also tend to be higher (indicating a better grade point average). Because study time and grade point average typically move in the same

Figure A-2

Using the Coordinate System

Grade point average is measured on the vertical axis and study time on the horizontal axis. Albert E., Alfred E., and their classmates are represented by various points. The graph shows that students who study more tend to get higher grades.



direction, we say that these two variables have a **positive correlation**. By contrast, if we were to graph party time and grades, we would likely find that higher party time is associated with lower grades. Because these variables typically move in opposite directions, we say that they have a **negative correlation**. In either case, the coordinate system makes the correlation between two variables easy to see.

Curves in the Coordinate System

Students who study more do tend to get higher grades, but other factors also influence a student's grades. Previous preparation is an important factor, for instance, as is talent, attention from teachers, or even eating a good breakfast. A scatter plot like Figure A-2 does not attempt to isolate the effect that studying has on grades from the effects of other variables. Often, however, economists prefer looking at how one variable affects another, holding all other possible variables constant.

To see how this is done, consider one of the most important graphs in economics: the **demand curve**. The demand curve traces the effect of a good's price on the quantity that consumers want to buy. Before showing a demand curve, however, consider Table A-1, which shows how the number of novels that Emma buys depends on her income and on the price of novels. When novels are cheap, Emma buys a lot of them. As they become more expensive, she instead borrows books from the library or goes to the movies rather than read. Similarly, at any price, Emma buys more novels when she has a higher income. That is, when her income increases, she spends part of the additional income on novels and part on other goods.

We now have three variables—the price of novels, income, and the number of novels purchased—which is more than can be shown in two dimensions. To put the information from Table A-1 in graphical form, we need to hold one of the three variables constant and trace out the relationship between the other two. Because the demand curve represents the relationship between price and quantity demanded, we hold Emma's income constant and show how the number of novels she buys varies with the price of novels.

Suppose that Emma's income is $40,000 per year. If we place the number of novels Emma buys on the x-axis and the price of novels on the y-axis, we can graphically represent the middle column of Table A-1. When the points that represent these

Table A-1

Novels Purchased by Emma

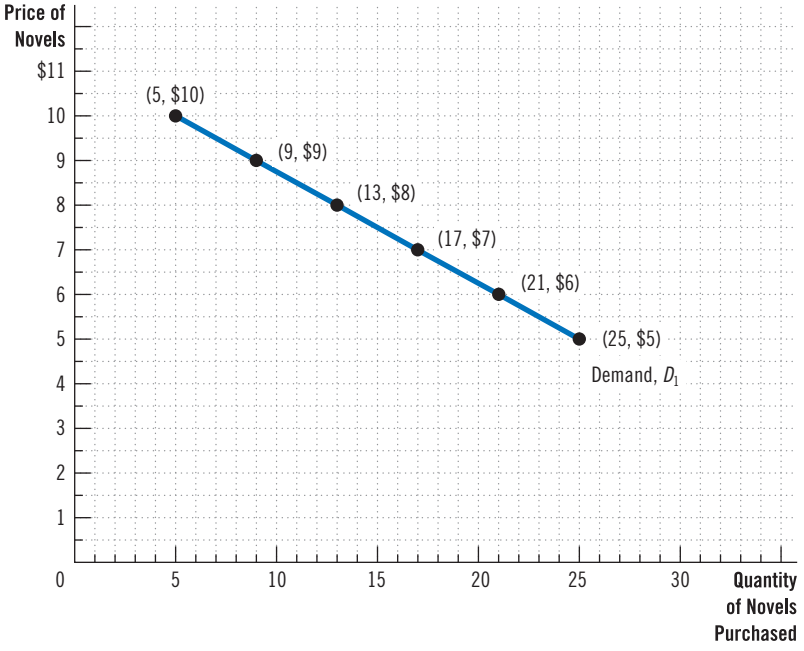
This table shows the number of novels Emma buys at various incomes and prices. For any given level of income, the data on price and quantity demanded can be graphed to produce Emma's demand curve for novels, as shown in Figures A-3 and A-4.

Price	For $30,000 Income:	For $40,000 Income:	For $50,000 Income:
$10	2 novels	5 novels	8 novels
9	6	9	12
8	10	13	16
7	14	17	20
6	18	21	24
5	22	25	28
	Demand curve, D_3	Demand curve, D_1	Demand curve, D_2

Figure A-3

Demand Curve

The line D_1 shows how Emma's purchases of novels depend on the price of novels when her income is held constant. Because the price and the quantity demanded are negatively related, the demand curve slopes down.



entries from the table—(5 novels, $10), (9 novels, $9), and so on—are connected, they form a line. This line, pictured in Figure A-3, is known as Emma's demand curve for novels; it tells us how many novels Emma buys at any price, holding income constant. The demand curve slopes down, indicating that a lower price increases the quantity of novels demanded. Because the quantity of novels demanded and the price move in opposite directions, we say that the two variables are **negatively related**. (Conversely, when two variables move in the same direction, the curve relating them slopes up, and we say that the variables are **positively related**.)

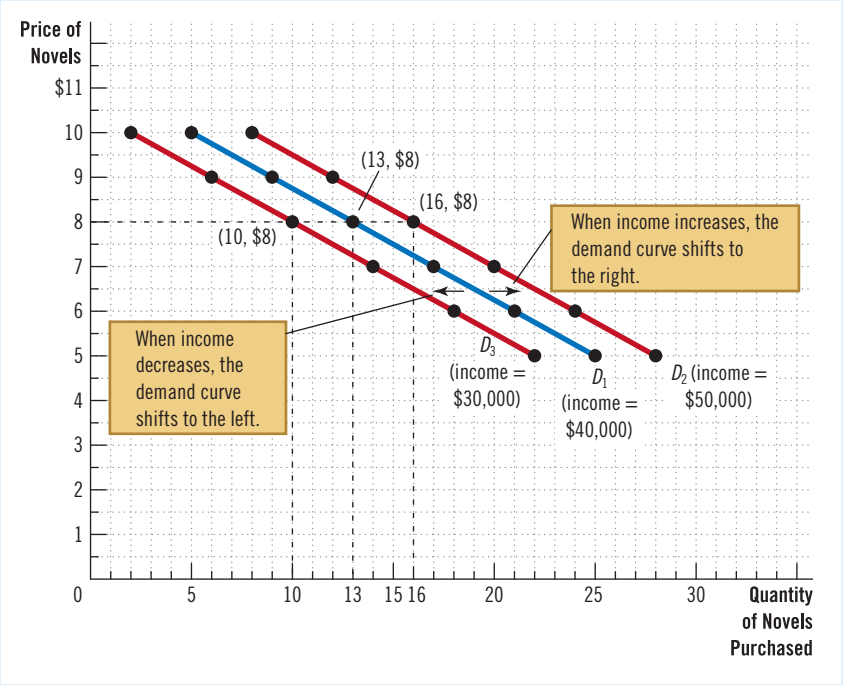
Now suppose Emma's income rises to $50,000 per year. At any price, Emma buys more novels than she did at her previous income. Just as we earlier drew Emma's demand curve for novels using the entries from the middle column of Table A-1, we now draw a new demand curve using the entries from the right column of the table. This new demand curve (curve D_2) is pictured alongside the old one (curve D_1) in Figure A-4; the new curve is a similar line drawn farther to the right. We therefore say that Emma's demand curve for novels **shifts** to the right when her income increases. Likewise, if Emma's income were to fall to $30,000 per year, she would buy fewer novels at any price, and her demand curve would shift to the left (to curve D_3).

In economics, it is important to distinguish between **movements along a curve** and **shifts of a curve**. As Figure A-3 shows, if Emma earns $40,000 per year and each novel costs $8, she buys 13 novels per year. If the price of novels falls to $7, Emma increases her purchases to 17 novels per year. The demand curve, however, stays fixed in the same place. Emma still buys the same number of novels **at each price**, but as the price falls, she moves along her demand curve from left to right.

Figure A-4

Shifting Demand Curves

The location of Emma's demand curve for novels depends on how much income she earns. The more she earns, the more novels she buys at any price, and the farther to the right her demand curve lies. Curve D_1 represents Emma's original demand curve, based on an income of $40,000 per year. If her income rises to $50,000 per year, her demand curve shifts to D_2. If her income falls to $30,000 per year, her demand curve shifts to D_3.



By contrast, if the price of novels remains fixed at $8 but her income rises to $50,000, Emma increases her purchases of novels from 13 to 16 per year. Because Emma buys more novels **at each price**, her demand curve shifts out, as shown in Figure A-4.

There is a simple way to tell when it is necessary to shift a curve: **When a relevant variable that is not named on either axis changes, the curve shifts**. Income is on neither the *x*-axis nor the *y*-axis of the graph, so when Emma's income changes, her demand curve shifts. The same is true for any change that affects Emma's purchasing habits, with the sole exception of a change in the price of novels. If, for instance, the public library closes and Emma must buy all the books she wants to read, she will demand more novels at each price, and her demand curve will shift to the right. Or, if the price of movies falls and Emma spends more time watching them and less time reading books, she will demand fewer novels at each price, and her demand curve will shift to the left. By contrast, when a variable on an axis of the graph changes, the curve does not shift. We read the change as a movement along the curve.

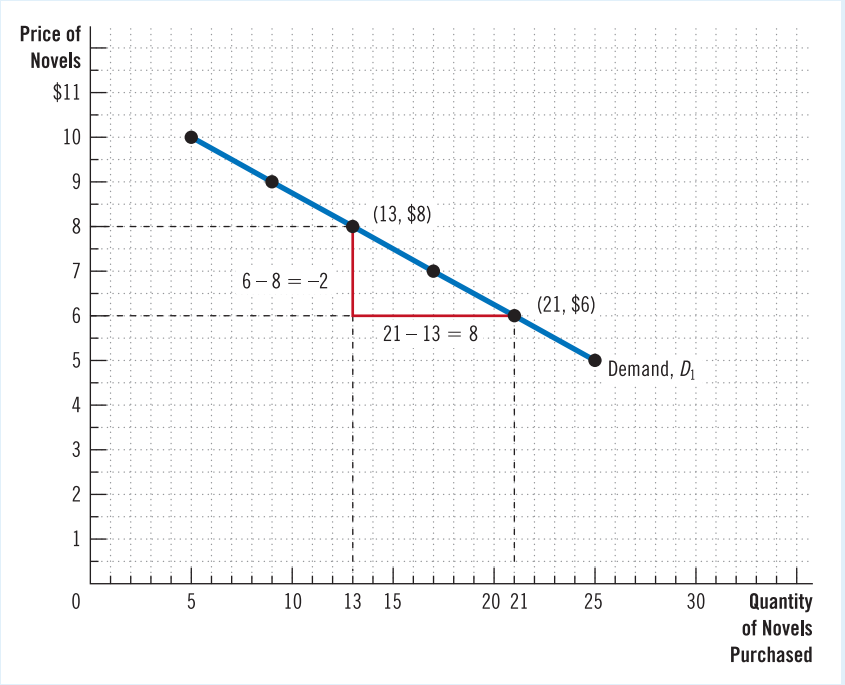
Slope

One question we might want to ask about Emma is how much her purchasing habits respond to changes in price. Look at the demand curve pictured in Figure A-5. If this curve is very steep, Emma buys nearly the same number of novels whether they are cheap or expensive. If the curve is much flatter, the number of novels she buys is more sensitive to price changes. To answer questions about how much one variable responds to changes in another, we can use the concept of **slope**.

Figure A-5

Calculating the Slope of a Line

To calculate the slope of the demand curve, look at the changes in the *x*- and *y*-coordinates as we move from the point (13 novels, $8) to the point (21 novels, $6). The slope of the line is the ratio of the change in the *y*-coordinate (–2) to the change in the *x*-coordinate (+8), which equals –¼.



The slope of a line is the ratio of the vertical distance covered to the horizontal distance covered as we move along the line. This definition is usually written in mathematical symbols as follows:

$$\text{slope} = \frac{\Delta y}{\Delta x},$$

where the Greek letter Δ (delta) stands for the change in a variable. In other words, the slope of a line is equal to the "rise" (change in y) divided by the "run" (change in x).

For an upward-sloping line, the slope is a positive number because the changes in x and y move in the same direction: If x increases, so does y, and if x decreases, so does y. For a fairly flat upward-sloping line, the slope is a small positive number. For a steep upward-sloping line, the slope is a large positive number.

For a downward-sloping line, the slope is a negative number because the changes in x and y move in opposite directions: If x increases, y decreases, and if x decreases, y increases. For a fairly flat downward-sloping line, the slope is a small negative number. For a steep downward-sloping line, the slope is a large negative number.

A horizontal line has a slope of zero because, in this case, the y-variable never changes. A vertical line is said to have an infinite slope because the y-variable can take any value without the x-variable changing at all.

What is the slope of Emma's demand curve for novels? First of all, because the curve slopes down, we know the slope will be negative. To calculate a numerical value for the slope, choose two points on the line. With Emma's income at $40,000, she buys 13 novels at a price of $8 or 21 novels at a price of $6. When we apply the

slope formula, we are concerned with the change between these two points. In other words, we are concerned with the difference between them, which tells us that we will have to subtract one set of values from the other, as follows:

$$\text{slope} = \frac{\Delta y}{\Delta x} = \frac{\text{second } y\text{-coordinate} - \text{first } y\text{-coordinate}}{\text{second } x\text{-coordinate} - \text{first } x\text{-coordinate}} = \frac{6 - 8}{21 - 13} = \frac{-2}{8} = \frac{-1}{4}$$

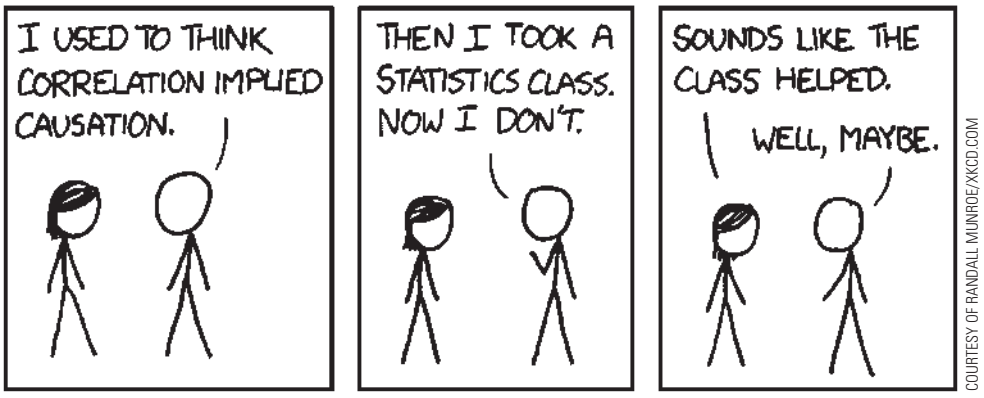
Figure A-5 shows graphically how this calculation works. Try computing the slope of Emma's demand curve using two different points. You should get the same result, – ¼. One of the properties of a straight line is that it has the same slope everywhere. This is not true of other types of curves, which are steeper in some places than in others.

The slope of Emma's demand curve tells us something about how responsive her purchases are to changes in the price. A small slope (a negative number close to zero) means that Emma's demand curve is relatively flat; in this case, she adjusts the number of novels she buys substantially in response to a price change. A larger slope (a negative number farther from zero) means that Emma's demand curve is relatively steep; in this case, she adjusts the number of novels she buys only slightly in response to a price change.

Cause and Effect

Economists often use graphs to advance an argument about how the economy works. In other words, they use graphs to argue about how one set of events **causes** another set of events. With a graph like the demand curve, there is no doubt about the cause and effect. Because we are varying price and holding all other variables constant, we know that changes in the price of novels cause changes in the quantity Emma demands. Remember, however, that our demand curve came from a hypothetical example. When graphing data from the real world, it is often more difficult to establish how one variable affects another.

The first problem is that it is difficult to hold everything else constant when studying the relationship between two variables. If we are not able to hold other variables constant, we might decide that one variable on our graph is causing changes in the other variable when those changes are actually being caused by a third **omitted variable** not pictured on the graph. Even if we have identified the correct two variables to look at, we might run into a second problem—**reverse causality**. In other words, we might decide that A causes B when, in fact, B causes A. The



COURTESY OF RANDALL MUNROE/XKCD.COM

omitted-variable and reverse-causality traps require us to proceed with caution when using graphs to draw conclusions about causes and effects.

Omitted Variables To see how omitting a variable can lead to a deceptive graph, consider an example. Imagine that the government, spurred by public concern about the large number of deaths from cancer, commissions an exhaustive study from Big Brother Statistical Services, Inc. Big Brother examines many of the items found in people's homes to see which of them are associated with the risk of cancer. Big Brother reports a strong relationship between two variables: the number of cigarette lighters that a household owns and the probability that someone in the household will develop cancer. Figure A-6 shows this relationship.

What should we make of this result? Big Brother advises a quick policy response. It recommends that the government discourage the ownership of cigarette lighters by taxing their sale. It also recommends that the government require warning labels: "Big Brother has determined that this lighter is dangerous to your health."

In judging the validity of Big Brother's analysis, one question is key: Has Big Brother held constant every relevant variable except the one under consideration? If the answer is no, the results are suspect. An easy explanation for Figure A-6 is that people who own more cigarette lighters are more likely to smoke cigarettes and that cigarettes, not lighters, cause cancer. If Figure A-6 does not hold constant the amount of smoking—and it doesn't because Big Brother never looked at that variable—it does not tell us the true effect of owning a cigarette lighter.

This story illustrates an important principle: When you see a graph used to support an argument about cause and effect, it is important to ask whether the movements of an omitted variable could explain the results you see.

Reverse Causality Economists can also make mistakes about causality by misreading its direction. To see how this is possible, suppose the Association of American Anarchists commissions a study of crime in America and arrives at Figure A-7, which plots the number of violent crimes per thousand people in major cities against the number of police officers per thousand people. The Anarchists note the curve's upward slope and argue that because police increase rather than decrease the amount of urban violence, law enforcement should be abolished.

Figure A-7, however, does not prove the Anarchists' point. The graph simply shows that more dangerous cities have more police officers. The explanation may be that more dangerous cities hire more police. In other words, rather than police

Figure A-6

Graph with an Omitted Variable

The upward-sloping curve shows that members of households with more cigarette lighters are more likely to develop cancer. Yet we should not conclude that ownership of lighters causes cancer because the graph does not take into account the number of cigarettes smoked.

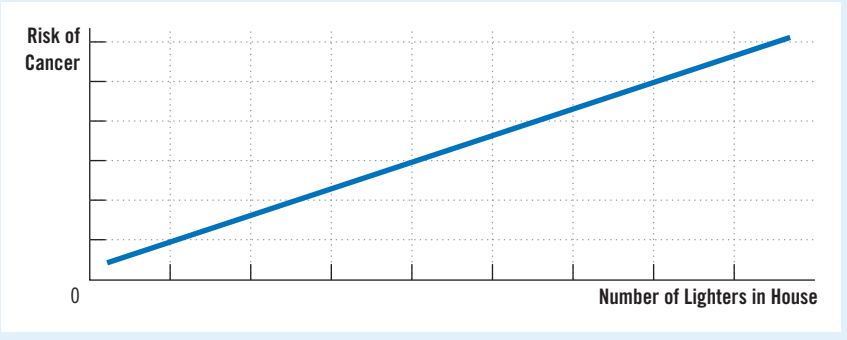
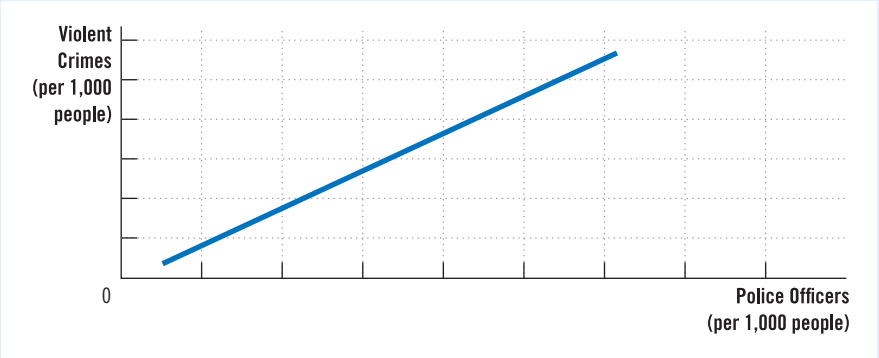


Figure A-7

Graph Suggesting Reverse Causality

The upward-sloping curve shows that cities with a higher concentration of police are more dangerous. Yet the graph does not tell us whether police cause crime or crime-plagued cities hire more police.



causing crime, crime may cause police. We could avoid the danger of reverse causality by running a controlled experiment. In this case, we would randomly assign different numbers of police to different cities and then examine the correlation between police and crime. Without such an experiment, establishing the direction of causality is difficult at best.

It might seem that we could determine the direction of causality by examining which variable moves first. If crime increases and then the police force expands, we reach one conclusion. If the police force expands and then crime increases, we reach the other conclusion. This approach, however, is also flawed: Often, people change their behavior not in response to a change in their present conditions but in response to a change in their **expectations** about future conditions. A city that expects a major crime wave in the future, for instance, might hire more police now. This problem is easier to see in the case of babies and minivans. Couples often buy a minivan in anticipation of the birth of a child. The minivan comes before the baby, but we wouldn't want to conclude that the sale of minivans causes the population to grow!

There is no complete set of rules that says when it is appropriate to draw causal conclusions from graphs. Yet just keeping in mind that cigarette lighters don't cause cancer (omitted variable) and that minivans don't cause larger families (reverse causality) will keep you from falling for many faulty economic arguments.

Chapter

3

Interdependence and the Gains from Trade

Consider a typical day. You wake up and pour juice from oranges grown in Florida and coffee from beans harvested in Brazil. Over breakfast, you read a news report edited in New York on a tablet made in China. You get dressed in clothes made of cotton grown in Georgia and sewn in factories in Thailand. You ride to class on a bicycle made of parts manufactured in half a dozen countries around the world. Then you open your economics textbook written by an author living in Massachusetts, published by a company located in Ohio, and printed on paper made from trees grown in Oregon.

Every day, you rely on many people, most of whom you have never met, to provide you with goods and services. Such interdependence is possible because people trade with one another. The people providing these things to you are not acting out of generosity, nor is some government agency directing them to satisfy your desires. Instead, people provide you and other consumers with the goods and services they produce because they get something in return.

BLACKCAT/E+/GETTY IMAGES, GOWITHSTOCK/SHUTTERSTOCK.COM

In later chapters, we examine how an economy coordinates the activities of millions of people with varying tastes and abilities. As a starting point, this chapter considers the reasons for economic interdependence. One of the **Ten Principles of Economics** in Chapter 1 is that trade can make everyone better off. We now examine this principle more closely. What exactly do people gain when they trade with one another? Why do people become interdependent?

The answers to these questions are key to understanding the global economy. Most countries today import from abroad many of the goods and services they consume, and they export to foreign customers much of what they produce. The analysis in this chapter explains interdependence not only among individuals but also among nations. As we will see, the gains from trade are much the same whether you are buying a haircut from your local barber or a T-shirt made on the other side of the globe.

3-1 A Parable for the Modern Economy

To understand how people benefit when they rely on one another for goods and services, let's examine a simple economy. Imagine that there are only two goods in the world: meat and potatoes. And there are only two people: a cattle rancher named Ruby and a potato farmer named Frank. Both Ruby and Frank would like to eat a diet of both meat and potatoes.

The gains from trade are clearest if Ruby can produce only meat and Frank can produce only potatoes. In one scenario, Frank and Ruby could choose to have nothing to do with each other. But after several months of eating beef roasted, broiled, seared, and grilled, Ruby might decide that self-sufficiency is not all it's cracked up to be. Frank, who has been eating potatoes mashed, fried, baked, and scalloped, would likely agree. It is easy to see that trade would allow both of them to enjoy greater variety: Each could then have a steak with a baked potato or a burger with fries.

Although this scene shows most simply how everyone can benefit from trade, the gains would be similar if Frank and Ruby were each capable of producing the other good, but only at great cost. Suppose, for example, that Ruby can grow potatoes, but her land is not well suited for it. Similarly, suppose that Frank can raise cattle and produce meat but is not good at it. In this case, Frank and Ruby benefit by specializing in what they do best and then trading with each other.

The gains are less obvious, however, when one person is better at producing **everything**. For example, imagine that Ruby is better at raising cattle **and** at growing potatoes. In this case, should Ruby remain self-sufficient? Or is there still a reason for her to trade with Frank? Let's look more closely at the factors that affect such a decision.

3-1a Production Possibilities

Suppose that Frank and Ruby each work 8 hours per day and can use this time to grow potatoes, raise cattle, or engage in a combination of the two. The table in Figure 1 shows the amount of time each person requires to produce 1 ounce of each good. Frank produces an ounce of potatoes in 15 minutes and an ounce of meat in 60 minutes. Ruby, who is better in both activities, can produce an ounce of potatoes in 10 minutes and an ounce of meat in 20 minutes. The last two columns show how much they can each produce if they devote all 8 hours to producing only meat or potatoes.

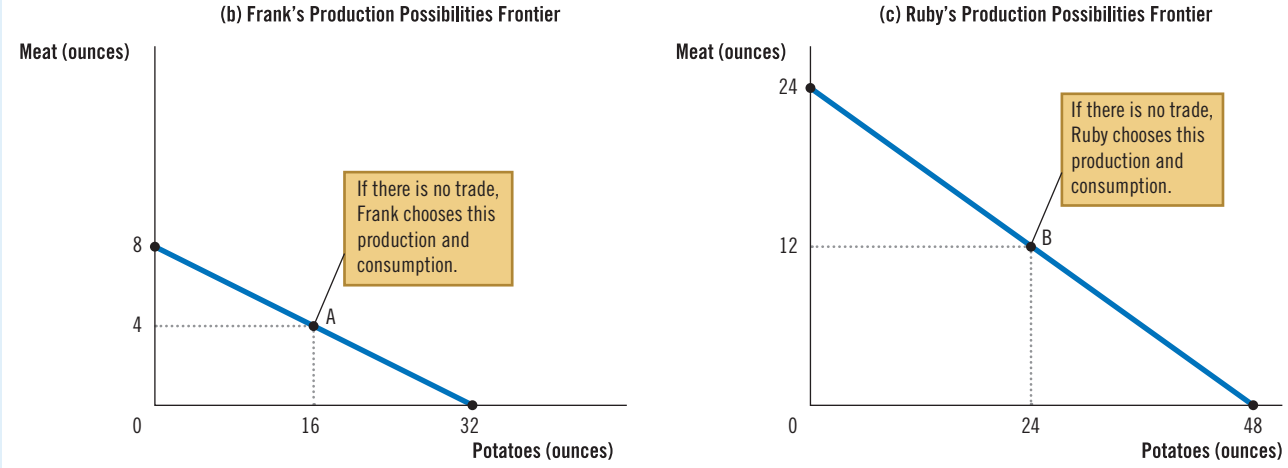
Figure 1

The Production Possibilities Frontier

Panel (a) shows the production opportunities available to Frank the farmer and Ruby the rancher. Panel (b) shows the combinations of meat and potatoes that Frank can produce. Panel (c) shows the combinations of meat and potatoes that Ruby can produce. Both production possibilities frontiers assume that Frank and Ruby each work 8 hours per day. If there is no trade, their production possibilities frontiers are also their consumption possibilities frontiers.

(a) Production Opportunities

	Minutes Needed to Make 1 Ounce of:		Amount Produced in 8 Hours	
	Meat	**Potatoes**	**Meat**	**Potatoes**
Frank the farmer	60 min/oz	15 min/oz	8 oz	32 oz
Ruby the rancher	20 min/oz	10 min/oz	24 oz	48 oz



Panel (b) of Figure 1 illustrates the amounts of meat and potatoes that Frank can produce. If he spends all 8 hours growing potatoes, Frank produces 32 ounces of potatoes (measured on the horizontal axis) and no meat. If he spends all of his time raising cattle, he produces 8 ounces of meat (measured on the vertical axis) and no potatoes. If Frank divides his time equally between the two activities, spending 4 hours on each, he produces 16 ounces of potatoes and 4 ounces of meat. The figure shows these three outcomes and all others in between.

This graph is Frank's production possibilities frontier. As we discussed in Chapter 2, a production possibilities frontier shows the mixes of output that an economy can produce. It illustrates one of the **Ten Principles of Economics** in Chapter 1: People face trade-offs. Here, Frank faces a trade-off between producing meat and producing potatoes.

You may recall that the production possibilities frontier in Chapter 2 was drawn bowed out. In that case, the rate at which society could trade one good for the other depended on the amounts that were being produced. Here, however, Frank can switch between the production of meat and potatoes (summarized in

Figure 1) at a constant rate. When Frank cuts 1 hour from producing meat and adds 1 hour to producing potatoes, he reduces his meat output by 1 ounce and raises his potato output by 4 ounces, and this is true regardless of how much he is already producing. As a result, the production possibilities frontier is a straight line.

Panel (c) of Figure 1 shows Ruby's production possibilities frontier. If she only grows potatoes, Ruby produces 48 ounces of them and no meat. If she only raises cattle, she produces 24 ounces of meat and no potatoes. If Ruby divides her time equally, spending 4 hours on each activity, she produces 24 ounces of potatoes and 12 ounces of meat. Once again, the production possibilities frontier shows all possible outcomes.

If Frank and Ruby remain self-sufficient instead of trading with each other, each consumes exactly what he or she produces. In this case, the production possibilities frontier is also the consumption possibilities frontier. That is, without trade, Figure 1 shows the possible combinations of meat and potatoes that Frank and Ruby can each produce and then consume.

These production possibilities frontiers are useful in showing the trade-offs that Frank and Ruby face, but they do not tell us what each will choose to do. For that, we need to know something about their dietary preferences. Suppose that Frank and Ruby choose the combinations identified by points A and B in Figure 1. Based on his opportunities and tastes, Frank decides to produce and consume 16 ounces of potatoes and 4 ounces of meat, while Ruby decides to produce and consume 24 ounces of potatoes and 12 ounces of meat.

3-1b Specialization and Trade

After several years of eating combination B, Ruby gets an idea and visits Frank:

Ruby: Frank, my friend, have I got a deal for you! We can improve life for both of us. You should stop producing meat altogether and just grow potatoes. According to my calculations, if you do that for 8 hours a day, you'll produce 32 ounces of potatoes. You can then give me 15 of those 32 ounces, and I'll give you 5 ounces of meat in return. It's great! You'll eat 17 ounces of potatoes and 5 ounces of meat every day instead of the 16 ounces of potatoes and 4 ounces of meat you eat now. With my plan, you'll have more of **both** foods. [To illustrate her point, Ruby shows Frank panel (a) of Figure 2.]

Frank: (sounding skeptical): That seems like a good deal for me. But I don't understand why you are offering it. If the deal is so good for me, it can't be good for you too.

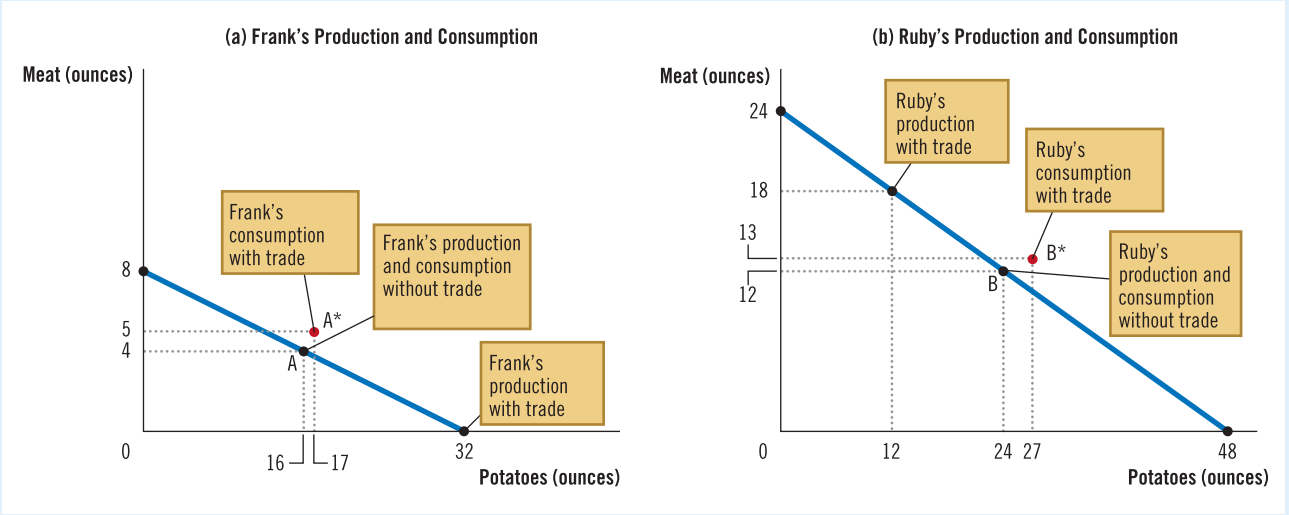
Ruby: Oh, but it is! Suppose I spend 6 hours a day raising cattle and 2 hours growing potatoes. Then I can produce 18 ounces of meat and 12 ounces of potatoes. After I give you 5 ounces of my meat in exchange for 15 ounces of your potatoes, I'll end up with 13 ounces of meat and 27 ounces of potatoes instead of the 12 ounces of meat and 24 ounces of potatoes that I have now. So I will also consume more of both foods than I do now. [She points out panel (b) of Figure 2.]

Frank: I don't know. . . . This sounds too good to be true.

Figure 2

How Trade Expands the Set of Consumption Opportunities

The proposed trade offers Frank and Ruby a combination of meat and potatoes that would be impossible without trade. In panel (a), Frank consumes at point A* rather than point A. In panel (b), Ruby consumes at point B* rather than point B. Trade allows each to consume more meat and more potatoes.



(c) The Gains from Trade: A Summary

	Frank		Ruby	
	Meat	**Potatoes**	**Meat**	**Potatoes**
Without Trade:				
Production and Consumption	4 oz	16 oz	12 oz	24 oz
With Trade:				
Production	0 oz	32 oz	18 oz	12 oz
Trade	Gets 5 oz	Gives 15 oz	Gives 5 oz	Gets 15 oz
Consumption	5 oz	17 oz	13 oz	27 oz
GAINS FROM TRADE:				
Increase in Consumption	+1 oz	+1 oz	+1 oz	+3 oz

Ruby: It's really not complicated. Here—I've summarized my proposal for you in a simple table. [Ruby shows Frank a copy of the table at the bottom of Figure 2.]

Frank: (after pausing to study the table): These calculations seem correct, but I am puzzled. How can this deal make us both better off?

Ruby: Trade allows each of us to do what we do best. You spend more time growing potatoes and less raising cattle. I spend more time raising cattle and less growing potatoes. Thanks to specialization and trade, each of us can consume more meat and more potatoes without working more hours.

QuickQuiz

1. Before Frank and Ruby engage in trade, each consumes
   a. at a point inside his or her production possibilities frontier.
   b. at a point on his or her production possibilities frontier.
   c. at a point outside his or her production possibilities frontier.
   d. the same amounts of meat and potatoes as the other.

2. After Frank and Ruby trade, each of them consumes
   a. at a point inside his or her production possibilities frontier.
   b. at a point on his or her production possibilities frontier.
   c. at a point outside his or her production possibilities frontier.
   d. the same amounts of meat and potatoes as the other.

Answers are at the end of the chapter.

3-2 Comparative Advantage: The Driving Force of Specialization

Ruby's explanation of the gains from trade, though correct, poses a puzzle: If Ruby is better at both raising cattle and growing potatoes, how can Frank ever specialize in doing what he does best? Frank doesn't seem to do anything best. To solve this puzzle, we need to look at the principle of **comparative advantage**.

As a first step, consider the following question: In our example, who can produce potatoes at a lower cost—Frank or Ruby? There are two possible answers, and in them lies the key to understanding the gains from trade.

3-2a Absolute Advantage

absolute advantage
the ability to produce a good using fewer inputs than another producer

One way to answer the question about the cost of producing potatoes is to compare the inputs each producer requires. Economists use the term **absolute advantage** when comparing the productivity of one person, firm, or nation to that of another. The producer that requires a smaller quantity of inputs to produce a good is said to have an absolute advantage in producing that good.

In our simple example, time is the only input, so it is all we need to examine to determine absolute advantage. Ruby has an absolute advantage in producing both meat and potatoes because she requires less time than Frank to produce a unit of either good. She needs only 20 minutes to produce an ounce of meat, while Frank needs 60 minutes. Similarly, it takes Ruby only 10 minutes to produce an ounce of potatoes, while it takes Frank 15 minutes. Thus, if cost is measured in terms of the quantity of inputs, Ruby produces potatoes at a lower cost.

3-2b Opportunity Cost and Comparative Advantage

opportunity cost
whatever must be given up to obtain some item

There is another way to look at the cost of producing potatoes. Rather than focusing on the inputs required, we can examine the opportunity costs. Recall from Chapter 1 that the **opportunity cost** of an item is what you give up to get that item. We have assumed that Frank and Ruby each work 8 hours a day. Time spent producing potatoes takes away from time available for producing meat. When reallocating time between the two goods, Ruby and Frank give up

units of one good to produce units of the other, moving along the production possibilities frontier. The opportunity cost measures the trade-off that each producer faces.

First, consider Ruby's opportunity cost. According to the table in panel (a) of Figure 1, it takes her 10 minutes to grow 1 ounce of potatoes, time that she isn't using to produce meat. Because Ruby needs 20 minutes to produce 1 ounce of meat, 10 minutes would yield ½ ounce of meat. Hence, Ruby's opportunity cost of producing 1 ounce of potatoes is ½ ounce of meat.

Next, consider Frank's situation. Producing 1 ounce of potatoes takes him 15 minutes. Because he needs 60 minutes to produce 1 ounce of meat, 15 minutes would yield ¼ ounce of meat. Hence, Frank's opportunity cost of producing 1 ounce of potatoes is ¼ ounce of meat.

Table 1 shows the opportunity costs of meat and potatoes for each of them. Notice that the opportunity cost of meat is the inverse of the opportunity cost of potatoes. Because 1 ounce of potatoes costs Ruby ½ ounce of meat, it is also true that 1 ounce of meat costs her 2 ounces of potatoes. Similarly, because 1 ounce of potatoes costs Frank ¼ ounce of meat, 1 ounce of meat costs him 4 ounces of potatoes.

Economists use the term **comparative advantage** when describing the opportunity costs faced by two producers. The producer who gives up less of the other good to produce Good X has the smaller opportunity cost of producing Good X and is said to have a comparative advantage in producing it. In our example, Frank has a lower opportunity cost of producing potatoes than Ruby: An ounce of potatoes costs Frank only ¼ ounce of meat but costs Ruby ½ ounce of meat. Conversely, Ruby has a lower opportunity cost of producing meat than Frank: An ounce of meat costs Ruby 2 ounces of potatoes but costs Frank 4 ounces of potatoes. Thus, Frank has a comparative advantage in growing potatoes, and Ruby has a comparative advantage in producing meat.

comparative advantage the ability to produce a good at a lower opportunity cost than another producer

Although it is possible for a person to have an absolute advantage in both goods (as Ruby does in our example), it is impossible for a person to have a comparative advantage in both goods. Because the opportunity cost of one good is the inverse of the opportunity cost of the other, if a person's opportunity cost of one good is relatively high, his or her opportunity cost of the other good must be relatively low. Unless two people have the same opportunity cost, one person will have a comparative advantage in one good, and the other person will have a comparative advantage in the other good.

Table 1

The Opportunity Cost of Meat and Potatoes

	Opportunity Cost of:	
	1 oz of Meat	1 oz of Potatoes
Frank the farmer	4 oz potatoes	¼ oz meat
Ruby the rancher	2 oz potatoes	½ oz meat

3-2c Comparative Advantage and Trade

The gains from specialization and trade are based on comparative advantage. When people produce goods in which they have a comparative advantage, total production rises. The economic pie grows larger. Depending on how this bounty is divided, everyone can be better off.

Once they start to trade, Frank spends more time growing potatoes, and Ruby works more on producing meat. Total potato production increases from 40 to 44 ounces, and total meat production increases from 16 to 18 ounces. Frank and Ruby share the benefits of the greater production.

The gains are reflected in the implicit prices that the trading partners pay each other. Because Frank and Ruby have different opportunity costs, they both get a bargain. That is, each of them benefits from trade by obtaining a good at a price that is lower than his or her opportunity cost of that good.

Consider the deal from Frank's viewpoint. He receives 5 ounces of meat in exchange for 15 ounces of potatoes. In other words, Frank buys each ounce of meat for a price of 3 ounces of potatoes. This price of meat is lower than his opportunity cost of an ounce of meat, which is 4 ounces of potatoes. Frank benefits from the deal because he gets to buy meat at a good price.

Now consider Ruby's viewpoint. She gets 15 ounces of potatoes in exchange for 5 ounces of meat. That is, the price of an ounce of potatoes is ⅓ ounce of meat. This price of potatoes is lower than her opportunity cost of an ounce of potatoes, which is ½ ounce of meat. Ruby benefits because she gets to buy potatoes at a good price.

The story of Ruby the rancher and Frank the farmer has a simple moral: **Trade can benefit everyone because it allows people to specialize in the activities in which they have a comparative advantage.**

3-2d The Price of the Trade

The principle of comparative advantage helps to explain the gains from specialization and trade, but it raises a couple of related questions: What determines the price at which trade takes place? How are the gains shared between the trading parties? The precise answers to these questions are beyond the scope of this chapter, but here is a general rule: **For both parties to gain from trade, the price at which they trade must lie between their opportunity costs.**

In our example, Frank and Ruby agreed to trade at a rate of 3 ounces of potatoes per ounce of meat. This price is between Ruby's opportunity cost (2 ounces of potatoes per ounce of meat) and Frank's opportunity cost (4 ounces of potatoes per ounce of meat). The price need not be exactly in the middle for both parties to gain, but it must be somewhere between 2 and 4.

Consider what would happen at prices outside this range. If meat was priced below 2 ounces of potatoes, both Frank and Ruby would want to buy meat because it would cost less than each of their opportunity costs. Similarly, if meat was priced above 4 ounces of potatoes, both would want to sell meat because the price would be above their opportunity costs. But they cannot both be buyers of meat, nor can they both be sellers. Someone must take the other side of the deal. Trade doesn't work at these prices.

A mutually advantageous trade can be struck at prices between 2 and 4. In this range, Ruby wants to sell meat to buy potatoes, and Frank wants to sell potatoes to buy meat. They both get to buy a good at a price below their opportunity cost of that good. In the end, they specialize in the good in which they have a comparative advantage, and, as a result, both of them are better off.

FYI

The Legacy of Adam Smith and David Ricardo

Here is how the great economist Adam Smith put the argument for the gains from trade:

> It is a 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 tailor 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 tailor. The farmer attempts to make neither the one nor the other, but employs those different artificers. All of them find it for their interest to employ their whole industry in a way in which they have some advantage over their neighbors, and to purchase with a part of its produce, or what is the same thing, with the price of part of it, whatever else they have occasion for.

BETTMANN/GETTY IMAGES

David Ricardo

This quotation is from Smith's 1776 book *The Wealth of Nations*, which was a landmark in the analysis of trade and economic interdependence.

Smith's book inspired David Ricardo, a millionaire stockbroker, to become an economist. In his 1817 book *On the Principles of Political Economy and Taxation*, Ricardo developed the principle of comparative advantage as we know it today. He considered an example with two goods (wine and cloth) and two countries (England and Portugal). He showed that both countries could gain by opening trade and specializing.

While Ricardo's theory is the starting point of modern international economics, his defense of free trade was not merely an academic exercise. He put his findings to work as a member of the British Parliament, where he opposed the Corn Laws, which restricted grain imports.

The conclusions of Smith and Ricardo on the gains from trade have held up well over time. Although economists often disagree on questions of policy, they are nearly united in their support of free trade. Moreover, the central argument has not changed much in the past two centuries. Even though the field of economics has broadened its scope and refined its theories, economists' opposition to trade restrictions is still based largely on the principle of comparative advantage. ■

QuickQuiz

3. In an hour, Mateo can wash 2 cars or mow 1 lawn, and Sophia can wash 3 cars or mow 1 lawn. Who has the absolute advantage in car washing, and who has it in lawn mowing?
   a. Mateo in washing, Sophia in mowing
   b. Sophia in washing, Mateo in mowing
   c. Mateo in washing, neither in mowing
   d. Sophia in washing, neither in mowing
4. Between Mateo and Sophia, who has the comparative advantage in car washing, and who has it in lawn mowing?
   a. Mateo in washing, Sophia in mowing
   b. Sophia in washing, Mateo in mowing
   c. Mateo in washing, neither in mowing
   d. Sophia in washing, neither in mowing
5. When Mateo and Sophia produce efficiently and make a mutually beneficial trade based on comparative advantage,
   a. Mateo mows more, and Sophia washes more.
   b. Mateo washes more, and Sophia mows more.
   c. Mateo and Sophia both wash more.
   d. Mateo and Sophia both mow more.

Answers are at the end of the chapter.

3-3 Applications of Comparative Advantage

The principle of comparative advantage explains interdependence and the gains from trade. Because interdependence is so prevalent, the principle of comparative advantage has many applications. Here are two examples, one fanciful and one of great practical importance.

ADRIAN DENNIS/AFP/GETTY IMAGES

Naomi Osaka may be good at pushing a lawnmower, but it's not her comparative advantage.

3-3a Should Naomi Osaka Mow Her Own Lawn?

Naomi Osaka is a great athlete. One of the best tennis players of the current era, she can run faster and hit a ball harder than most other people. Most likely, she is talented at other physical activities as well. For example, let's imagine that Osaka can mow her lawn faster than anyone else. But just because she **can** mow her lawn quickly, does this mean she **should**? If she enjoys mowing as a favorite form of relaxation, then of course she should. Otherwise, she can reach a better outcome by applying the concepts of opportunity cost and comparative advantage.

Let's say that Osaka can mow her lawn in 2 hours. In those same 2 hours, she could film a television commercial and earn $30,000. By contrast, Hari, the boy next door, can mow Osaka's lawn in 4 hours. In those same 4 hours, Hari could work at McDonald's and earn $50.

Osaka has an absolute advantage in mowing lawns because she can do the work in less time. Yet because her opportunity cost of mowing the lawn is $30,000 and Hari's is only $50, Hari has a comparative advantage in mowing lawns.

The gains from trade here are tremendous. Rather than mowing her own lawn, Osaka should film the commercial and hire Hari to mow the lawn. As long as Osaka pays Hari more than $50 and less than $30,000, both are better off.

3-3b Should the United States Trade with Other Countries?

Just as individuals can benefit from specialization and trade with one another, so can countries. Many of the goods that Americans enjoy are produced abroad, and many of the goods produced in the United States are sold abroad. Goods produced abroad and sold domestically are **imports**. Those produced domestically and sold abroad are **exports**.

imports
goods produced abroad and sold domestically

exports
goods produced domestically and sold abroad

Let's focus on the United States and Japan, both of which produce food and cars. Imagine that they produce cars equally well: An American worker and a Japanese worker can each produce one car per month. By contrast, because the United States has more fertile land, it is better at producing food: A U.S. worker can produce 2 tons of food per month, while a Japanese worker can produce only 1 ton of food per month.

The principle of comparative advantage states that each good should be produced by the country with the lower opportunity cost of producing that good. Because the opportunity cost of a car is 2 tons of food in the United States but only 1 ton of food in Japan, Japan has a comparative advantage in producing cars. Japan should produce more cars than it wants for its own use and export some of them to the United States. Similarly, because the opportunity cost of a ton of food is 1 car in Japan but only ½ car in the United States, the United States has a comparative advantage in producing food. The United States should produce more food than it wants to consume and export some to Japan. Through specialization and trade, both countries can have more food and more cars.

To be sure, the issues involved in trade among nations are more complex than this simple example suggests. Most importantly, each country has many people, and trade affects them in different ways. When the United States exports food and imports cars, the impact on an American farmer is not the same as the

impact on an American autoworker. As a result, international trade can make some individuals worse off, even as it makes the country as a whole better off. Yet this example teaches an important lesson: Contrary to the opinions sometimes voiced by politicians and pundits, international trade is not like war, in which some countries win and others lose. Trade allows all countries to achieve greater prosperity.

3-4 Conclusion

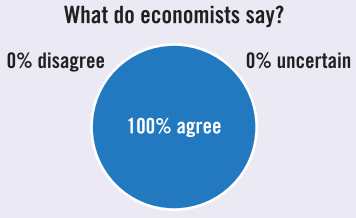
The benefits of living in an interdependent economy are enormous. When Americans buy tube socks from China, when residents of Maine drink orange juice from Florida, and when homeowners hire local kids to mow their lawns, the same economic forces are at work. The principle of comparative advantage shows that trade can make everyone better off.

Having seen why interdependence is desirable, you might ask how it is possible. How do free societies coordinate the diverse activities of all the people involved in their economies? What ensures that goods and services will get from those who should be producing them to those who should be consuming them? In a world with only two people, such as Ruby and Frank, the answer is simple: They can bargain and allocate resources directly. In the real world, with billions of people, the task is far more complex. As we will see in the next chapter, most economies allocate resources using the market forces of supply and demand.

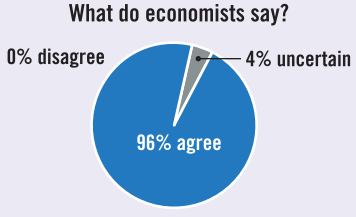
Ask the Experts

Trade between China and the United States

"Trade with China makes most Americans better off because, among other advantages, they can buy goods that are made or assembled more cheaply in China."



"Some Americans who work in the production of competing goods, such as clothing and furniture, are made worse off by trade with China."



Source: IGM Economic Experts Panel, June 19, 2012.

QuickQuiz

6. A nation will typically import those goods in which
   a. the nation has an absolute advantage.
   b. the nation has a comparative advantage.
   c. other nations have an absolute advantage.
   d. other nations have a comparative advantage.
7. Suppose that in the United States, producing an aircraft takes 10,000 hours of labor, and producing a shirt takes 2 hours of labor. In China, producing an aircraft takes 40,000 hours of labor, and producing a shirt takes 4 hours of labor. What will these nations trade?
   a. China will export aircraft, and the United States will export shirts.
   b. China will export shirts, and the United States will export aircraft.
   c. Both nations will export shirts.
   d. There are no gains from trade in this situation.
8. Kayla can cook dinner in 30 minutes and wash the laundry in 20 minutes. Her roommate takes twice as long to do each task. How should the roommates allocate the work?
   a. Kayla should do more of the cooking based on her comparative advantage.
   b. Kayla should do more of the washing based on her comparative advantage.
   c. Kayla should do more of the washing based on her absolute advantage.
   d. There are no gains from trade in this situation.

Answers are at the end of the chapter.

Economics within a Marriage

An economist argues that you shouldn't always unload the dishwasher just because you're better at it than your partner.

You're Dividing the Chores Wrong

By Emily Oster

No one likes doing chores. In happiness surveys, housework is ranked down there with commuting as activities that people enjoy the least. Maybe that's why figuring out who does which chores usually prompts, at best, tense discussion in a household and, at worst, outright fighting.

If everyone is good at something different, assigning chores is easy. If your partner is great at grocery shopping and you are great at the laundry, you're set. But this isn't always—or even usually—the case. Often one person is better at everything. (And let's be honest, often that person is the woman.) Better at the laundry, the grocery shopping, the cleaning, the cooking. But does that mean she should have to do everything?

Before my daughter was born, I both cooked and did the dishes. It wasn't a big deal, it didn't take too much time, and honestly I was a lot better at both than my husband. His cooking repertoire extended only to eggs and chili, and when I left him in charge of the dishwasher, I'd often find he had run it "full" with one pot and eight forks.

After we had a kid, we had more to do and less time to do it in. It seemed like it was time for some reassignments. But, of course, I was still better at doing both things. Did that mean I should do them both?

I could have appealed to the principle of fairness: We should each do half. I could have appealed to feminism—surveys show that women more often than not get the short end of the chore stick. In time-use data, women do about 44 minutes more housework than men (2 hours and 11 minutes versus 1 hour and 27 minutes). Men outwork women only in the areas of "lawn" and "exterior maintenance." I could have suggested he do more chores to rectify this imbalance, to show our daughter, in the *Free to Be You and Me* style, that Mom and Dad are equal and that housework is fun if we do it together! I could have simply smashed around the pans in the dishwasher while sighing loudly in the hopes he would notice and offer to do it himself.

But luckily for me and my husband, I'm an economist, so I have more effective tools than passive aggression. And some basic economic principles provided the answer. We needed to divide the chores because it is simply not **efficient** for the best cook and dishwasher to do all the cooking and dishwashing. The economic principle at play here is increasing marginal cost. Basically, people get worse when they are tired. When I teach my students this principle, I explain it in the context of managing their employees. Imagine you have a good employee and a not-so-good one. Should you make the good employee do literally everything?

Usually, the answer is no. Why not? It's likely that the not-so-good employee is better at 9 a.m. after a full night of sleep than the good employee is at 2 a.m. after a 17-hour workday. So you want to give at least a few tasks to your worse guy. The same principle applies in your household. Yes, you (or your spouse) might be better at everything. But anyone doing the laundry at 4 a.m. is likely to put the red towels in with the white T-shirts. Some task splitting is a good idea. How much depends on how fast people's skills decay.

To "optimize" your family efficiency (every economist's ultimate goal—and yours, too), you want to equalize effectiveness on the final task each person is doing. Your partner does the dishes, mows the lawn, and makes

Chapter in a Nutshell

- Each person consumes goods and services produced by many other people both in the United States and around the world. Interdependence and trade are desirable because they allow everyone to enjoy a greater quantity and variety of goods and services.
- There are two ways to compare the abilities of two people to produce a good. The person who can produce the good with the smaller quantity of inputs is said to have an **absolute advantage** in producing the good. The person who has the lower opportunity cost of producing the good is said to have a **comparative advantage**. The gains from trade are based on comparative advantage, not absolute advantage.
- Trade makes everyone better off because it allows people to specialize in those activities in which they have a comparative advantage.
- The principle of comparative advantage applies to countries as well as to people. Economists use the principle of comparative advantage to advocate free trade among countries.

the grocery list. You do the cooking, laundry, shopping, cleaning, and paying the bills. This may seem imbalanced, but when you look at it, you see that by the time your partner gets to the grocery-list task, he is wearing thin and starting to nod off. It's all he can do to figure out how much milk you need. In fact, he is just about as good at that as you are when you get around to paying the bills, even though that's your fifth task.

If you then made your partner also do the cleaning—so it was an even four and four—the house would be a disaster, since he is already exhausted by his third chore while you are still doing fine. This system may well end up meaning one person does more, but it is unlikely to result in one person doing everything.

Once you've decided you need to divide up the chores in this way, how should you decide who does what? One option would be randomly assigning tasks; another would be having each person do some of everything. One spousal-advice website I read suggested you should divide tasks based on which ones you like the best. None of these are quite right. (In the last case, how would anyone ever end up with the job of cleaning the bathroom?)

To decide who does what, we need more economics. Specifically, the principle of comparative advantage. Economists usually talk about this in the context of trade. Imagine Finland is better than Sweden at making both reindeer hats and snowshoes. But they are much, much better at the hats and only a little better at the snowshoes. The overall world production is maximized when Finland makes hats and Sweden makes snowshoes.

DON EMMERT/AFP/GETTY IMAGES

Emily Oster

We say that Finland has an **absolute advantage** in both things but a **comparative advantage** only in hats. This principle is part of the reason economists value free trade, but that's for another column (and probably another author). But it's also a guideline for how to trade tasks in your house. You want to assign each person the tasks on which he or she has a comparative advantage. It doesn't matter that you have an absolute advantage in everything. If you are much, much better at the laundry and only a little better at cleaning the toilet, you should do the laundry and have your spouse get out the scrub brush. Just explain that it's efficient!

In our case, it was easy. Other than using the grill—which I freely admit is the husband domain—I'm much, much better at cooking. And I was only moderately better at the dishes. So he got the job of cleaning up after meals, even though his dishwasher loading habits had already come under scrutiny. The good news is another economic principle I hadn't even counted on was soon in play: **learning by doing**. As people do a task, they improve at it. Eighteen months into this new arrangement the dishwasher is almost a work of art: neat rows of dishes and everything carefully screened for "top-rack only" status. I, meanwhile, am forbidden from getting near the dishwasher. Apparently, there is a risk that I'll "ruin it." ■

Questions to Discuss

1. In your family, do you think tasks are divided among family members according to comparative advantage? If so, how? If not, how might the allocation of tasks be improved?
2. Do you think being married to an economist would facilitate family harmony or just the opposite?

Ms. Oster is a professor of economics at Brown University.

Source: *Slate*, November 21, 2012.

Key Concepts

absolute advantage, p. 50
comparative advantage, p. 51
exports, p. 54
imports, p. 54
opportunity cost, p. 50

Questions for Review

1. Under what conditions is the production possibilities frontier linear rather than bowed out?
2. Explain how absolute advantage and comparative advantage differ.

3. Give an example in which one person has an absolute advantage in doing something, but another person has a comparative advantage.
4. Is absolute advantage or comparative advantage more important for trade? Explain your reasoning using the example in your answer to question 3.
5. If two parties trade based on comparative advantage and both gain, in what range must the price implicit in the trade lie?
6. Why do economists oppose policies that restrict trade among nations?

Problems and Applications

1. Maria can read 20 pages of economics in an hour. She can also read 50 pages of sociology in an hour. She spends 5 hours per day studying.
   a. Draw Maria's production possibilities frontier for reading economics and sociology.
   b. What is Maria's opportunity cost of reading 100 pages of sociology?
2. American and Japanese workers can each produce 4 cars per year. An American worker can produce 10 tons of grain per year, while a Japanese worker can produce 5 tons of grain per year. To keep things simple, assume that each country has 100 million workers.
   a. For this situation, construct a table analogous to the table in Figure 1.
   b. Graph the production possibilities frontiers for the American and Japanese economies.
   c. For the United States, what is the opportunity cost of a car? Of grain? For Japan, what is the opportunity cost of a car? Of grain? Put this information in a table analogous to Table 1.
   d. Which country has an absolute advantage in producing cars? In producing grain?
   e. Which country has a comparative advantage in producing cars? In producing grain?
   f. Without trade, half of each country's workers produce cars, and half produce grain. What quantities of cars and grain does each country produce?
   g. Starting from a position without trade, give an example in which trade makes each country better off.
3. Diego and Darnell are roommates. They spend most of their time studying (of course), but they leave some time for their favorite activities: making pizza and brewing root beer. Diego takes 4 hours to brew a gallon of root beer and 2 hours to make a pizza. Darnell takes 6 hours to brew a gallon of root beer and 4 hours to make a pizza.
   a. What is each roommate's opportunity cost of making a pizza? Who has the absolute advantage in making pizza? Who has the comparative advantage in making pizza?
   b. If Diego and Darnell trade foods with each other, who will trade pizza in exchange for root beer?
   c. The price of pizza can be expressed in terms of gallons of root beer. What is the highest price at which pizza can be traded that would make both roommates better off? What is the lowest price? Explain.
4. Suppose that there are 10 million workers in Canada and that each of these workers can produce either 2 cars or 30 bushels of wheat in a year.
   a. What is the opportunity cost of producing a car in Canada? What is the opportunity cost of producing a bushel of wheat in Canada? Explain the relationship between the opportunity costs of the two goods.
   b. Draw Canada's production possibilities frontier. If Canada chooses to consume 10 million cars, how much wheat can it consume without trade? Label this point on the production possibilities frontier.
   c. Now suppose that the United States offers to buy 10 million cars from Canada in exchange for 20 bushels of wheat per car. If Canada continues to consume 10 million cars, how much wheat does this deal allow Canada to consume? Label this point on your diagram. Should Canada accept the deal?
5. England and Scotland both produce scones and sweaters. Suppose that an English worker can produce 50 scones per hour or 1 sweater per hour. Suppose that a Scottish worker can produce 40 scones per hour or 2 sweaters per hour.
   a. Which country has the absolute advantage in the production of each good? Which country has the comparative advantage?
   b. If England and Scotland decide to trade, which commodity will Scotland export to England? Explain.
   c. If a Scottish worker could produce only 1 sweater per hour, would Scotland still gain from trade? Would England still gain from trade? Explain.

6. The following table describes the production possibilities of two cities in the country of Baseballia:

	Pairs of Red Socks per Worker per Hour	Pairs of White Socks per Worker per Hour
Boston	3	3
Chicago	2	1

   a. Without trade, what is the price of white socks (in terms of red socks) in Boston? What is the price in Chicago?
   b. Which city has an absolute advantage in the production of each color sock? Which city has a comparative advantage in the production of each color sock?
   c. If the cities trade with each other, which color sock will each export?
   d. What is the range of prices at which mutually beneficial trade can occur?

7. A German worker takes 400 hours to produce a car and 2 hours to produce a case of wine. A French worker takes 600 hours to produce a car and *X* hours to produce a case of wine.
   a. For what values of *X* will gains from trade be possible? Explain.
   b. For what values of *X* will Germany export cars and import wine? Explain.

8. Suppose that in a year, an American worker can produce 100 shirts or 20 computers, and a Chinese worker can produce 100 shirts or 10 computers.
   a. For each country, graph the production possibilities frontier. Suppose that without trade, the workers in each country spend half their time producing each good. Identify this point in your graphs.
   b. If these countries were open to trade, which country would export shirts? Give a specific numerical example and show it on your graphs. Which country would benefit from trade? Explain.
   c. Explain at what price of computers (in terms of shirts) the two countries might trade.
   d. Suppose that China catches up with American productivity so that a Chinese worker can produce 100 shirts or 20 computers in a year. What pattern of trade would you predict now? How does this advance in Chinese productivity affect the economic well-being of the two countries' citizens?

9. Are the following statements true or false? Explain your answer in each case.
   a. "Two countries can achieve gains from trade even if one of the countries has an absolute advantage in the production of all goods."
   b. "Certain talented people have a comparative advantage in everything they do."
   c. "If a certain trade is good for one person, it can't be good for the other one."
   d. "If a certain trade is good for one person, it is always good for the other one."
   e. "If trade is good for a country, it must be good for everyone in the country."

Quick**Quiz Answers**

1. **b** 2. **c** 3. **d** 4. **b** 5. **a** 6. **d** 7. **b** 8. **d**

Chapter

4

The Market Forces of Supply and Demand

When a cold snap hits Florida, the price of orange juice rises in supermarkets throughout the United States. When the weather turns warm in New England every summer, the price of hotel rooms in the Caribbean plummets. When a war breaks out in the Middle East, the price of gasoline in the United States rises, and the price of a used SUV falls. What do these events have in common? They all show the workings of supply and demand.

Supply and **demand** are the two words economists use most often—and for good reason. They are the forces that make market economies work, determining the quantity of each good produced and the price at which it is sold. If you want to know how events and policies affect the economy, study supply and demand.

This chapter introduces the theory of supply and demand. It considers how buyers and sellers behave and interact, how supply and demand determine prices, and how prices allocate scarce resources.

SHIFT DRIVE/SHUTTERSTOCK.COM, ERDARKE+/GETTY IMAGES

4-1 Markets and Competition

The terms **supply** and **demand** refer to the behavior of people as they interact in competitive markets. Let's first discuss the meaning of the terms **market** and **competition**.

4-1a What Is a Market?

market
a group of buyers and sellers of a particular good or service

A **market** is a group of buyers and sellers of a good or service. The buyers determine the demand for the product, and the sellers determine the supply of the product.

Markets take many forms. Some are highly organized. In the markets for wheat and corn, buyers and sellers meet at a specific time and place, knowing how much of these agricultural commodities they are willing to buy and sell at various prices. An auctioneer keeps the process orderly by arranging sales and (most importantly) finding the prices that bring the buying and selling into balance.

More often, markets are less organized than that. For example, consider the market for ice cream in a particular town. Ice-cream buyers do not all meet at any one time or place. The sellers are in several locations and offer somewhat different toppings and flavors. No auctioneer calls out the price of a sundae. Each seller posts a price for an ice-cream cone, and each buyer decides how many to buy at each store. Nonetheless, these consumers and producers of ice cream are closely connected. The buyers are choosing among the various sellers to satisfy their cravings, and the sellers are all trying to attract the same buyers to make their businesses succeed. Even though they do not look as organized, the ice-cream buyers and sellers form a market.

4-1b What Is Competition?

The ice-cream market, like many markets in the economy, is highly competitive. Buyers know that there are several sellers from which to choose, and sellers are aware that each of their products is similar to those offered by others. As a result, the price of ice cream and the quantity sold are determined not by any single buyer or seller but by all the buyers and sellers as they interact in the marketplace.

competitive market
a market in which there are many buyers and many sellers so each has a negligible impact on the market price

Economists use the term **competitive market** to describe a market in which there are so many buyers and sellers that each has little effect on the market price. Each seller has limited control over the price because many other sellers are offering similar products. A seller has little reason to charge less than the going price, and if the seller charges more, buyers will go elsewhere. Similarly, no single buyer can influence the price because each buyer purchases only a small amount.

In this chapter, we keep things simple by assuming that markets are **perfectly competitive**. In this ideal form of competition, a market has two characteristics: (1) The goods offered for sale are all exactly the same, and (2) the buyers and sellers are so numerous that no single buyer or seller has any influence over the market price. Because buyers and sellers in perfectly competitive markets must accept the price the market determines, they are said to be **price takers**. At the market price, buyers can buy all they want, and sellers can sell all they want.

There are some markets in which the assumption of perfect competition applies perfectly. In the wheat market, for example, there are thousands of farmers who sell wheat and millions of consumers who use wheat and wheat products. Because no single buyer or seller can influence the price of wheat, each takes the market price as given.

Not all goods and services are sold in perfectly competitive markets. For example, some markets have only one seller, and this seller sets the price. A seller in such a market is called a **monopoly**. A local cable television company, for instance, is a monopoly if residents of the town have only one company from which to buy cable service. Many other markets fall between the extremes of perfect competition and monopoly.

But perfectly competitive markets are a useful place to start. They are the easiest to analyze because everyone participating in them takes the price as given by market conditions. Moreover, because some degree of competition is present in most markets, many of the lessons learned studying supply and demand under perfect competition apply to more complex markets as well.

QuickQuiz

1. The best definition of a market is
   a. a store that offers a variety of goods and services.
   b. a place where buyers meet and an auctioneer calls out prices.
   c. a group of buyers and sellers of a good or service.
   d. a venue where the sole supplier of a good offers its product.
2. In a perfectly competitive market,
   a. each seller tries to distinguish itself by offering a better product than its rivals.
   b. each seller takes the price of its product as set by market conditions.
   c. each seller tries to undercut the prices charged by its rivals.
   d. one seller has successfully outcompeted its rivals, so no other sellers remain.
3. The market for which product best fits the definition of a perfectly competitive market?
   a. eggs
   b. tap water
   c. movies
   d. computer operating systems

Answers are at the end of the chapter.

4-2 Demand

Let's begin our study of markets by examining the behavior of buyers, particularly those who love ice cream. (And who doesn't?)

4-2a The Demand Curve: The Relationship between Price and Quantity Demanded

The **quantity demanded** of any good is the amount that buyers are willing and able to purchase. Many things determine the quantity demanded of a good, but one determinant plays a central role: its price. If the price of ice cream rose to $20 per scoop, most people would buy less. They might buy frozen yogurt instead. If the price of ice cream fell to $0.50 per scoop, they might buy more. This relationship between price and quantity demanded is true for most goods. In fact, it is so pervasive that economists call it the **law of demand**: Other things being equal, when the price of a good rises, the quantity demanded falls, and when the price falls, the quantity demanded rises.

quantity demanded
the amount of a good that buyers are willing and able to purchase

law of demand
the claim that, other things being equal, the quantity demanded of a good falls when the price of the good rises

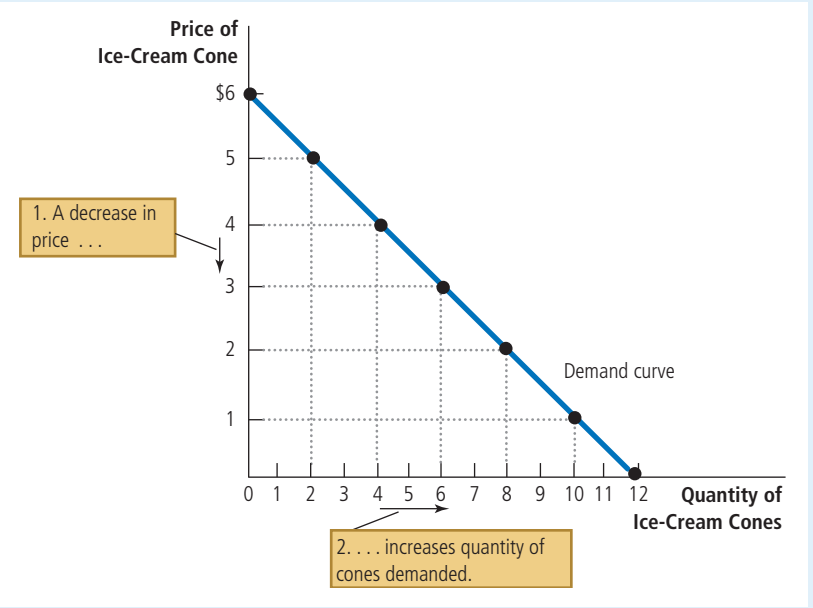
The table in Figure 1 shows how many ice-cream cones Catherine would buy each month at different prices. If ice-cream cones are free, Catherine buys 12 cones per month. At $1 per cone, she buys 10 per month. As the price rises further, she

Figure 1

Catherine's Demand Schedule and Demand Curve

The demand schedule is a table that shows the quantity demanded at each price. The demand curve, which graphs this schedule, illustrates how the quantity demanded changes as the price varies. Because a lower price increases the quantity demanded, the demand curve slopes downward.

Price of Ice-Cream Cone	Quantity of Cones Demanded
$0	12 cones
1	10
2	8
3	6
4	4
5	2
6	0



demand schedule
a table that shows the relationship between the price of a good and the quantity demanded

demand curve
a graph of the relationship between the price of a good and the quantity demanded

buys fewer and fewer cones. When the price reaches $6, Catherine doesn't buy any ice cream at all. This table is called a **demand schedule**. It shows the relationship between the price of a good and the quantity demanded, holding constant everything else that influences how much of the good a consumer wants to buy.

The graph in Figure 1 uses the numbers from the table to illustrate the law of demand. By convention, the price of ice cream is on the vertical axis, and the quantity demanded is on the horizontal axis. The line relating price and quantity demanded is the **demand curve**. It slopes downward because, other things being equal, a lower price increases the quantity demanded.

4-2b Market Demand versus Individual Demand

The demand curve in Figure 1 shows an individual's demand for a product. But to analyze how markets work, it's important to know the **market demand**, the sum of all the individual demands for a particular good or service.

The table in Figure 2 shows the demand schedules for ice cream for two people—Catherine and Nicholas. At any price, Catherine's demand schedule shows how many cones she buys, and Nicholas's shows the same information for him. The market demand at each price is the sum of their individual demands.

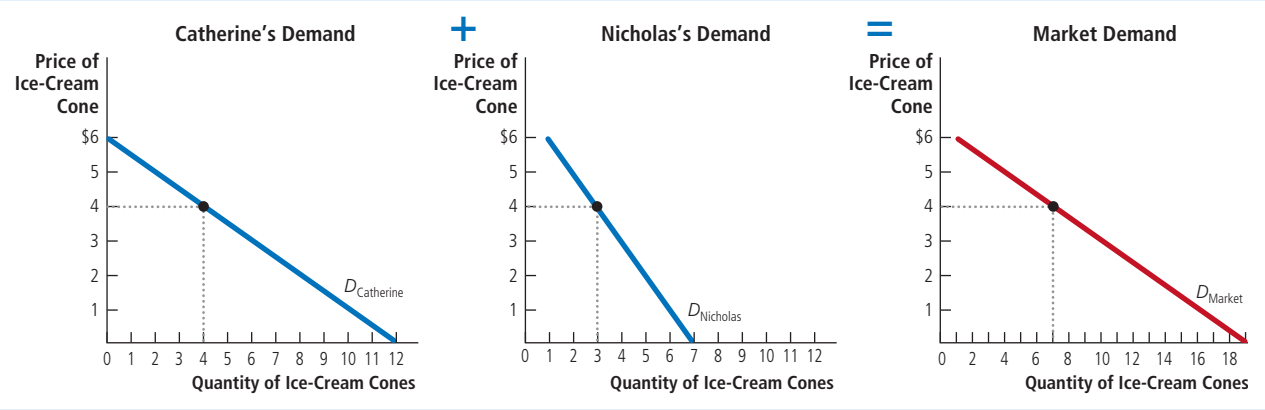
The graph in Figure 2 shows the demand curves for these demand schedules. To obtain the market demand curve, we add the individual demand curves **horizontally**. That is, to find the total quantity demanded at any price, we add the individual quantities demanded, which are found on the horizontal axis of

Figure 2

Market Demand as the Sum of Individual Demands

The quantity demanded in a market is the sum of the quantities demanded by all the buyers at each price. Thus, the market demand curve is found by adding the individual demand curves horizontally. At a price of $4, Catherine demands 4 ice-cream cones, and Nicholas demands 3, so the quantity demanded in the market at this price is 7 cones.

Price of Ice-Cream Cone	Catherine		Nicholas		Market
$0	12	+	7	=	19 cones
1	10		6		16
2	8		5		13
3	6		4		10
4	4		3		7
5	2		2		4
6	0		1		1



the individual demand curves. The market demand curve is crucial for analyzing how markets function. It shows how the total quantity demanded of a good varies as its price changes, holding constant all the other factors that affect consumer purchases.

4-2c Shifts in the Demand Curve

Because the market demand curve holds other things constant, it need not be stable over time. If something happens to alter the quantity demanded at any given price, the demand curve shifts.

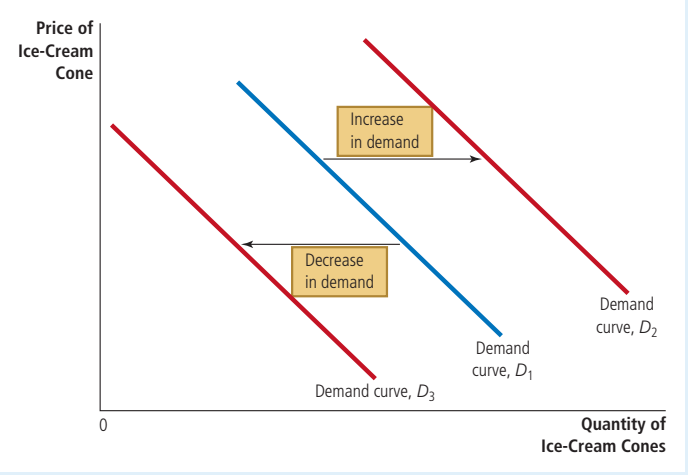
For example, suppose the American Medical Association discovers that people who regularly eat ice cream live longer, healthier lives. Such a marvelous discovery would raise the demand for ice cream. At any price, buyers would now want to purchase more ice cream, and the demand curve for ice cream would shift.

Figure 3 illustrates shifts in demand. A change that increases the quantity demanded at every price, such as this wondrous but imaginary discovery, shifts the demand

Figure 3

Shifts in the Demand Curve

A change that increases the quantity that buyers want to purchase at any price shifts the demand curve to the right. A change that decreases the quantity that buyers want to purchase at any price shifts the demand curve to the left.



curve to the right and is called an **increase in demand**. A change that reduces the quantity demanded at every price shifts the demand curve to the left and is called a **decrease in demand**.

Changes in many variables can shift the demand curve, including:

Income What would happen to your demand for ice cream if you lost your job one summer? It would most likely fall because you have less money to spend on things like ice cream. If the demand for something falls when income falls, that good is called a **normal good**.

normal good
a good for which, other things being equal, an increase in income leads to an increase in demand

Normal goods are the norm, but not all goods are normal goods. If the demand for something rises when income falls, that good is called an **inferior good**. An example of an inferior good might be bus rides. As your income falls, you are less likely to buy a car or take an Uber and more likely to ride a bus.

inferior good
a good for which, other things being equal, an increase in income leads to a decrease in demand

Prices of Related Goods Suppose that the price of frozen yogurt declines. The law of demand says that you will buy more of it. At the same time, you may buy less ice cream. Because ice cream and frozen yogurt are both cold, sweet, creamy desserts, they satisfy similar cravings. When a fall in the price of one good, like frozen yogurt, reduces the demand for another good, like ice cream, the two goods are called **substitutes**. Substitutes are often pairs of goods that are used in place of each other, such as hot dogs and hamburgers, sweaters and sweatshirts, and movie tickets and video streaming services.

substitutes
two goods for which an increase in the price of one leads to an increase in the demand for the other

Now suppose that the price of hot fudge declines. According to the law of demand, you will buy more hot fudge. Yet in this case, you may be inclined to buy more ice cream as well because ice cream and hot fudge go well together. When a fall in the price of one good, like hot fudge, raises the demand for another good, like ice cream, the two goods are called **complements**. Complements are often pairs of goods that are used together, such as electricity and air conditioners, computers and software, and peanut butter and jelly.

complements
two goods for which an increase in the price of one leads to a decrease in the demand for the other

Table 1

Variables That Influence Buyers

This table lists the variables that affect how much of any good consumers choose to buy. Notice the special role that the price of the good plays: A change in that price represents a movement along the demand curve, while a change in one of the other variables shifts the curve.

Variable	A Change in This Variable . . .
Price of the good itself	Represents a movement along the demand curve
Income	Shifts the demand curve
Prices of related goods	Shifts the demand curve
Tastes	Shifts the demand curve
Expectations	Shifts the demand curve
Number of buyers	Shifts the demand curve

Tastes If you like pistachio ice cream, you will buy more of it. While individual tastes, like preferences for ice-cream flavors, are critically important for explaining demand, economists typically don't try to explain them. This is because they are unique to you, though affected by historical and psychological forces. Economists do, however, examine what happens when tastes change.

Expectations Your views about the future may affect your demand for something today. If you expect a higher income next month, you may choose to save less now and spend more on ice cream today. If you believe ice cream will be cheaper tomorrow, you may be reluctant to buy a cone at today's price.

Number of Buyers In addition to the factors that influence the behavior of individual buyers, market demand depends on how many of these buyers there are. If Peter were to join Catherine and Nicholas as an ice-cream consumer, the quantity demanded would be higher at every price, and market demand would increase.

Summary The demand curve shows what happens to the quantity demanded of a good as its price varies, holding constant all the other variables that influence buyers. When one of these other variables changes, the quantity demanded at each price changes, and the demand curve shifts. Table 1 lists the variables that influence how much of a good consumers choose to buy.

If you have trouble remembering whether you need to shift or move along the demand curve, it helps to recall a lesson from the appendix to Chapter 2. A curve shifts when there is a change in a relevant variable that is not measured on either axis. Because the price is on the vertical axis, a change in price represents a movement along the demand curve. By contrast, income, the prices of related goods, tastes, expectations, and the number of buyers are not measured on either axis, so a change in one of these variables shifts the demand curve.

PABLO DEL RIO SOTELO/SHUTTERSTOCK.COM

Case Study

Two Ways to Reduce Smoking

Because smoking can harm you and those around you, policymakers often want to reduce the amount that people smoke. Consider two paths for achieving this goal.

One way to reduce smoking is to shift the demand curve for cigarettes and other tobacco products. This can be done through public service announcements, mandatory health warnings on cigarette packages, and the prohibition of cigarette advertising on television, all of which are aimed at reducing the quantity of cigarettes demanded at any price. When successful, these policies shift the demand curve for cigarettes to the left, as in panel (a) of Figure 4.

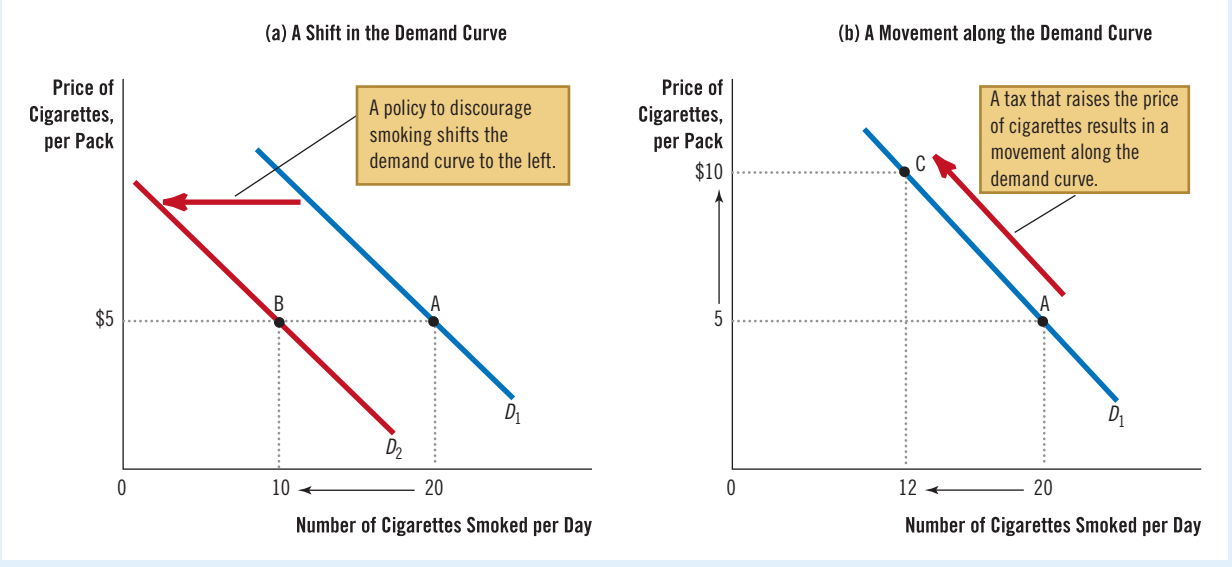
Another way to discourage smoking is to raise the price of cigarettes. When the government taxes cigarettes, the companies that make and sell them pass most of the tax on to consumers in the form of higher prices. Because people tend to buy less when the price rises, this policy also reduces smoking. But this approach does not shift the demand curve. Instead, the change appears as a movement along the same curve to a point with a higher price and lower quantity, as in panel (b) of Figure 4.

How much does the amount of smoking respond to changes in the price of cigarettes? Economists have studied what happens when the tax on cigarettes changes. They have found that a ten percent price increase causes a four percent reduction in the quantity demanded. Teenagers are especially sensitive to the price of cigarettes: A ten percent price increase causes a 12 percent drop in teenage smoking.

Figure 4

Shifts in the Demand Curve versus Movements along the Demand Curve

When warnings on cigarette packages persuade smokers to smoke less, the demand curve for cigarettes shifts to the left. In panel (a), the curve shifts from D_1 to D_2. At a price of $5 per pack, the quantity demanded falls from 20 to 10 cigarettes per day, as reflected by the shift from point A to point B. By contrast, when a tax raises the price of cigarettes, the demand curve does not shift. Instead, there is a movement to a different point on the demand curve. In panel (b), when the price rises from $5 to $10, the quantity demanded falls from 20 to 12 cigarettes per day, as reflected by the movement from point A to point C.



A related question is how the price of cigarettes affects the demand for other products, such as marijuana. Opponents of cigarette taxes sometimes argue that tobacco and marijuana are substitutes, so high cigarette prices encourage marijuana use. By contrast, many experts on substance abuse view tobacco as a "gateway drug," leading young people to experiment with other harmful substances. Most studies of the data are consistent with this latter view. They find that lower cigarette prices are associated with greater use of marijuana. In other words, tobacco and marijuana appear to be complements rather than substitutes. ●

QuickQuiz

4. A change in which of the following will NOT shift the demand curve for hamburgers?
   a. the price of hot dogs
   b. the price of hamburgers
   c. the price of hamburger buns
   d. the income of hamburger consumers
5. Which of the following will shift the demand curve for pizza to the right?
   a. an increase in the price of hamburgers, a substitute for pizza
   b. an increase in the price of root beer, a complement to pizza
   c. the departure of college students when they leave for summer vacation
   d. a decrease in the price of pizza
6. If pasta is an inferior good, then the demand curve shifts to the _____ when _____ rises.
   a. right; the price of pasta
   b. right; consumers' income
   c. left; the price of pasta
   d. left; consumers' income

Answers are at the end of the chapter.

4-3 Supply

Buyers are only half the story of how markets work. Sellers are the other half. Let's now consider the sellers of ice cream.

4-3a The Supply Curve: The Relationship between Price and Quantity Supplied

The **quantity supplied** of any good or service is the amount that sellers are willing and able to sell. There are many determinants of the quantity supplied, but again, price plays a special role. When the price of ice cream is high, selling ice cream is very profitable, so the quantity supplied is large. Sellers work long hours, buy many ice-cream machines, and hire many workers. By contrast, when the price is low, the business is less profitable, so sellers produce less. Some sellers may even shut down, reducing their quantity supplied to zero. This relationship between price and the quantity supplied is called the **law of supply**: Other things being equal, when the price of a good rises, the quantity supplied also rises, and when the price falls, the quantity supplied falls as well.

The table in Figure 5 shows the quantity of ice-cream cones supplied each month by Ben, an ice-cream seller, at various prices of ice cream. At a price below $2, Ben does not supply any ice cream at all. As the price rises, he supplies a greater and greater quantity. This table is called the **supply schedule**. It shows the relationship between the price of a good and the quantity supplied, holding constant everything else that influences how much of the good producers want to sell.

The graph in Figure 5 uses numbers from the table to illustrate the law of supply. The curve relating price and the quantity supplied is the **supply curve**. The supply

quantity supplied
the amount of a good that sellers are willing and able to sell

law of supply
the claim that, other things being equal, the quantity supplied of a good rises when the price of the good rises

supply schedule
a table that shows the relationship between the price of a good and the quantity supplied

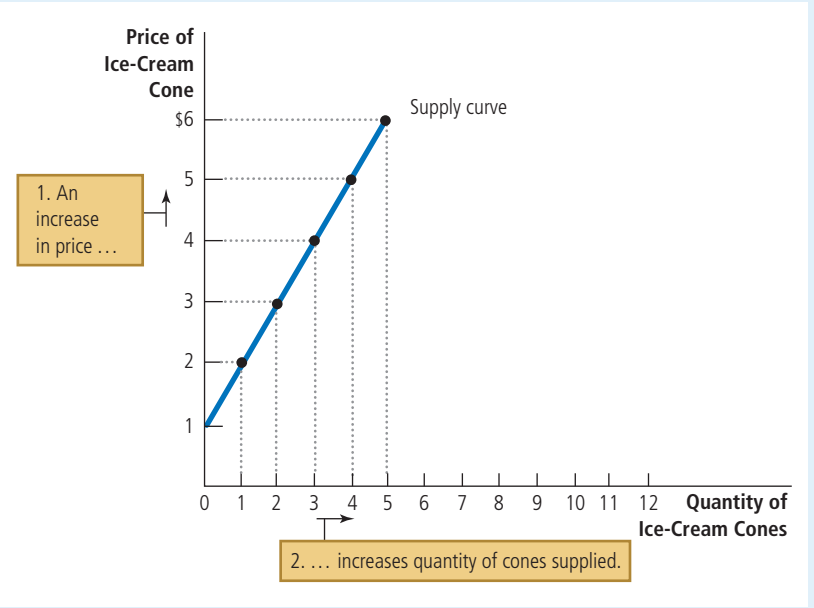
supply curve
a graph of the relationship between the price of a good and the quantity supplied

Figure 5

Ben's Supply Schedule and Supply Curve

The supply schedule is a table that shows the quantity supplied at each price. The supply curve, which graphs the supply schedule, illustrates how the quantity supplied changes as a good's price varies. Because a higher price increases the quantity supplied, the supply curve slopes upward.

Price of Ice-Cream Cone	Quantity of Cones Supplied
$0	0 cones
1	0
2	1
3	2
4	3
5	4
6	5



curve slopes upward because, other things being equal, a higher price means a greater quantity supplied.

4-3b Market Supply versus Individual Supply

Just as market demand is the sum of the demands of all buyers, market supply is the sum of the supplies of all sellers. The table in Figure 6 shows the supply schedules for the market's two ice-cream producers—Ben and Jerry. At any price, Ben's supply schedule tells us the quantity that Ben supplies, and Jerry's supply schedule tells us how much Jerry supplies. The market supply is the sum of the two individual supplies.

The graph in Figure 6 shows the supply curves that correspond to the supply schedules. As with demand curves, the market supply curve is obtained by adding the individual supply curves **horizontally**. That is, to find the total quantity supplied at any price, we add the individual quantities, which are located on the horizontal axis of the individual supply curves. The market supply curve shows how the total quantity supplied varies as the price varies, holding constant all other factors that influence producers' decisions about how much to sell.

4-3c Shifts in the Supply Curve

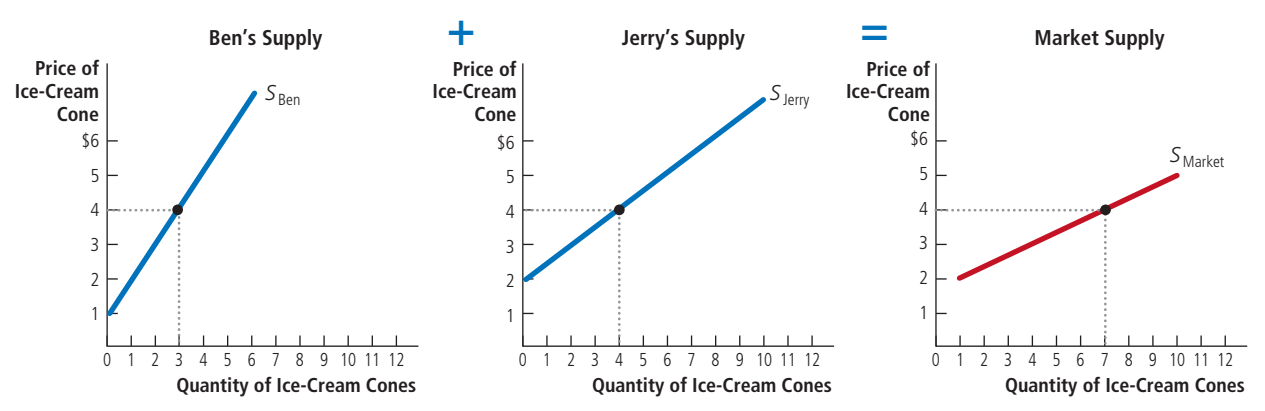
Because a market supply curve holds constant all the variables other than price that affect quantity supplied, it can move over time. When one of these other variables

Figure 6

Market Supply as the Sum of Individual Supplies

The quantity supplied in a market is the total quantity supplied by all sellers at each price. You can build the market supply curve by adding the individual supply curves horizontally. At a price of $4, Ben supplies 3 ice-cream cones, and Jerry supplies 4 ice-cream cones, so the total quantity supplied at the price of $4 is 7 cones.

Price of Ice-Cream Cone	Ben		Jerry		Market
$0	0	+	0	=	0 cones
1	0		0		0
2	1		0		1
3	2		2		4
4	3		4		7
5	4		6		10
6	5		8		13



changes, the quantity that producers want to sell at any price changes, and the supply curve shifts.

For example, suppose the price of sugar falls. Because sugar is an input in the production of ice cream, the lower price of sugar makes selling ice cream more profitable. This increases the ice-cream supply: At any price, sellers are willing to produce more. As a result, the supply curve shifts to the right.

Figure 7 illustrates shifts in supply. A change that raises the quantity supplied at every price, such as a fall in the price of sugar, shifts the supply curve to the right and is called an **increase in supply**. A change that reduces the quantity supplied at every price shifts the supply curve to the left and is called a **decrease in supply**.

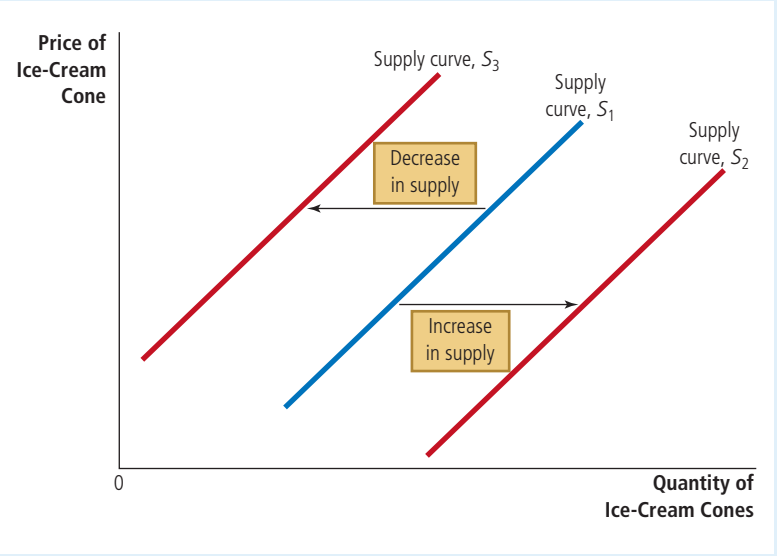
Many variables can shift the supply curve. The most important ones include:

Input Prices Ice-cream sellers use various inputs to make their product: cream, sugar, flavoring, ice-cream machines, the buildings in which the ice cream is made, and the labor of the workers who mix the ingredients and operate the machines. When the price of one or more of these inputs rises, producing ice cream becomes

Figure 7

Shifts in the Supply Curve

A change that raises the quantity that sellers want to produce at any price shifts the supply curve to the right. A change that lowers the quantity that sellers want to produce at any price shifts the supply curve to the left.



less profitable, and firms supply less ice cream. If input prices rise substantially, a firm might shut down and supply no ice cream at all. Thus, the supply of a good moves in the opposite direction of the prices of inputs.

Technology The technology for turning inputs into output is another determinant of supply. The invention of mechanized ice-cream machines, for example, reduced the labor needed to make ice cream. By reducing producers' costs, this advance in technology increased the supply. In the long run, such changes in technology are among the most potent forces affecting market outcomes.

Expectations The amount that ice-cream makers supply may depend on their expectations about the future. For example, if they expect the price to rise, they may put some of their current production into storage and supply less to the market today.

Number of Sellers In addition to the factors that influence the behavior of individual sellers, market supply depends on how many sellers there are in the market. If Ben or Jerry retires from the ice-cream business, the market supply falls. If Edy starts a new ice-cream business, the market supply rises.

Summary The supply curve shows what happens to the quantity supplied when a good's price varies, holding constant all the other variables that influence sellers. When one of these other variables changes, the quantity supplied at each price changes, and the supply curve shifts. Table 2 lists the variables that influence how much producers choose to sell.

Once again, to remember whether you need to shift or move along the supply curve, keep this in mind: A curve shifts only when there is a change in a relevant variable that isn't named on either axis. Price is on the vertical axis, so a change in price is represented by a movement along the supply curve. By contrast, because input prices, technology, expectations, and the number of sellers are not measured on either axis, a change in one of these variables shifts the supply curve.

Table 2

Variables That Influence Sellers

This table lists the variables that affect how much of any good producers choose to sell. Notice the special role that the price of the good plays: A change in that price represents a movement along the supply curve, while a change in one of the other variables shifts the curve.

Variable	A Change in This Variable . . .
Price of the good itself	Represents a movement along the supply curve
Input prices	Shifts the supply curve
Technology	Shifts the supply curve
Expectations	Shifts the supply curve
Number of sellers	Shifts the supply curve

QuickQuiz

7. What event moves pizza suppliers up along a given supply curve?
   a. an increase in the price of pizza
   b. an increase in the price of root beer, a complement to pizza
   c. a decrease in the price of cheese, an input to pizza
   d. a kitchen fire that destroys a popular pizza joint
8. What event shifts the supply curve for pizza to the right?
   a. an increase in the price of pizza
   b. an increase in the price of root beer, a complement to pizza
   c. a decrease in the price of cheese, an input to pizza
   d. a kitchen fire that destroys a popular pizza joint
9. Movie tickets and video streaming services are substitutes. If the price of video streaming increases, what happens in the market for movie tickets?
   a. The supply curve shifts to the left.
   b. The supply curve shifts to the right.
   c. The demand curve shifts to the left.
   d. The demand curve shifts to the right.

Answers are at the end of the chapter.

4-4 Supply and Demand Together

Let's now combine supply and demand to see how they determine the price and quantity of a good sold in a market.

4-4a Equilibrium

Figure 8 shows the market supply curve and market demand curve together. Notice that there is one point at which the supply and demand curves intersect. This point is the market's **equilibrium**. The price at this intersection is the **equilibrium price**, and the quantity is the **equilibrium quantity**. Here, the equilibrium price is $4.00 per cone, and the equilibrium quantity is 7 cones.

The dictionary defines **equilibrium** as a situation in which forces are in balance. This sense of balance is key to the concept of a market equilibrium. **At the equilibrium price, the quantity of the good that buyers are willing and able to buy exactly balances the quantity that sellers are willing and able to sell.** The equilibrium price is sometimes called the **market-clearing price** because, at this price, everyone in the market has been satisfied: Buyers have bought all they want to buy, and sellers have sold all they want to sell.

equilibrium
a situation in which the market price has reached the level at which the quantity supplied equals the quantity demanded

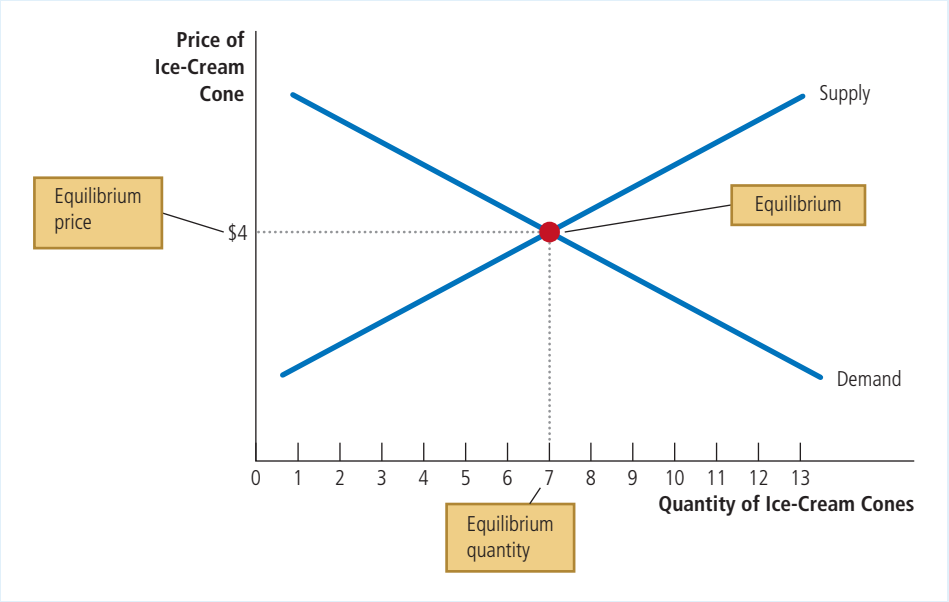
equilibrium price
the price that balances the quantity supplied and the quantity demanded

equilibrium quantity
the quantity supplied and the quantity demanded at the equilibrium price

Figure 8

The Equilibrium of Supply and Demand

The market's equilibrium is where the supply and demand curves intersect. At the equilibrium price, the quantity supplied equals the quantity demanded. Here, the equilibrium price is $4. At this price, 7 ice-cream cones are supplied, and 7 are demanded.



surplus
a situation in which the quantity supplied is greater than the quantity demanded

shortage
a situation in which the quantity demanded is greater than the quantity supplied

The actions of buyers and sellers move markets toward the equilibrium of supply and demand. To see why, consider what happens when the market price does not equal the equilibrium price.

Suppose first that the market price is above the equilibrium price, as in panel (a) of Figure 9. At a price of $5 per cone, the quantity supplied (10 cones) exceeds the quantity demanded (4 cones). There is a **surplus** of the good: Producers are unable to sell all they want at the going price. A surplus is sometimes called a situation of **excess supply**. When there is a surplus in the ice-cream market, sellers find their freezers increasingly full of ice cream they would like to sell but cannot. They respond by cutting prices. Falling prices, in turn, increase the quantity demanded and decrease the quantity supplied. These changes represent movements **along** the supply and demand curves, not shifts in the curves. Prices continue to fall until the market reaches the equilibrium.

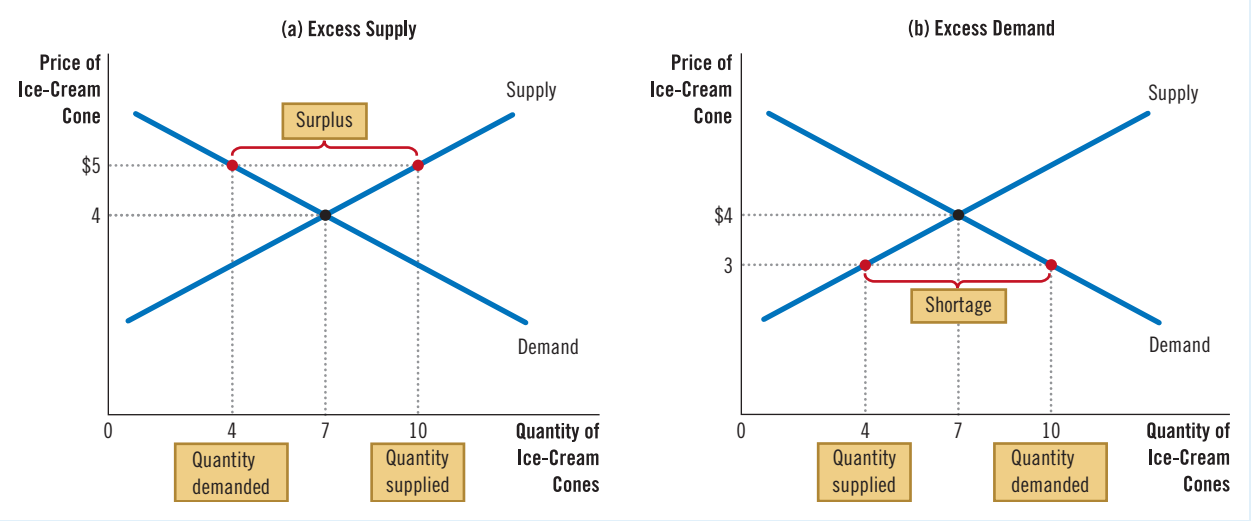
Suppose now that the market price is below the equilibrium price, as in panel (b) of Figure 9. In this case, the price is $3 per cone, and the quantity demanded exceeds the quantity supplied. There is a **shortage** of the good: Consumers are unable to buy all they want at the going price. A shortage is sometimes called a situation of **excess demand**. When a shortage occurs in the ice-cream market, buyers must wait in long lines for a chance to buy one of the few cones available. With too many buyers chasing too few goods, sellers can raise prices without losing sales. These price increases cause the quantity demanded to fall and the quantity supplied to rise. Again, these changes are depicted as movements **along** the supply and demand curves, and they move the market closer to equilibrium.

Regardless of where the price starts, the activities of buyers and sellers push the market price toward equilibrium. Once the market reaches equilibrium, all buyers and sellers are satisfied in the sense that they can buy and sell the amount they want at the going price. At that point, there is no further upward or downward

Figure 9

Markets Not in Equilibrium

In panel (a), there is a surplus. Because the market price of $5 is above the equilibrium price, the quantity supplied (10 cones) exceeds the quantity demanded (4 cones). Producers try to increase sales by cutting the price, moving it toward its equilibrium level. In panel (b), there is a shortage. Because the market price of $3 is below the equilibrium price, the quantity demanded (10 cones) exceeds the quantity supplied (4 cones). With too many buyers chasing too few goods, producers raise the price. In both cases, the price adjustment moves the market toward the equilibrium of supply and demand.

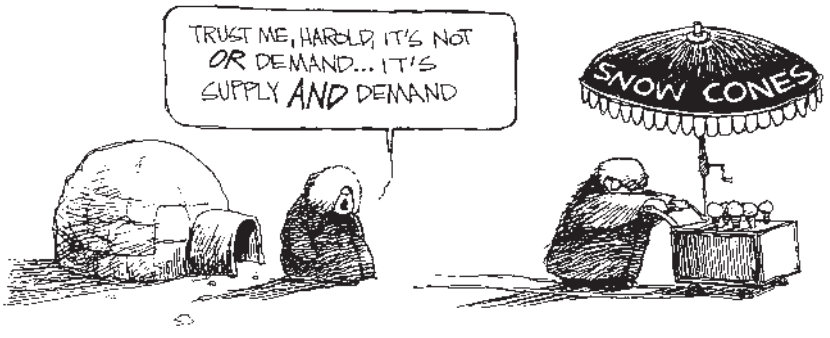


pressure on the price. How quickly equilibrium is reached varies from market to market depending on how quickly prices adjust. In most well-functioning markets, surpluses and shortages are only temporary because prices quickly move toward their equilibrium levels. This phenomenon is so pervasive that it is called the **law of supply and demand**: The price of any good adjusts to bring the quantity supplied and the quantity demanded into balance.

law of supply and demand
the claim that the price of any good adjusts to bring the quantity supplied and the quantity demanded of that good into balance

4-4b Three Steps to Analyzing Changes in Equilibrium

Supply and demand together determine a market's equilibrium, which in turn determines the price and quantity of the good that buyers purchase and sellers



produce. The equilibrium price and quantity depend on the positions of the supply and demand curves. When an event shifts one of these curves, the equilibrium changes, resulting in a new price and a new quantity exchanged between buyers and sellers.

When analyzing how an event affects the market's equilibrium, we proceed in three steps. First, we decide if the event shifts the supply curve, the demand curve, or both. Second, we decide whether the curve shifts to the right or to the left. Third, we use a supply-and-demand diagram to compare the initial equilibrium with the new one, which shows how the shift affects the equilibrium price and quantity. Table 3 summarizes these three steps. To see how this works, let's consider a few events that might affect the market for ice cream.

Example: A Shift in Demand Changes the Market Equilibrium Suppose that this summer's weather is exceptionally hot. How does this affect the ice-cream market? To answer this question, let's follow our three steps.

1. The weather affects the demand curve by changing consumers' taste for ice cream. That is, it alters the amount that people want to buy at any price. The supply curve remains the same because the weather does not directly affect the firms that sell ice cream.
2. Because hot weather makes a cool treat more appealing, people want more ice cream. Figure 10 shows this increase in demand as a rightward shift in the demand curve from D_1 to D_2. This shift indicates that the quantity demanded is higher at every price.
3. At the old price of $4, there is now an excess demand for ice cream, and this shortage induces firms to raise the price. As Figure 10 shows, the increase in demand raises the equilibrium price from $4 to $5 and the equilibrium quantity from 7 to 10 cones. In other words, the hot weather increases both the price of ice cream and the quantity sold.

Shifts in Curves versus Movements along Them When hot weather increases the demand for ice cream and drives up the price, the quantity that ice-cream makers supply rises, even though the supply curve remains the same. In this case, economists say there has been an increase in the quantity supplied but no change in supply.

Supply refers to the position of the supply curve, while the **quantity supplied** refers to the amount producers want to sell. In the summer heat, supply does not change because the weather does not affect how much producers want to sell at any price. Instead, the sultry weather makes consumers more eager to buy at any price,

Table 3

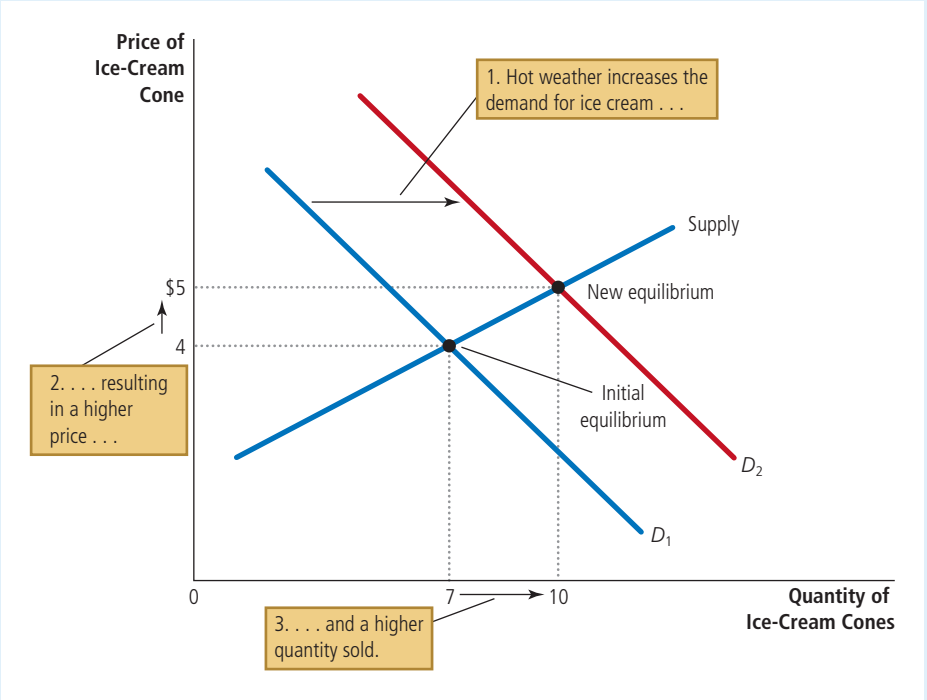
Three Steps for Analyzing Changes in Equilibrium

1. Decide if the event shifts the supply or demand curve (or perhaps both).
2. Decide in which direction the curve shifts.
3. Use a supply-and-demand diagram to see how the shift changes the equilibrium price and quantity.

Figure 10

How an Increase in Demand Affects the Equilibrium

An event that raises the quantity demanded at any price shifts the demand curve to the right. The equilibrium price and quantity both rise. Here, an abnormally hot summer causes buyers to demand more ice cream. The demand curve shifts from D_1 to D_2, causing the equilibrium price to increase from \$4 to \$5 and the equilibrium quantity to increase from 7 to 10 cones.



shifting the demand curve to the right. The increase in demand causes the equilibrium price to rise. When the price rises, the quantity supplied rises. This increase in quantity supplied is represented by the movement along the supply curve.

To summarize, a shift **in** the supply curve is called a "change in supply," and a shift **in** the demand curve is called a "change in demand." A movement **along** a fixed supply curve is called a "change in the quantity supplied," and a movement **along** a fixed demand curve is called a "change in the quantity demanded."

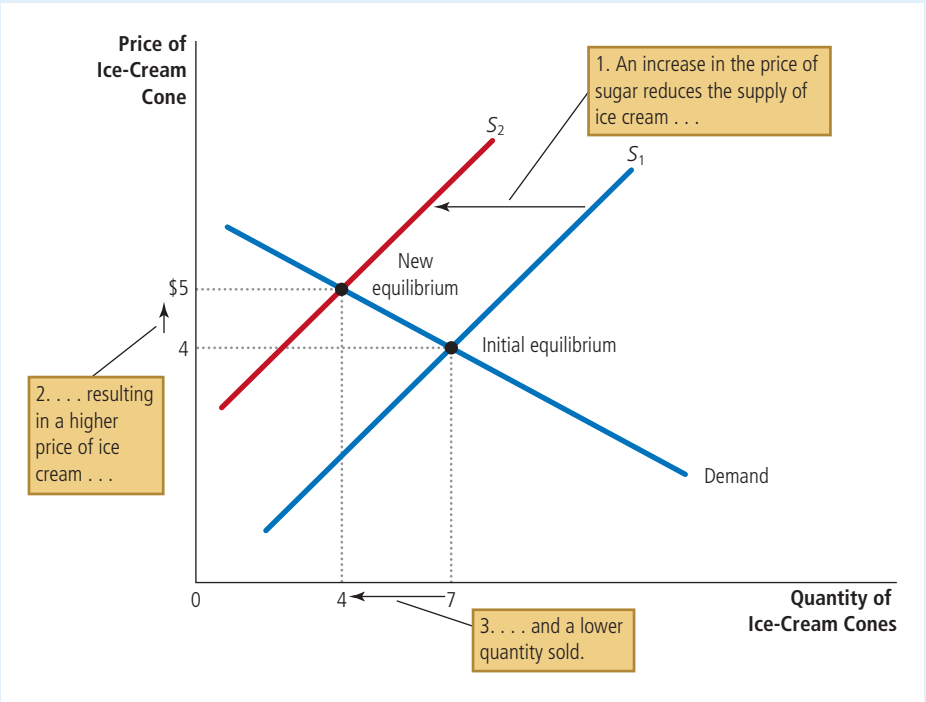
Example: A Shift in Supply Changes the Market Equilibrium One August, a hurricane destroys part of the sugarcane crop and drives up the price of sugar. How does this affect the market for ice cream? Again, we follow our three steps.

1. The increase in the price of sugar, an input for ice cream, raises the cost of producing ice cream. It therefore affects the supply curve. The demand curve does not change because the higher cost of inputs does not directly affect the amount of ice cream consumers want to buy.
2. Higher costs reduce the amount of ice cream that producers are willing and able to sell at every price. Figure 11 depicts this decrease in supply as a leftward shift in the supply curve from S_1 to S_2.
3. At the old price of \$4, there is now an excess demand for ice cream, and this shortage causes firms to raise the price. As Figure 11 shows, the shift in the supply curve raises the equilibrium price from \$4 to \$5 and lowers the equilibrium quantity from 7 to 4 cones. Because of the sugar price increase, the price of ice cream rises, and the quantity sold falls.

Figure 11

How a Decrease in Supply Affects the Equilibrium

An event that reduces the quantity supplied at any price shifts the supply curve to the left. The equilibrium price rises, and the equilibrium quantity falls. Here, an increase in the price of sugar (an input) causes sellers to supply less ice cream. The supply curve shifts from S_1 to S_2, causing the equilibrium price of ice cream to rise from $4 to $5 and the equilibrium quantity to fall from 7 to 4 cones.



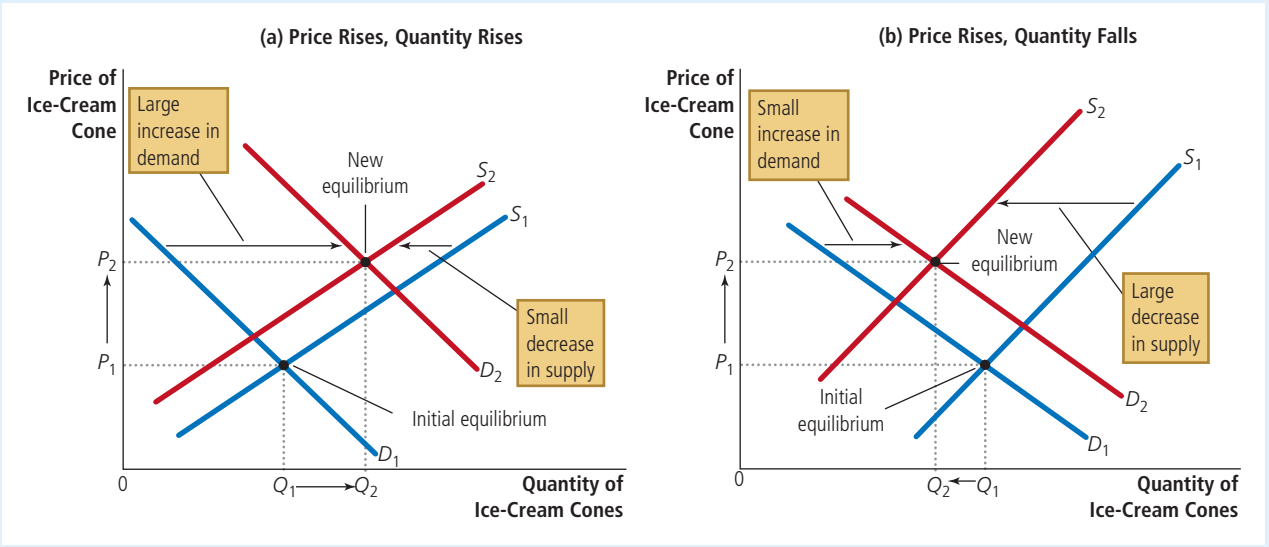
Example: Both Supply and Demand Shift In a series of unfortunate events, a heat wave and a hurricane strike in the same summer. To analyze this nasty combination, we turn again to our three steps.

1. Both curves must shift. The heat affects the demand curve because it alters the amount of ice cream that consumers want to buy at any price. At the same time, the hurricane alters the supply curve for ice cream: By driving up sugar prices, it changes the amount of ice cream that producers want to sell at any price.
2. The curves shift in the same directions as they did earlier: The demand curve shifts to the right, and the supply curve shifts to the left, as Figure 12 shows.
3. Two outcomes are possible, depending on the relative size of the demand and supply shifts. In both cases, the equilibrium price rises. In panel (a), where demand increases substantially while supply falls just a little, the equilibrium quantity also increases. But in panel (b), where supply falls substantially while demand rises just a little, the equilibrium quantity falls. Thus, these events certainly raise the price of ice cream, but their impact on the amount of ice cream sold is ambiguous (that is, it could go either way).

Figure 12

A Shift in Both Supply and Demand

A simultaneous increase in demand and decrease in supply yields two possible outcomes. In panel (a), the equilibrium price rises from P_1 to P_2, and the equilibrium quantity rises from Q_1 to Q_2. In panel (b), the equilibrium price again rises from P_1 to P_2, but the equilibrium quantity falls from Q_1 to Q_2.



Summary Supply and demand curves help to analyze a change in equilibrium. When an event shifts the supply curve, the demand curve, or perhaps both curves, these tools can predict how the event will alter the price and quantity sold in equilibrium. Table 4 shows the predicted outcome for any combination of shifts in the two curves. To ensure that you understand how to use the tools of supply and demand, pick a few of the table's entries and make sure you can explain to yourself the stated predictions.

Table 4

What Happens to Price and Quantity When Supply or Demand Shifts?

As a quick quiz, make sure you can explain at least a few of the entries in this table using a supply-and-demand diagram.

	No Change in Supply	An Increase in Supply	A Decrease in Supply
No Change in Demand	*P* same *Q* same	*P* down *Q* up	*P* up *Q* down
An Increase in Demand	*P* up *Q* up	*P* ambiguous *Q* up	*P* up *Q* ambiguous
A Decrease in Demand	*P* down *Q* down	*P* down *Q* ambiguous	*P* ambiguous *Q* down

In the News

Price Increases after Disasters

When a disaster strikes, many goods experience an increase in demand or a decrease in supply, putting upward pressure on prices. Not everyone thinks that's fair.

The Law of Supply and Demand Isn't Fair

By Richard Thaler

For an economist, one of the most jarring sights during the early weeks of the coronavirus crisis in the United States was the spectacle of bare shelves in sections of the supermarket.

There was no toilet paper or hand sanitizer. Pasta, flour and even yeast could be hard to find in the early weeks of social distancing as many people decided to take up baking. Of far greater concern, hospitals could not buy enough of the masks, gowns and ventilators required to safely treat Covid-19 patients.

What happened to the laws of supply and demand? Why didn't prices rise enough to clear the market, as economic models predict?

A paper that I wrote with my friends Daniel Kahneman, a psychologist, and Jack Knetsch, an economist, explored this problem. We found that the answer may be summed up with a single word, one you won't find in the standard supply-and-demand models: fairness. Basically, it just isn't socially acceptable to raise prices in an emergency.

We asked people questions about the actions of hypothetical firms. For example: "A hardware store has been selling snow shovels for $15. The morning after a blizzard, the store raises the price of snow shovels to $20."

Fully 82 percent of our respondents judged this to be unfair. The respondents were Canadians, known for their politeness, but the general findings have now been replicated and confirmed in studies around the world.

Most companies implicitly understand that abiding by the social norms of fairness should be part of their business model. In the current crisis, large retail chains have responded to the shortages of toilet paper not by raising the price but by limiting the amount each customer can buy. And Amazon and eBay prohibited what was viewed as price gouging on their sites.

We have seen similar behavior after hurricanes. As soon as a storm ends, there is typically enormous demand for goods like bottled water and plywood. Big retailers like Home Depot and Walmart anticipate this, sending trucks loaded with supplies to regions just outside the danger zone, ready to be deployed. Then, when it is safe, the stores provide water for free and sell the plywood at the list price **or lower**.

At the same time, some "entrepreneurs" are likely to behave differently. They see a disaster as an opportunity and so will fill up trucks with plywood near their homes, drive to the storm site and sell their goods for whatever price they can get.

It is not that large retailers are intrinsically more ethical than the entrepreneurs; it is simply that they have different time horizons. The large companies are playing a long game, and by behaving "fairly," they are hoping to retain customer loyalty after the emergency. The entrepreneurs are just interested in a quick buck.

Fairness norms help explain the breakdown of supply chains of medical equipment in the coronavirus crisis. Hospitals normally use buying associations that make long-term deals with wholesalers to provide essential supplies. The wholesalers generally want to preserve these relationships and realize that now would not be a good time to raise prices. Often, they are contractually obligated to supply items at prices negotiated before a spike in demand.

One current example is the N95 face mask. At the onset of the pandemic, hospitals had long-term contracts to buy them for about 35 cents each, an executive at a New York hospital told me. When the need for the masks surged, these suppliers were not allowed to raise the price, even if inclined to do so.

QuickQuiz

10. The discovery of a large new reserve of crude oil will shift the _____ curve for gasoline, leading to a _____ equilibrium price.
    a. supply; higher
    b. supply; lower
    c. demand; higher
    d. demand; lower

11. If the economy goes into a recession and incomes fall, what happens in the markets for inferior goods?
    a. Prices and quantities both rise.
    b. Prices and quantities both fall.
    c. Prices rise, and quantities fall.
    d. Prices fall, and quantities rise.

But others along the supply chain could make big profits by diverting masks to anyone willing to pay top dollar. That left hospitals in a bind. As the coronavirus spread in New York, the executive's hospital searched frantically for masks, eventually paying an overseas supplier $6 each, for hundreds of thousands of them, when the regular stock was desperately short.

When anyone tries to reap big profits in an emergency like this, it can look ugly. Consider the case of two brothers who began buying hand sanitizer, masks and other scarce commodities on March 1, the day of the first announcement of a Covid-19 death in the United States. After they sold some of their merchandise at big markups on Amazon and eBay, these outlets cut them off. Eventually, after considerable adverse publicity, the brothers decided to donate their supplies.

Notice that the brothers were making markets more "efficient," by buying low and selling high. If instead of arbitraging coronavirus supplies they had sold shares of airline and hotel companies and bought shares of Netflix and Zoom, they would simply have been considered smart traders. But while smart trading may be fine for investments, it is not considered fair when it involves essential goods during a pandemic.

One can argue that this social norm is harmful in that it prevents markets from doing their magic. For example, Tyler Cowen, the George Mason University economist, has said he wishes it were OK to raise prices for coronavirus essentials.

TAB62/SHUTTERSTOCK.COM

How much would you pay for this in an emergency?

"Higher prices discourage panic buying and increase the chance that the people who truly need particular goods and services have a greater chance of getting them," he wrote.

But which people "truly need" N95 masks? What is the right allocation of masks among well-endowed research hospitals, poorly funded municipal facilities, nursing homes and food processing plants? Supply and demand would tell us that the masks should simply go to the buyer who was willing and able to pay the most for them. But fairness tells us this can't be the only consideration.

As a practical matter for businesses, big and small, that want to keep operating for the long haul, it makes good sense to obey the law of fairness. If the next shortage is meat and a store owner realizes that there is only one package of pork chops left, it would be unwise to sell it at auction to the highest bidder. ■

Questions to Discuss

1. After the onset of a pandemic, do you think you would be more or less likely to find hand sanitizer for sale if the sellers were allowed to increase prices? Why?
2. If the sellers of scarce resources are not allowed to increase prices to equilibrate supply and demand after a disaster, how do you think these resources should be allocated among the population? What are the benefits of your proposal? What problems might arise with your proposal in practice?

Richard Thaler is a professor of economics at the University of Chicago. He won the Nobel prize in economics in 2017.

Source: *New York Times*, March 24, 2020.

12. What event might lead to an increase in the equilibrium price of jelly and a decrease in the equilibrium quantity of jelly sold?
    a. an increase in the price of peanut butter, a complement to jelly
    b. an increase in the price of Marshmallow Fluff, a substitute for jelly
    c. an increase in the price of grapes, an input into jelly
    d. an increase in consumers' incomes, as long as jelly is a normal good

13. An increase in _____ will cause a movement along a given supply curve, which is called a change in _____.
    a. supply; demand
    b. supply; quantity demanded
    c. demand; supply
    d. demand; quantity supplied

Answers are at the end of the chapter.

4-5 Conclusion: How Prices Allocate Resources

This chapter analyzed supply and demand in a single market. The discussion centered on the market for ice cream, but the lessons apply to most other markets as well. When you go to a store to buy something, you are contributing to the demand for that item. When you look for a job, you are contributing to the supply of labor services. Because supply and demand are such pervasive forces in market economies, the model of supply and demand is a powerful analytical tool.

"Two dollars" "—and seventy-five cents."

ROBERT J. DAY/THE NEW YORKER COLLECTION/THE CARTOON BANK

One of the **Ten Principles of Economics** in Chapter 1 is that markets are usually a good way to organize economic activity. It is still too early to judge whether market outcomes are good or bad, but this chapter has begun to show how markets work. In any economic system, scarce resources must be allocated among competing uses. Market economies harness the forces of supply and demand to serve that end. Supply and demand together determine the prices of the economy's many different goods and services. Prices, in turn, are the signals that guide the allocation of resources.

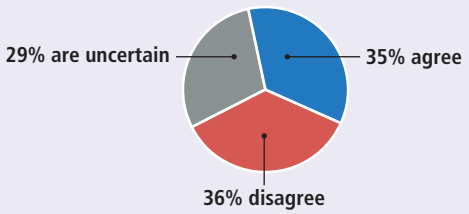
For example, consider the allocation of beachfront land. Because the amount of this land is limited, not everyone can enjoy the luxury of living by the beach. Who gets this resource? The answer is whoever is willing and able to pay the price. The price of beachfront land adjusts until the quantity of land demanded balances the quantity supplied. In market economies, prices are the mechanism for rationing scarce resources.

Similarly, prices determine who produces each good and how much is produced. For instance, consider farming. Because everyone needs food to survive, it is crucial that some people work on farms. What determines who is a farmer and who is not? In a free society, no government planning agency makes this decision to ensure an adequate food supply. Instead, the allocation of labor to farms is based on the job decisions of millions of workers. This decentralized system performs well because these decisions depend on prices. The prices of food and the wages of farmworkers (the price of their labor) adjust to ensure that enough people choose to be farmers.

Ask the Experts — Price Gouging

"Laws to prevent high prices for essential goods in short supply in a crisis would raise social welfare."

What do economists say?



Source: IGM Economic Experts Panel, May 26, 2020.

If a person had never seen a market economy in action, the whole idea might seem preposterous. Economies are enormous groups of people engaged in a multitude of interdependent activities. What prevents decentralized decision making from degenerating into chaos? What coordinates the actions of the millions of people with their varying abilities and desires? What ensures that what needs to be done is, in fact, done? The answer, in a word, is **prices**. If an invisible hand guides market economies, as Adam Smith famously suggested, the price system is the baton with which the invisible hand conducts the economic orchestra.

Chapter in a Nutshell

- Economists use the model of supply and demand to analyze competitive markets. In such markets, there are many buyers and sellers, each of whom has little or no influence on the market price.
- The demand curve for a good shows how the quantity demanded depends on the price. According to the law of demand, as the good's price falls, the quantity demanded rises. That's why the demand curve slopes downward.
- In addition to price, other determinants of how much consumers want to buy include income, the prices of substitutes and complements, tastes, expectations, and the number of buyers. When one of these factors changes, the quantity demanded at each price changes, and the demand curve shifts.
- The supply curve for a good shows how the quantity supplied depends on the price. According to the law of supply, as the good's price rises, the quantity supplied rises. That's why the supply curve slopes upward.
- In addition to price, other determinants of how much producers want to sell include input prices, technology, expectations, and the number of sellers. When one of these factors changes, the quantity supplied at each price changes, and the supply curve shifts.
- The intersection of the supply and demand curves represents the market equilibrium. At the equilibrium price, the quantity demanded equals the quantity supplied.
- The behavior of buyers and sellers naturally drives markets toward equilibrium. When the market price is above the equilibrium price, there is a surplus of the good, which causes the market price to fall. When the market price is below the equilibrium price, there is a shortage, which causes the market price to rise.
- To analyze how any event influences the equilibrium price and quantity in a market, use a supply-and-demand diagram and follow these three steps. First, decide if the event shifts the supply curve or the demand curve (or both). Second, decide in which direction the curve shifts. Third, compare the new equilibrium with the initial one.
- In market economies, prices are the signals that guide decisions and allocate scarce resources. For every good in the economy, the price ensures that supply and demand are in balance. The equilibrium price determines how much buyers choose to consume and how much sellers choose to produce.

Key Concepts

market, p. 62
competitive market, p. 62
quantity demanded, p. 63
law of demand, p. 63
demand schedule, p. 64
demand curve, p. 64
normal good, p. 66
inferior good, p. 66
substitutes, p. 66
complements, p. 66
quantity supplied, p. 69
law of supply, p. 69
supply schedule, p. 69
supply curve, p. 69
equilibrium, p. 73
equilibrium price, p. 73
equilibrium quantity, p. 73
surplus, p. 74
shortage, p. 74
law of supply and demand, p. 75

Questions for Review

1. What is a competitive market? Briefly describe a type of market that is **not** perfectly competitive.
2. What are the demand schedule and the demand curve, and how are they related? Why does the demand curve slope downward?
3. Does a change in consumers' tastes lead to a movement along the demand curve or to a shift in the demand curve? Does a change in price lead to a movement along the demand curve or to a shift in the demand curve? Explain your answers.
4. Harry's income declines, and as a result, he buys more pumpkin juice. Is pumpkin juice an inferior good or a normal good? What happens to Harry's demand curve for pumpkin juice?
5. What are the supply schedule and the supply curve, and how are they related? Why does the supply curve slope upward?
6. Does a change in producers' technology lead to a movement along the supply curve or to a shift in the supply curve? Does a change in price lead to a movement along the supply curve or to a shift in the supply curve?
7. Define the equilibrium of a market. Describe the forces that move a market toward its equilibrium.
8. Beer and pizza are complements because they are often enjoyed together. When the price of beer rises, what happens to the supply, demand, quantity supplied, quantity demanded, and price in the market for pizza?
9. Describe the role of prices in market economies.

Problems and Applications

1. Explain each of the following statements using supply-and-demand diagrams.
   a. "When a cold snap hits Florida, the price of orange juice rises in supermarkets throughout the United States."
   b. "When the weather turns warm in New England every summer, the price of hotel rooms in the Caribbean plummets."
   c. "When a war breaks out in the Middle East, the price of gasoline rises, and the price of a used SUV falls."
2. "An increase in the demand for notebooks raises the quantity of notebooks demanded but not the quantity supplied." Is this statement true or false? Explain.
3. Consider the market for minivans. For each of the events listed here, identify which of the determinants of demand or supply are affected. Also, indicate whether demand or supply increases or decreases. Then, draw a diagram to show the effect on the price and quantity of minivans.
   a. People decide to have more children.
   b. A strike by steelworkers raises steel prices.
   c. Engineers develop new automated machinery for the production of minivans.
   d. The price of sports utility vehicles rises.
   e. A stock market crash lowers people's wealth.
4. Consider the markets for video streaming services, TV screens, and tickets at movie theaters.
   a. For each pair, identify whether they are complements or substitutes:
      - video streaming and TV screens
      - video streaming and movie tickets
      - TV screens and movie tickets
   b. Suppose a technological advance reduces the cost of manufacturing TV screens. Draw a diagram to show what happens in the market for TV screens.
   c. Draw two more diagrams to show how the change in the market for TV screens affects the markets for video streaming and movie tickets.
5. Over the past 40 years, technological advances have reduced the cost of computer chips. How do you think this has affected the market for computers? For computer software? For typewriters?
6. Using supply-and-demand diagrams, show the effects of the following events on the market for sweatshirts.
   a. A hurricane in South Carolina damages the cotton crop.
   b. The price of leather jackets falls.
   c. All colleges require morning exercise in appropriate attire.
   d. New knitting machines are invented.

7. Ketchup is a complement (as well as a condiment) for hot dogs. If the price of hot dogs rises, what happens in the market for ketchup? For tomatoes? For tomato juice? For orange juice?

8. The market for pizza has the following demand and supply schedules:

Price	Quantity Demanded	Quantity Supplied
$4	135 pizzas	26 pizzas
5	104	53
6	81	81
7	68	98
8	53	110
9	39	121

   a. Graph the demand and supply curves. What are the equilibrium price and quantity in this market?
   b. If the actual price in this market were **above** the equilibrium price, what would drive the market toward equilibrium?
   c. If the actual price in this market were **below** the equilibrium price, what would drive the market toward equilibrium?

9. Consider the following events: Scientists reveal that eating oranges decreases the risk of diabetes, and at the same time, farmers use a new fertilizer that makes orange trees produce more oranges. Illustrate and explain what effect these changes have on the equilibrium price and quantity of oranges.

10. Because bagels and cream cheese are often eaten together, they are complements.
   a. We observe that both the equilibrium price of cream cheese and the equilibrium quantity of bagels have risen. What could be responsible for this pattern: a fall in the price of flour or a fall in the price of milk? Illustrate and explain your answer.
   b. Suppose instead that the equilibrium price of cream cheese has risen, but the equilibrium quantity of bagels has fallen. What could be responsible for this pattern: a rise in the price of flour or a rise in the price of milk? Illustrate and explain your answer.

11. Suppose that the price of basketball tickets at your college is determined by market forces. Currently, the demand and supply schedules are as follows:

Price	Quantity Demanded	Quantity Supplied
$4	10,000 tickets	8,000 tickets
8	8,000	8,000
12	6,000	8,000
16	4,000	8,000
20	2,000	8,000

   a. Draw the demand and supply curves. What is unusual about this supply curve? Why might this be true?
   b. What are the equilibrium price and quantity of tickets?
   c. Your college plans to increase total enrollment next year by 5,000 students. The additional students will have the following demand schedule:

Price	Quantity Demanded
$4	4,000 tickets
8	3,000
12	2,000
16	1,000
20	0

   Now add the old demand schedule and the demand schedule for the new students to calculate the new demand schedule for the entire college. What will be the new equilibrium price and quantity?

QuickQuiz Answers

1. **c** 2. **b** 3. **a** 4. **b** 5. **a** 6. **d** 7. **a** 8. **c** 9. **d** 10. **b** 11. **a** 12. **c** 13. **d**

Chapter

5

Elasticity and Its Application

Imagine that some event drives up the price of gasoline in the United States. It could be tensions in the Middle East that tighten the world supply of oil, a booming Chinese economy that boosts the world demand for oil, or a hike in the gasoline tax passed by Congress. How would U.S. consumers respond to the higher price?

It is easy to answer this question in a broad fashion: People would buy less gas. This follows from the law of demand in the previous chapter: Other things being equal, when the price of a good rises, the quantity demanded falls. But you might want a precise answer. By how much would gas purchases fall? This question can be answered using a concept called **elasticity**.

Elasticity is a measure of how much buyers and sellers respond to changes in market conditions. When studying how events or policies affect a market, we can discuss not only the direction of the effects but also their magnitude.

AZMANJAKA/E+/GETTY IMAGES, CARLOS FLORES/EYEEM/GETTY IMAGES

In the market for gasoline, studies typically find that the quantity demanded responds to gasoline prices more in the long run than in the short run. A 10 percent increase in the price reduces gasoline consumption by about 2.5 percent after a year but by about 6 percent after five years. About half of the long-run reduction comes from less driving, and half comes from switching to more fuel-efficient cars—and, increasingly, to electric cars that require no gasoline at all. Both responses are reflected in the demand curve and its elasticity.

5-1 The Elasticity of Demand

In Chapter 4, we noted that consumers usually buy more of a good when its price is lower, when their incomes are higher, when the prices of its substitutes are higher, or when the prices of its complements are lower. The discussion was qualitative, not quantitative. That is, it addressed whether quantity demanded rose or fell but not the size of the change. To measure how much consumers respond to changes in these variables, economists use the concept of **elasticity**.

elasticity
a measure of the responsiveness of the quantity demanded or quantity supplied to a change in one of its determinants

price elasticity of demand
a measure of how much the quantity demanded of a good responds to a change in its price, calculated as the percentage change in quantity demanded divided by the percentage change in price

5-1a The Price Elasticity of Demand and Its Determinants

The law of demand states that a fall in the price of a good raises the quantity demanded. The **price elasticity of demand** measures how much the quantity demanded responds to a change in the price. Demand for a good is said to be **elastic** if the quantity demanded responds substantially to price changes. Demand is said to be **inelastic** if the quantity demanded responds only slightly to price changes.

The price elasticity of demand for any good measures how willing consumers are to buy less of it as its price rises. Because a demand curve reflects the economic, social, and psychological forces that shape consumer preferences, there is no simple, universal rule for what determines a demand curve's elasticity. But there are some rules of thumb.

Availability of Close Substitutes Goods with close substitutes tend to have more elastic demand because it is easier for consumers to switch from those goods to others. For example, margarine is a common substitute for butter. Increase the price of butter by a small amount, and you will see its quantity demanded fall by a large amount, assuming the price of margarine is steady. By contrast, eggs don't have a close substitute, so the demand for them is less elastic. Increase the price of eggs by a small amount, and there will not be a big drop in the quantity of eggs demanded.

Necessities and Luxuries For a true necessity, a small price increase won't diminish the amount you purchase by very much: Another way of saying this is that necessities tend to have inelastic demands. For example, when the price of a doctor's visit rises, most people do not cut back sharply on the number of times they go to the doctor, although they might go a little less often. Luxuries are another matter. When the price of sailboats rises, the quantity of sailboats demanded falls substantially. The reason is that most people view sailboats as a luxury. Whether a good is a necessity or a luxury depends not on its intrinsic properties but on the buyer's preferences. For avid sailors with little concern about their health, sailboats might be a necessity with inelastic demand and doctor visits a luxury with elastic demand.

Defining the Market Broadly or Narrowly The elasticity of demand depends on how we draw the market's boundaries. Narrowly defined markets tend to have

more elastic demand than broadly defined ones. That's because it is easier to find close substitutes for narrowly defined goods. For example, food, a broad category, has a fairly inelastic demand because there are no good substitutes for food. Ice cream, a narrow category, has a more elastic demand because it is easy to substitute other desserts for ice cream. Vanilla ice cream, an even narrower category, has a very elastic demand because some other flavors of ice cream, such as sweet cream, are almost perfect substitutes for vanilla.

Time Horizon Demand tends to be more elastic over longer periods of time. When the price of gasoline rises, the quantity demanded falls only slightly in the first few months. As time passes, however, people buy more fuel-efficient or all-electric cars, arrange carpools, switch to public transportation, and move closer to work. Over several years, the quantity of gasoline demanded falls more substantially.

5-1b The Price Elasticity of Demand, with Numbers

Now that we have discussed the price elasticity of demand in general terms, let's be more precise about how it is measured. Economists compute the price elasticity of demand as the percentage change in the quantity demanded divided by the percentage change in the price. That is:

$$\text{Price elasticity of demand} = \frac{\text{Percentage change in quantity demanded}}{\text{Percentage change in price}}.$$

For example, suppose that after a 10 percent increase in the price of an ice-cream cone, you buy 20 percent fewer cones. We calculate your elasticity of demand this way:

$$\text{Price elasticity of demand} = \frac{20 \text{ percent}}{10 \text{ percent}} = 2.$$

In this example, the elasticity is 2. That means that the change in quantity demanded is proportionately twice as large as the change in price.

Because the quantity demanded of a good moves in the opposite direction as its price, the percentage change in quantity has the opposite sign as the percentage change in price. In this example, the percentage change in price is a **positive** 10 percent (reflecting an increase), and the percentage change in quantity demanded is a **negative** 20 percent (reflecting a decrease). For this reason, price elasticities of demand are sometimes reported as negative numbers. It is common practice, however, to drop the minus sign and report all price elasticities of demand as positive numbers. (Mathematicians call this the **absolute value**.) With this convention, which this book follows, a larger price elasticity implies a greater responsiveness of quantity demanded to price changes.

5-1c The Midpoint Method: A Better Way to Calculate Percentage Changes and Elasticities

If you try calculating the price elasticity of demand between two points on a demand curve, you will face an annoying problem: The elasticity from point A to point B seems different from the elasticity from point B to point A. Consider this example:

Point A : Price = \$4 Quantity = 120
Point B : Price = \$6 Quantity = 80

Going from point A to point B, the price rises by 50 percent and the quantity falls by 33 percent, indicating that the price elasticity of demand is 33/50, or 0.66. Going from point B to point A, the price falls by 33 percent and the quantity rises by 50 percent, indicating that the price elasticity of demand is 50/33, or 1.5. This difference arises because the percentage changes are calculated from a different base. Yet the underlying reality—the response of buyers to price changes—is identical whether moving from point A to point B or from point B to point A.

The **midpoint method** for calculating elasticities avoids this confusion. The standard procedure for calculating a percentage change is to divide the change by the initial level. But the midpoint method instead divides the change by the midpoint (or average) of the initial and final levels. For instance, \$5 is midway between \$4 and \$6. Therefore, according to the midpoint method, a change from \$4 to \$6 is considered a 40 percent rise because $(6 - 4)/5 \times 100 = 40$. Similarly, a change from \$6 to \$4 is considered a 40 percent fall.

Because the midpoint method gives the same answer regardless of the direction of change, it is often used when calculating the price elasticity of demand between two points. In our example, the midpoint between point A and point B is:

Midpoint: Price = \$5 Quantity = 100

According to the midpoint method, when going from point A to point B, the price rises by 40 percent and the quantity falls by 40 percent. Similarly, when going from point B to point A, the price falls by 40 percent and the quantity rises by 40 percent. In both directions, the price elasticity of demand equals 1.

The following formula expresses the midpoint method for calculating the price elasticity of demand between two points, denoted (Q_1, P_1) and (Q_2, P_2):

$$\text{Price elasticity of demand} = \frac{(Q_2 - Q_1)/\left[(Q_2 + Q_1)/2\right]}{(P_2 - P_1)/\left[(P_2 + P_1)/2\right]}.$$

The numerator is the percentage change in quantity using the midpoint method, and the denominator is the percentage change in price using the midpoint method. If you ever need to calculate elasticities, use this formula.

This book rarely performs such calculations. For most purposes here, what elasticity represents—the responsiveness of quantity demanded to price changes—is more important than how it is calculated.

5-1d The Variety of Demand Curves

Economists use elasticity to classify demand curves. When the quantity moves proportionately more than the price, the elasticity is greater than one, and demand is said to be **elastic**. When the quantity moves proportionately less than the price, the elasticity is less than one, and demand is said to be **inelastic**. Finally, when the percentage change in quantity equals the percentage change in price, the elasticity is exactly 1, and demand is said to have **unit elasticity**.

Because the price elasticity of demand measures how much quantity demanded responds to price changes, it is closely related to the slope of the demand curve. Here's a useful rule of thumb: The flatter the demand curve at a given point, the greater the price elasticity of demand. The steeper the demand curve at a given point, the smaller the price elasticity of demand.

Figure 1 shows five cases. In the extreme case of zero elasticity, shown in panel (a), demand is **perfectly inelastic**, and the demand curve is vertical. In this case, regardless of the price, the quantity demanded stays the same. As the elasticity rises, the demand curve gradually flattens, as shown in panels (b), (c), and (d). At the opposite extreme, in panel (e), demand is **perfectly elastic**. This occurs as the price elasticity of demand becomes so large that it approaches infinity. The demand curve becomes horizontal, showing that tiny changes in the price lead to huge changes in the quantity demanded.

If you have trouble keeping straight which curve is **elastic** and which **inelastic**, here's a memory trick: **I**nelastic curves, such as in panel (a) of Figure 1, look like the letter I. (Economists call it a curve, but when it's perfectly inelastic, it's really a vertical line.) This is not a deep insight, but it might help on your next exam.

5-1e Total Revenue and the Price Elasticity of Demand

When studying changes in supply or demand in a market, one variable we often want to study is **total revenue**, the amount paid by buyers and received by sellers of the good. In mathematical form, total revenue is $P \times Q$, the price of the good times the quantity of the good sold. Figure 2 shows total revenue graphically. The height of the box under the demand curve is P, and the width is Q. The area of this box, $P \times Q$, equals the total revenue in this market. In Figure 2, where $P = \$4$ and $Q = 100$, total revenue is $\$4 \times 100$, or $\$400$.

total revenue
the amount paid by buyers and received by the sellers of a good, calculated as the price of the good times the quantity sold

FYI

A Few Elasticities from the Real World

We have talked about what elasticity means, what determines it, and how it is calculated. Beyond these general ideas, you might ask for a specific number. How much, precisely, does a good's price influence its quantity demanded?

To answer this question, economists collect market data and apply statistical techniques to estimate the price elasticity of demand. Here are some price elasticities of demand, obtained from various studies, for a range of goods:

Eggs	0.1	▲ Very inelastic (quantity demanded responds little to price changes)
Healthcare	0.2	
Cigarettes	0.4	
Rice	0.5	
Housing	0.7	
Beef	1.6	
Peanut Butter	1.7	
Restaurant Meals	2.3	
Cheerios	3.7	
Mountain Dew	4.4	▼ Very elastic (quantity demanded responds strongly to price changes)

These numbers are fun to think about, and they can be useful when comparing markets, but take them with a grain of salt. One reason is that the statistical techniques used to obtain them require some assumptions about the world, and these assumptions might not be true in practice. (The part of economics called econometrics studies these statistical techniques.) Another reason is that the price elasticity of demand need not be the same at all points on a demand curve, as we will see in the case of a linear demand curve. For both reasons, don't be surprised when different studies report different price elasticities of demand for the same good. ■

Figure 1

The Price Elasticity of Demand

The price elasticity of demand determines whether the demand curve is steep or flat. Note that all percentage changes are calculated using the midpoint method.

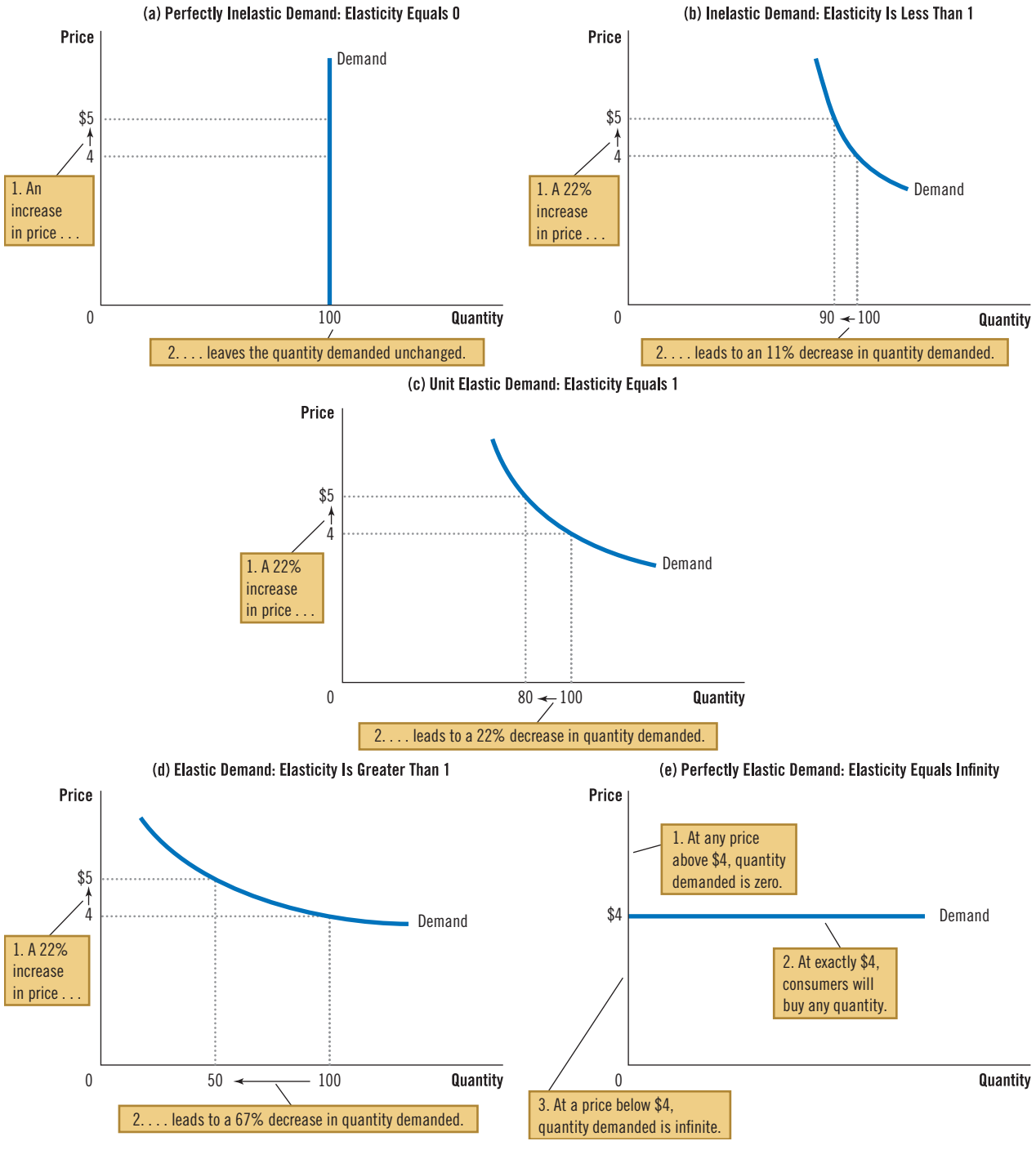
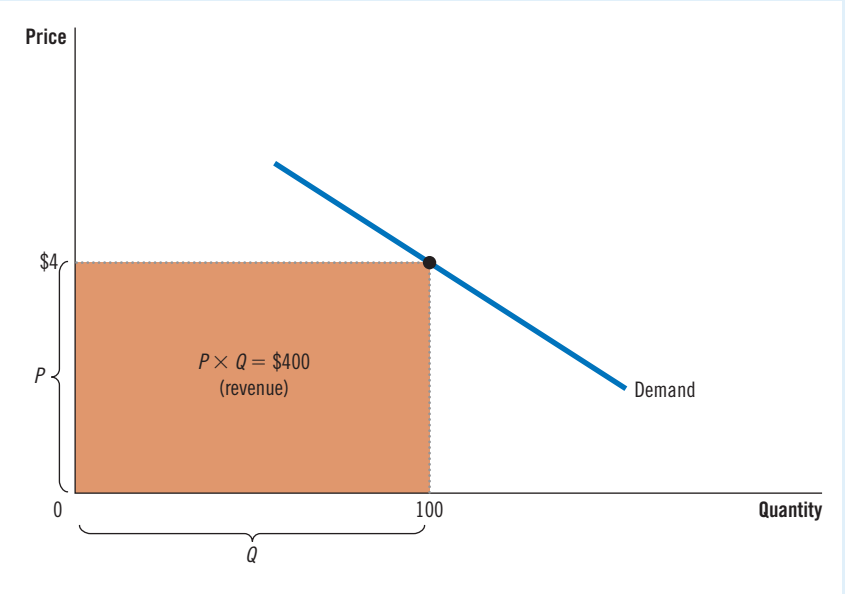


Figure 2

Total Revenue

The area of the box under the demand curve, $P \times Q$, equals the total amount paid by buyers as well as the total revenue received by sellers. Here, at a price of \$4, the quantity demanded is 100, and total revenue is \$400.



How does total revenue change as one moves along the demand curve? The answer depends on the price elasticity of demand. If demand is inelastic, as in panel (a) of Figure 3, then an increase in the price causes an increase in total revenue. Here, an increase in the price from \$4 to \$5 causes the quantity demanded to fall from 100 to 90, so total revenue rises from \$400 to \$450. An increase in the price raises $P \times Q$ because the fall in Q is proportionately smaller than the rise in P. In other words, the extra revenue from selling units at a higher price (represented by area A in the figure) more than offsets the decline in revenue from selling fewer units (represented by area B).

The opposite result occurs when demand is elastic: An increase in the price causes total revenue to decline. In panel (b) of Figure 3, for instance, when the price rises from \$4 to \$5, the quantity demanded falls from 100 to 70, so total revenue falls from \$400 to \$350. Because demand is elastic, the reduction in quantity demanded is so great that it more than offsets the price increase. That is, an increase in the price reduces $P \times Q$ because the fall in Q is proportionately greater than the rise in P. In this case, the extra revenue from selling units at a higher price (area A) is smaller than the decline in revenue from selling fewer units (area B).

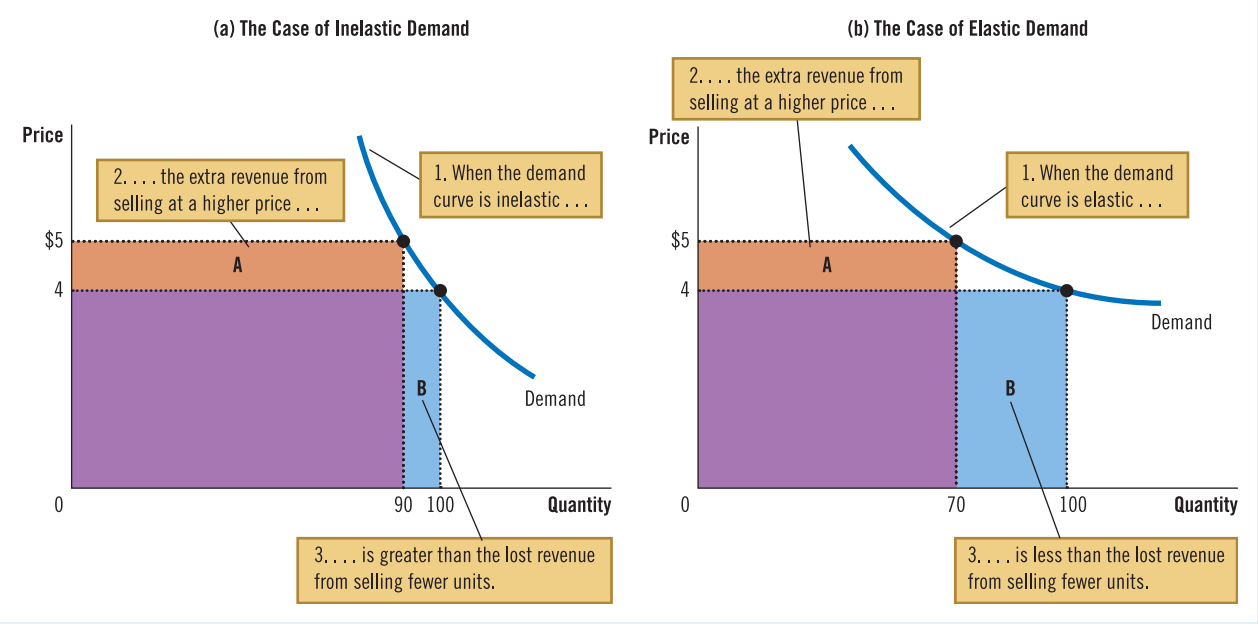
The examples in this figure illustrate some general rules:

- When demand is inelastic (a price elasticity less than one), the price and total revenue move in the same direction: If the price increases, total revenue also increases.
- When demand is elastic (a price elasticity greater than one), the price and total revenue move in opposite directions: If the price increases, total revenue decreases.
- If demand is unit elastic (a price elasticity exactly equal to 1), total revenue remains constant when the price changes.

Figure 3

How Total Revenue Changes When Price Changes

The impact of a price change on total revenue (price times quantity) depends on the elasticity of demand. In panel (a), the demand curve is inelastic. A price increase leads to a proportionately smaller decrease in quantity demanded, so total revenue increases. Here, the price increases from $4 to $5, and the quantity demanded falls from 100 to 90. Total revenue rises from $400 to $450. In panel (b), the demand curve is elastic. A price increase leads to a proportionately larger decrease in quantity demanded, so total revenue decreases. Here, the price increases from $4 to $5, and the quantity demanded falls from 100 to 70. Total revenue falls from $400 to $350.



5-1f Elasticity and Total Revenue along a Linear Demand Curve

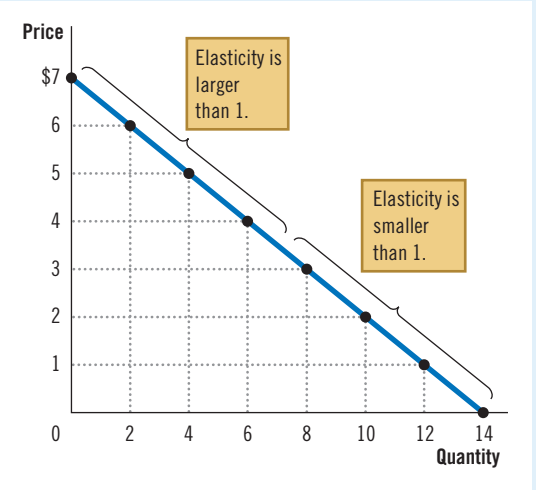
Let's examine how elasticity varies along a linear demand curve, as shown in Figure 4. Because the demand curve is a straight line, it has a constant slope. Slope is defined as "rise over run," which here is the ratio of the change in price ("rise") to the change in quantity ("run"). In this case, the demand curve's slope is constant because each $1 increase in the price causes the same decrease of two units in the quantity demanded.

Even though the slope of a linear demand curve is constant, the elasticity is not. This is because the slope is the ratio of **changes** in the two variables, while the elasticity is the ratio of **percentage changes** in them. You can see this in the table for Figure 4, which shows the demand schedule for the linear demand curve in the graph. The table uses the midpoint method to calculate the price elasticity of demand. It illustrates this fundamental idea: **At points with a low price and high quantity, a linear demand curve is inelastic. At points with a high price and low quantity, a linear demand curve is elastic.**

Figure 4

Elasticity along a Linear Demand Curve

The slope of a linear demand curve is constant, but its elasticity is not. The price elasticity of demand is calculated using the demand schedule and the midpoint method. At points with a low price and high quantity, the demand curve is inelastic. At points with a high price and low quantity, it is elastic.



Price	Quantity	Total Revenue (Price × Quantity)	Percentage Change in Price	Percentage Change in Quantity	Elasticity	Description
$7	0	$0				
			15	200	13.0	Elastic
6	2	12				
			18	67	3.7	Elastic
5	4	20				
			22	40	1.8	Elastic
4	6	24				
			29	29	1.0	Unit elastic
3	8	24				
			40	22	0.6	Inelastic
2	10	20				
			67	18	0.3	Inelastic
1	12	12				
			200	15	0.1	Inelastic
0	14	0				

The explanation comes from the arithmetic of percentage changes. When the price is low and consumers are buying a lot, a $1 price increase and two-unit reduction in quantity demanded constitute a large percentage increase in the price and a small percentage decrease in quantity demanded, resulting in a small elasticity. When the price is high and consumers are not buying much, the same $1 price increase and two-unit reduction in quantity demanded constitute a small percentage increase in the price and a large percentage decrease in quantity demanded, resulting in a large elasticity.

The table presents total revenue at each point on the demand curve. These numbers illustrate the relationship between total revenue and elasticity. When the price is $1, for instance, demand is inelastic, and a price increase to $2 raises total revenue. When the price is $5, demand is elastic, and a price increase to $6 reduces total revenue. Between $3 and $4, demand is exactly unit elastic and total revenue is the same at these two prices.

In short, the price elasticity of demand need not be the same at all points on a demand curve. A demand curve with constant elasticity is possible, but it is a special case. A linear demand curve never has a constant elasticity.

5-1g Other Demand Elasticities

In addition to the price elasticity of demand, economists use other elasticities to describe the behavior of buyers in a market.

income elasticity of demand
a measure of how much the quantity demanded of a good responds to a change in consumers' income, calculated as the percentage change in quantity demanded divided by the percentage change in income

The Income Elasticity of Demand The **income elasticity of demand** measures how the quantity demanded changes as consumer income changes. It is calculated as the percentage change in quantity demanded divided by the percentage change in income. That is:

$$\text{Income elasticity of demand} = \frac{\text{Percentage change in quantity demanded}}{\text{Percentage change in income}}.$$

As Chapter 4 discussed, most goods are **normal goods**: Higher income increases the quantity demanded. Because quantity demanded and income move in the same direction, normal goods have positive income elasticities. A few goods, such as bus rides, are **inferior goods**: This doesn't mean that anything is wrong with them, just that higher income reduces the quantity demanded. Because quantity demanded and income move in opposite directions, inferior goods have negative income elasticities.

Even among normal goods, income elasticities vary substantially in size. Necessities such as food tend to have small income elasticities because consumers buy some of these goods even when their incomes are low. (**Engel's Law**, named for the 19th-century statistician who discovered it, says that as a family's income rises, the percent of its income spent on food declines, indicating an income elasticity less than one.) By contrast, luxuries such as diamond jewelry and sailboats tend to have large income elasticities because most consumers feel that they can do without them altogether when their incomes decline.

cross-price elasticity of demand
a measure of how much the quantity demanded of one good responds to a change in the price of another good, calculated as the percentage change in the quantity demanded of the first good divided by the percentage change in the price of the second good

The Cross-Price Elasticity of Demand The **cross-price elasticity of demand** measures how the quantity demanded of one good responds to a change in the price of another. It is calculated as the percentage change in the quantity demanded of good one divided by the percentage change in the price of good two. That is:

$$\text{Cross-price elasticity of demand} = \frac{\text{Percentage change in quantity demanded of good one}}{\text{Percentage change in the price of good two}}$$

Whether the cross-price elasticity is positive or negative depends on whether the two goods are substitutes or complements. As discussed in Chapter 4, **substitutes** are goods that are typically used in place of one another, such as hamburgers and hot dogs. When hot dog prices increase, people grill more hamburgers instead. Because the price of hot dogs and the quantity of hamburgers demanded move in the same direction, the cross-price elasticity is positive. Conversely, **complements** are goods that are typically used together, such as computers and software. In this case, the cross-price elasticity is negative, indicating that an increase in the price of computers reduces the quantity of software demanded.

QuickQuiz

1. A good tends to have a small price elasticity of demand if
   a. the good is a necessity.
   b. there are many close substitutes.
   c. the market is narrowly defined.
   d. the long-run response is being measured.
2. An increase in a good's price reduces the total amount consumers spend on the good if the _____ elasticity of demand is _____ than one.
   a. income; less
   b. income; greater
   c. price; less
   d. price; greater
3. A linear, downward-sloping demand curve is
   a. inelastic.
   b. unit elastic.
   c. elastic.
   d. inelastic at some points and elastic at others.
4. The citizens of Rohan spend a higher fraction of their income on food than do the citizens of Gondor. The reason could be that
   a. Rohan has lower food prices, and the price elasticity of demand is zero.
   b. Rohan has lower food prices, and the price elasticity of demand is 0.5.
   c. Rohan has lower income, and the income elasticity of demand is 0.5.
   d. Rohan has lower income, and the income elasticity of demand is 1.5.

Answers are at the end of the chapter.

5-2 The Elasticity of Supply

The discussion of supply in Chapter 4 noted that producers of a good offer to sell more of it when its price rises. To turn from qualitative to quantitative statements about quantity supplied, economists again use the concept of elasticity.

5-2a The Price Elasticity of Supply and Its Determinants

The law of supply states that higher prices increase the quantity supplied. The **price elasticity of supply** measures how much the quantity supplied responds to changes in the price. Supply is said to be **elastic** if the quantity supplied responds substantially to price changes and **inelastic** if the quantity supplied responds only slightly.

price elasticity of supply
a measure of how much the quantity supplied of a good responds to a change in its price, calculated as the percentage change in quantity supplied divided by the percentage change in price

The price elasticity of supply depends on the flexibility of sellers to change the amount they produce. Beachfront land has an inelastic supply: As Mark Twain once advised, "Buy land, they're not making it anymore." Manufactured goods, such as books, cars, and televisions, have elastic supplies because firms that produce them can run their factories longer in response to higher prices.

In most markets, supply is more elastic in the long run than in the short run. The reason is simple: Over short periods, firms can produce more by running longer shifts, but they can't easily change the size of their factories. Thus, in the short run, the quantity supplied is not very responsive to changes in the price. Over longer periods of time, firms can build new factories or close old ones. In addition, new firms can enter a market, and old ones can exit. Thus, in the long run, the quantity supplied responds substantially to price changes.

5-2b The Price Elasticity of Supply, with Numbers

This is generally how the price elasticity of supply works, but let's be more precise. Economists calculate it as the percentage change in quantity supplied divided by the percentage change in price. That is:

$$\text{Price elasticity of supply} = \frac{\text{Percentage change in quantity supplied}}{\text{Percentage change in price}}.$$

For example, suppose that an increase in the price of milk from $2.85 to $3.15 a gallon raises the amount that dairy farmers produce from 9,000 to 11,000 gallons per month. Using the midpoint method, we calculate the percentage change in price as:

$$\text{Percentage change in price} = (3.15 - 2.85)/3.00 \times 100 = 10 \text{ percent}.$$

Similarly, we calculate the percentage change in quantity supplied as:

$$\text{Percentage change in quantity supplied} = (11{,}000 - 9{,}000)/10{,}000 \times 100 = 20 \text{ percent}.$$

In this case, the price elasticity of supply is:

$$\text{Price elasticity of supply} = \frac{20 \text{ percent}}{10 \text{ percent}} = 2.$$

In this example, the elasticity of 2 indicates that the quantity supplied changes proportionately twice as much as the price.

5-2c The Variety of Supply Curves

The appearance of supply curves reflects the price elasticity of supply. Figure 5 shows five cases. In the extreme case of zero elasticity, in panel (a), supply is **perfectly inelastic**, and the supply curve is vertical. In this case, the quantity supplied is the same regardless of the price. As the elasticity rises, the supply curve flattens, showing that the quantity supplied responds more to changes in the price. At the opposite extreme, in panel (e), supply is **perfectly elastic**. This occurs as the price elasticity of supply approaches infinity and the supply curve becomes horizontal, meaning that tiny price changes lead to large changes in the quantity supplied.

In some markets, the elasticity of supply is not constant but varies over the supply curve. Figure 6 shows a typical case for an industry in which firms have factories with limited production capacity. For low levels of quantity supplied, the elasticity of supply is high, indicating that firms respond substantially to price changes. In this region of the supply curve, firms have additional capacity for production, such as plants and equipment that are idle for all or part of the day. Small increases in the price make it profitable for firms to start using this idle capacity. But as the quantity supplied rises, firms approach the limits of their current capacity. Increasing production further may require the construction of new factories. But the price must rise substantially to justify the extra expense, so in this range, supply is less elastic.

Figure 6 shows how this works. When the price rises from $3 to $4 (a 29 percent increase, using the midpoint method), the quantity supplied rises from 100 to 200

Figure 5

The Price Elasticity of Supply

The price elasticity of supply determines whether the supply curve is steep or flat. Note that all percentage changes are calculated using the midpoint method.

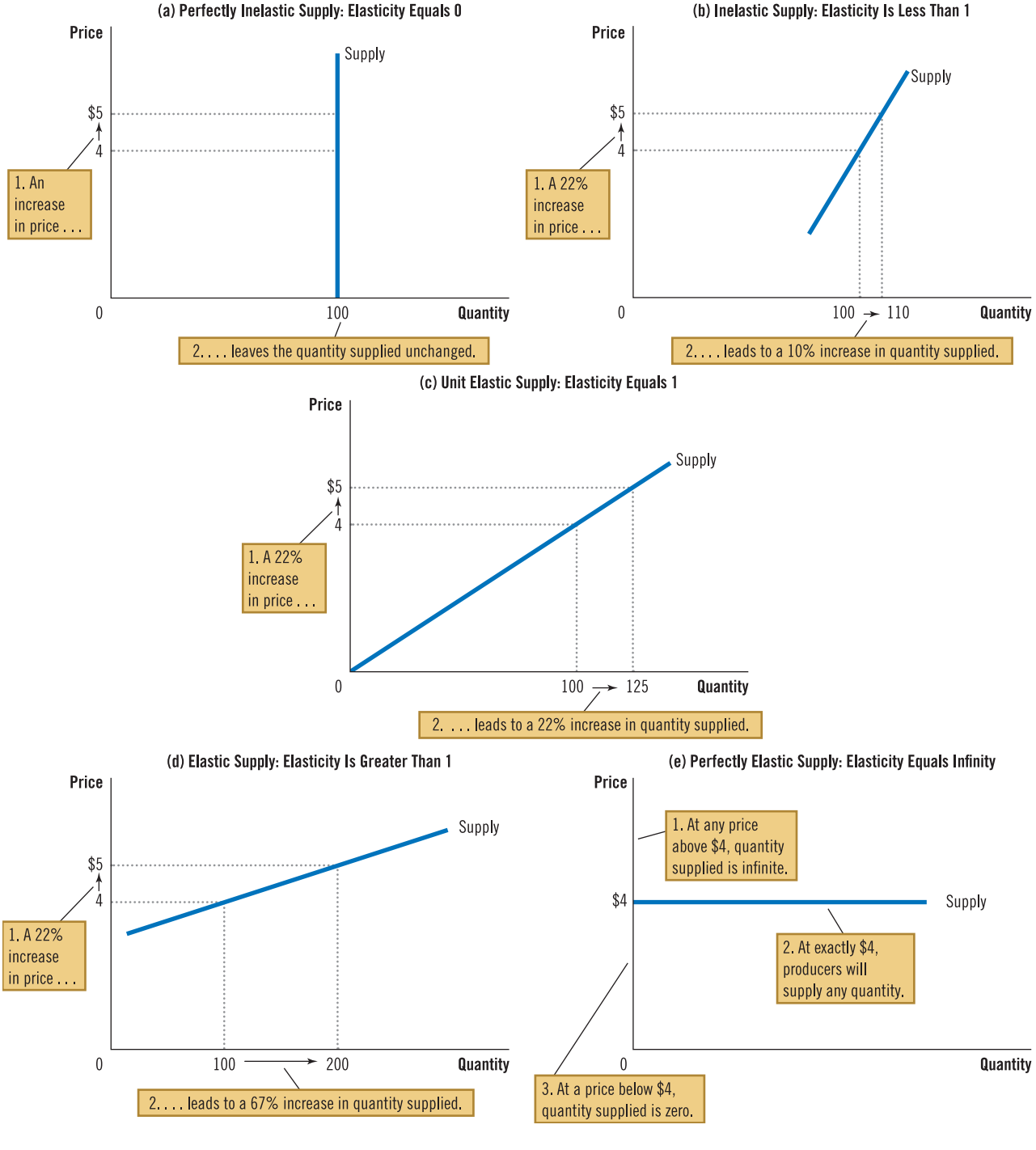
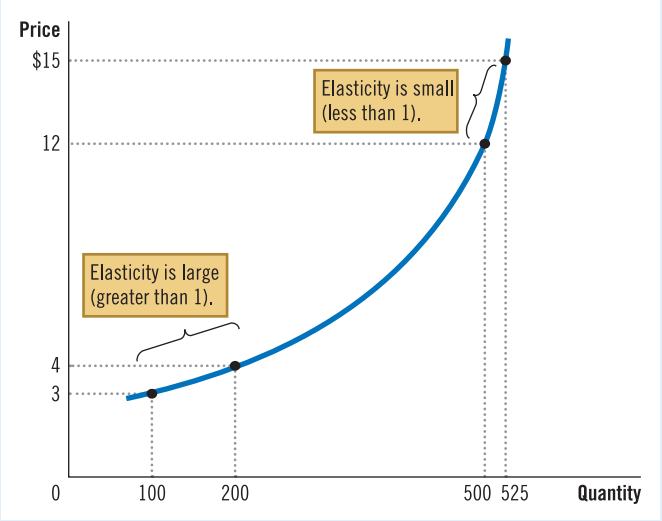


Figure 6

How the Price Elasticity of Supply Can Vary

Because firms often have a maximum capacity for production, the elasticity of supply may be very high at low levels of quantity supplied and very low at high levels of quantity supplied. Here, an increase in the price from $3 to $4 increases the quantity supplied from 100 to 200. Because the 67 percent increase in quantity supplied (calculated with the midpoint method) is larger than the 29 percent increase in price, the supply curve in this range is elastic. By contrast, when the price rises from $12 to $15, the quantity supplied rises only from 500 to 525. Because the 5 percent increase in quantity supplied is smaller than the 22 percent increase in price, the supply curve in this range is inelastic.



(a 67 percent increase). Because the quantity supplied changes proportionately more than the price, the supply curve has an elasticity greater than one. By contrast, when the price rises from $12 to $15 (a 22 percent increase), the quantity supplied rises from 500 to 525 (a 5 percent increase). In this case, the quantity supplied moves proportionately less than the price, so the elasticity is less than one.

QuickQuiz

5. The price of a good rises from $16 to $24, and the quantity supplied rises from 90 to 110 units. Calculated with the midpoint method, the price elasticity of supply is
   a. 1/5.
   b. 1/2.
   c. 2.
   d. 5.
6. If the price elasticity of supply is zero, the supply curve is
   a. upward sloping.
   b. horizontal.
   c. vertical.
   d. fairly flat at low quantities but steeper at larger quantities.
7. The ability of firms to enter and exit a market over time means that, in the long run,
   a. the demand curve is more elastic.
   b. the demand curve is less elastic.
   c. the supply curve is more elastic.
   d. the supply curve is less elastic.

Answers are at the end of the chapter.

5-3 Three Applications of Supply, Demand, and Elasticity

Can good news for farming be bad news for farmers? Why has OPEC, the international oil cartel, failed to keep the price of oil high? Does drug interdiction increase or decrease drug-related crime? These questions might seem to have little in common. Yet they are all about markets, and markets are all subject to the forces of supply and demand.

5-3a Can Good News for Farming Be Bad News for Farmers?

Imagine you're a Kansas wheat farmer. Because all your income comes from selling wheat, you make your land as productive as possible. You monitor weather and soil conditions, check the fields for pests and disease, and study the latest advances in farm technology. The more wheat you grow, the more you will harvest, and the higher your sales and standard of living will be.

One day, Kansas State University announces a major discovery. Researchers have developed a hybrid of wheat that increases production per acre by 20 percent. How should you react? Should you grow this new hybrid? Does this discovery make you better or worse off than you were before?

Recall the three steps from Chapter 4. First, examine whether the supply or demand curve shifts. Second, consider the direction of any shift. Third, use a supply-and-demand diagram to see how the market equilibrium changes.

In this case, the discovery of the hybrid affects the supply curve. Because the hybrid increases production per acre, farmers are willing to supply more wheat at any price. In other words, the supply curve shifts to the right. The demand curve remains the same because consumers' desire to buy wheat products at any price is not affected by the discovery of the hybrid. Figure 7 shows such a change. When the supply curve shifts from S_1 to S_2, the quantity of wheat sold increases from 100 to 110, and the price falls from \$3 to \$2.

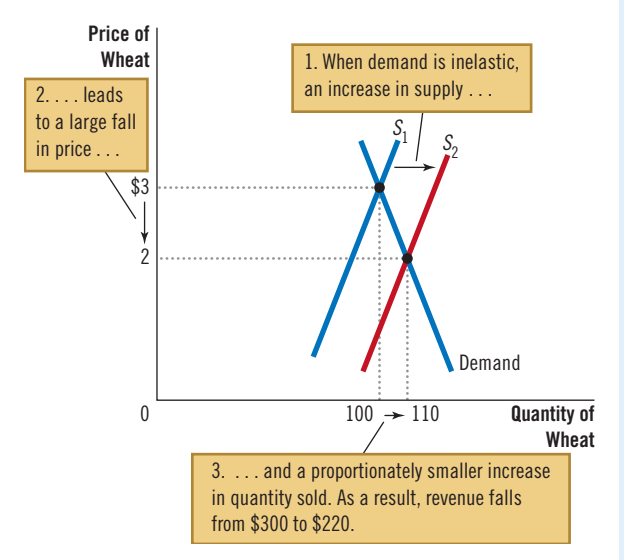
Does this discovery make farmers better off? Consider what happens to their total revenue, which can be written as $P \times Q$, the price of the wheat times the quantity sold. The discovery affects farmers in conflicting ways. The hybrid allows farmers to produce more wheat (Q rises), but each bushel of wheat sells for less (P falls).

The price elasticity of demand determines whether total revenue rises or falls. Wheat is a central ingredient in many people's diet. Demand for such basic foodstuffs is usually inelastic because they are relatively inexpensive and have few good substitutes. When the demand curve is inelastic, as in Figure 7, a decrease in the

Figure 7

A Supply Increase in the Market for Wheat

When a technological advance increases the wheat supply from S_1 to S_2, the price falls. Because the demand for wheat is inelastic, the increase in quantity from 100 to 110 is proportionately smaller than the decrease in the price from \$3 to \$2. As a result, farmers' total revenue falls from \$300 (\$3 × 100) to \$220 (\$2 × 110).



price causes total revenue to fall. You can see this in the figure: The price of wheat falls substantially, while the quantity sold rises only slightly. Total revenue falls from $300 to $220. In short, the new hybrid reduces farmers' total revenue.

If farmers are made worse off by the discovery of this hybrid, why do they adopt it? The answer goes to the heart of how competitive markets work. Because each farmer represents only a small part of the market for wheat, each takes the price as given. For any given price, it is better to produce and sell more wheat, which is accomplished using the new hybrid. Yet when all farmers do this, the supply of wheat increases, the price falls, and farmers are worse off.

This example may seem hypothetical, but it helps explain a major change in the U.S. economy. Two hundred years ago, most Americans lived on farms. Knowledge about agricultural methods was so primitive that most Americans had to be farmers to produce enough food to feed the nation's population. But over time, advances in farm technology increased the amount of food that each farmer could produce. The increase in food supply, together with the inelastic demand for food, caused farm revenues to fall, which led people to leave farming.

A few numbers show the magnitude of the change in the United States. In 1900, about 12 million people worked on farms, representing 40 percent of the labor force. In 2020, about 3 million people worked on farms, representing 2 percent of the labor force. Despite the large drop in the number of farmers, U.S. farms fed a population that increased more than fourfold, thanks to rising productivity.

This analysis helps explain a curious public policy: Certain government programs try to help farmers by inducing them **not** to plant crops. The purpose is to reduce the supply of farm products and thereby raise prices. With inelastic demand for their goods, farmers as a group receive greater total revenue if they supply a smaller crop to the market. Without government intervention, no individual farmer would choose to leave land fallow because each takes the market price as given. Less planting simply means lower earnings. But if all farmers can be persuaded to plant less together, the market price rises, and they can all be better off. Taxpayer-funded subsidies can help in that persuasion.

The interests of farmers, however, may not coincide with the interests of society as a whole. Improvement in farm technology can be bad for farmers because it makes them increasingly unnecessary, but it is good for consumers who pay less for food. Similarly, a government policy aimed at reducing the supply of farm products may raise the incomes of farmers, but it does so at the expense of consumers who pay higher prices and taxpayers who bear the cost of the subsidies.

BALOO/WWW.CARTOONSTOCK.COM

5-3b Why Has OPEC Failed to Keep the Price of Oil High?

Many of the most disruptive events for the world's economies originated in the world market for oil. In the 1970s, members of the Organization of Petroleum Exporting Countries (OPEC) decided to raise the world price of oil to increase their incomes. These countries accomplished this goal by agreeing to jointly reduce the amount of oil they supplied. As a result, the price of oil (adjusted for overall inflation) rose more than 50 percent from 1973 to 1974. Then, a few years later, OPEC did the same thing again. From 1979 to 1981, the price of oil approximately doubled.

Yet OPEC found it difficult to maintain such a high price. From 1982 to 1985, the price of oil steadily declined about 10 percent per year. Dissatisfaction and disarray soon prevailed among the OPEC countries. In 1986, cooperation among OPEC members completely broke down, and the price of oil plunged 45 percent. In 1990, the price of oil (adjusted for overall inflation) was back to where it began in 1970, and it stayed at that low level throughout most of the 1990s.

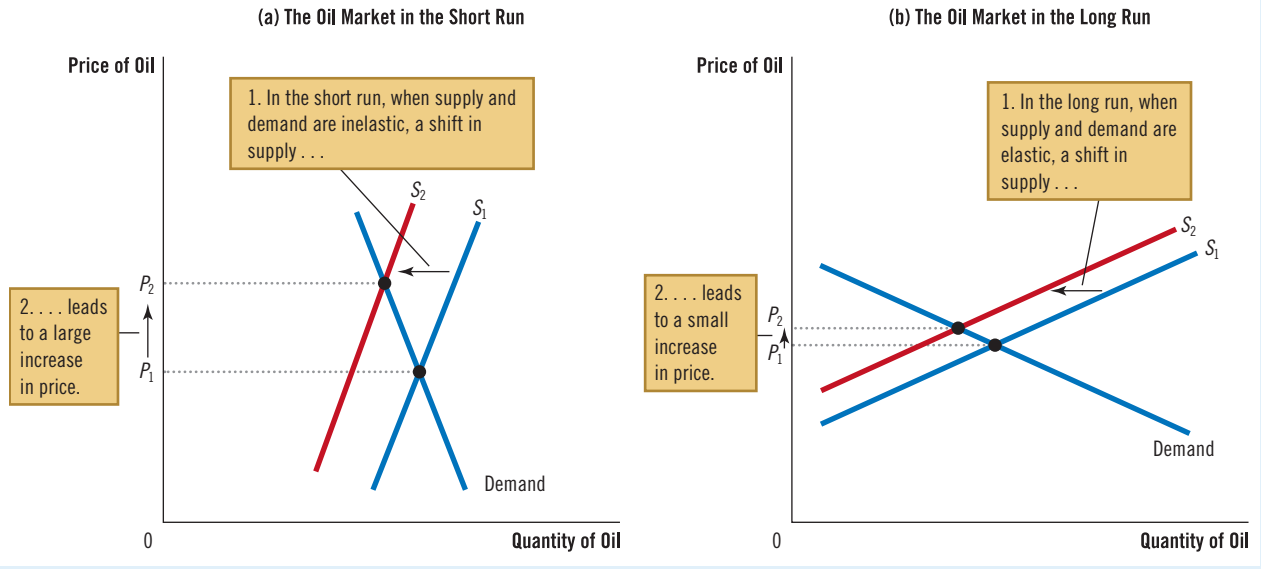
The OPEC episodes of the 1970s and 1980s show how supply and demand can behave differently in the short run and in the long run. In the short run, both the supply and demand for oil are relatively inelastic. Supply is inelastic because the quantity of known oil reserves and the capacity for oil extraction cannot be changed quickly. Demand is inelastic because buying habits do not respond immediately to changes in the price. That's why the short-run supply and demand curves are steep, as in panel (a) of Figure 8. When the supply of oil shifts from S_1 to S_2, the price increase from P_1 to P_2 is large.

The situation is very different in the long run. Over extended periods, producers of oil outside OPEC respond to high prices by increasing oil exploration and by

Figure 8

A Reduction in Supply in the World Market for Oil

When the supply of oil falls, the response depends on the time horizon. In the short run, supply and demand are relatively inelastic, as in panel (a). The shift in the supply curve from S_1 to S_2 leads to a substantial price increase. In the long run, however, supply and demand are relatively elastic, as in panel (b). In this case, the same size shift in the supply curve (S_1 to S_2) causes a smaller price increase.



building new extraction capacity. Consumers respond with greater conservation, such as by replacing old, inefficient cars with newer, efficient ones. As panel (b) of Figure 8 shows, the long-run supply and demand curves are more elastic. In the long run, the shift in the supply curve from S_1 to S_2 causes a much smaller increase in the price.

This is why OPEC was able to keep the price of oil high only in the short run. When OPEC countries agreed to reduce their production of oil, they shifted the supply curve to the left. Even though each OPEC member sold less oil, the price rose by so much in the short run that OPEC incomes rose. In the long run, however, supply and demand are more elastic. As a result, the same reduction in supply, measured by the horizontal shift in the supply curve, caused a smaller increase in the price. OPEC learned that raising prices is easier in the short run than in the long run.

During the first two decades of the 21st century, the price of oil fluctuated substantially once again, but the main driving force was not OPEC supply restrictions. Instead, booms and busts in economies around the world caused demand to fluctuate, while advances in fracking technology caused large increases in supply. Going forward, a main driving force in the oil market will be the move away from fossil fuels, motivated by concerns about global climate change.

5-3c Does Drug Interdiction Increase or Decrease Drug-Related Crime?

The use of illegal drugs, such as heroin, fentanyl, cocaine, ecstasy, and methamphetamine, has plagued the United States for decades. Drug use has several pernicious effects. Addiction can ruin the lives of drug users and their families. Addicts often turn to robbery and other violent crimes to support their habit, and when they are apprehended, they may spend long stretches in prison. In response to the drug epidemic, the U.S. government has devoted billions of dollars each year to reduce the flow of drugs into the country. The tools of supply and demand are useful in examining the effects of this policy of drug interdiction.

Suppose the government increases the number of federal agents devoted to stopping drug trafficking. What is the big picture in the market for illegal drugs? As usual, the answer comes from the three steps. First, consider whether the supply or demand curve shifts. Second, consider the direction of the shift. Third, examine how the shift affects the equilibrium price and quantity.

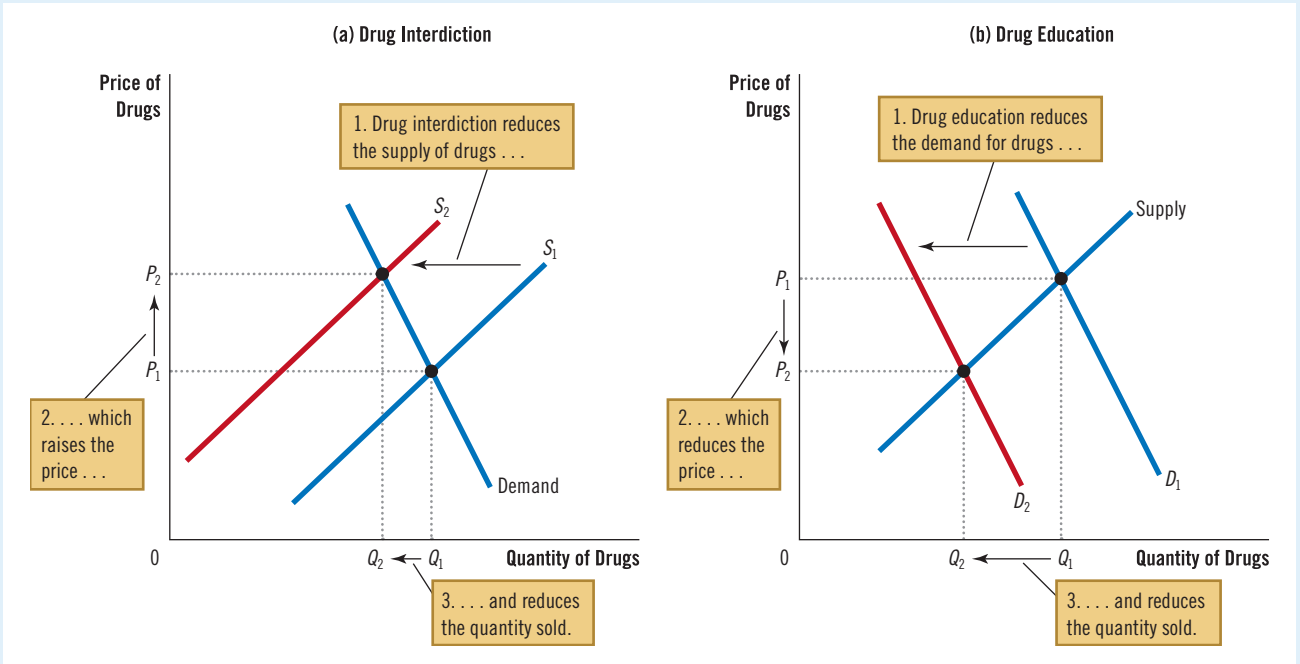
The direct impact of interdiction is on drug sellers rather than drug buyers. When the government stops drugs from entering the country and arrests smugglers, it increases the cost of selling drugs and, other things being equal, reduces the quantity of drugs supplied at any price. The demand for drugs—the amount buyers want at any price—remains the same. As in panel (a) of Figure 9, interdiction shifts the supply curve to the left from S_1 to S_2 without changing the demand curve. The equilibrium price of drugs rises from P_1 to P_2, and the equilibrium quantity falls from Q_1 to Q_2. The fall in the equilibrium quantity shows that drug interdiction reduces drug use.

But what about the amount of drug-related crime? Consider the total amount that drug users pay for the drugs they buy. Because few addicts are likely to quit in response to a higher price, it is likely that the demand for drugs is inelastic, as shown in the figure. If demand is inelastic, then an increase in the price raises total revenue in the drug market. That is, because drug interdiction raises the price of drugs proportionately more than it reduces drug use, it increases the total amount of money that drug users pay for drugs. Addicts would now have an even greater

Figure 9

Policies to Reduce the Use of Illegal Drugs

Drug interdiction reduces the supply of drugs from S_1 to S_2, as in panel (a). If the demand for drugs is inelastic, then the total amount paid by drug users rises, even as the amount of drug use falls. By contrast, drug education reduces the demand for drugs from D_1 to D_2, as in panel (b). Because both price and quantity fall, the amount paid by drug users falls.



need for quick cash. The result is inescapable: Drug interdiction could increase drug-related crime. (And that doesn't even consider the well-documented economic and social damage resulting from the enforcement of drug laws on certain communities, particularly among people of color.)

Because of the adverse effect of drug interdiction, some analysts argue for alternative approaches. One is the legalization of some less dangerous drugs, such as marijuana. For drugs that remain illegal, rather than trying to reduce the supply, policymakers might try to reduce the demand through drug education. Successful drug education has the effects shown in panel (b) of Figure 9. The demand curve shifts to the left from D_1 to D_2. As a result, the equilibrium quantity falls from Q_1 to Q_2, and the equilibrium price falls from P_1 to P_2. Total revenue, $P \times Q$, also falls. In contrast to drug interdiction, drug education can reduce both drug use and drug-related crime.

Advocates of drug interdiction might argue that the long-run effects of this policy are different from the short-run effects because the elasticity of demand depends on the time horizon. The demand for drugs is probably inelastic over short periods because higher prices do not substantially affect drug use by established addicts. But it may be more elastic over longer periods because higher prices would discourage experimentation with drugs among the young and, over time, lead to fewer drug addicts. In this case, drug interdiction would increase drug-related crime in the short run but decrease it in the long run.

In the News

Elasticity of Supply and Demand in the Ride-share Market

According to economist Austan Goolsbee, the supply of ride-shares is more elastic than the demand, and this fact holds the key to the future of this market.

Passengers May Pay a Lot More. Drivers Won't Accept Much Less.

By Austan Goolsbee

Uber and Lyft, the two leading ride-share companies, have lost a great deal of money and don't project a profit any time soon.

Yet they are both trading on public markets with a combined worth of more than $80 billion. Investors presumably expect that these companies will someday find a path to profitability, which leaves us with a fundamental question: Will that extra money come mainly from higher prices paid by consumers or from lower wages paid to drivers?

Old-fashioned economics provides an answer: Passengers, not drivers, are likely to be the main source of financial improvement, at least within the next few years, mainly because of something called "relative price sensitivity."

This conclusion may seem to run counter to popular wisdom. Wall Street analysts have suggested that Uber and Lyft will need to squeeze their drivers. Those workers are quite concerned about the possibility. Thousands went on a one-day strike before Uber's initial public offering in May to demand higher pay and more benefits.

And Lyft, in documents filed in connection with its own I.P.O., said it hoped to use autonomous vehicles, which don't need a wage, for a majority of its rides within 10 years. But rather than debate the plausibility—or cost—of amassing an autonomous fleet in a decade, let's consider what is possible over the short term.

Economic theory predicts that sensitivity to price changes determines who will pay more. And it turns out that passengers aren't very sensitive to price, while drivers are.

Yes, surge pricing—the practice of raising prices when demand is high—makes many people **feel** irate. But what people actually **do** is what is important. The most comprehensive study of rider behavior in the marketplace found that riders didn't change their behavior much when prices surged. (Like most major quantitative studies about Uber, it relied on the company's data and included the participation of an Uber employee.)

Passengers were what economists call "inelastic," meaning demand for rides fell by less than prices rose. For every 10 percent increase in price, demand fell by only about 5 percent.

Drivers, on the other hand, are quite sensitive to prices—that is, their wages—largely because there are so many people who are ready to start driving at any time. If prices change, people enter or exit the market, pushing the average wage to what is known as the "market rate."

That's what always happens when there are no barriers to entry in a market. In 1848, for example, at the start of the California gold rush, the first miners made about $20 per day, on average. The historical data shows that was at least 10 times more than the wage for workers doing what I would classify as similar activities—stone cutting and brick laying—in New York at that time.

QuickQuiz

8. An increase in the supply of grain will reduce the total revenue grain producers receive if
   a. the supply curve is inelastic.
   b. the supply curve is elastic.
   c. the demand curve is inelastic.
   d. the demand curve is elastic.
9. In competitive markets, farmers adopt new technologies that will eventually reduce their revenue because
   a. each farmer is a price taker.
   b. farmers are short-sighted.
   c. regulation requires the use of best practices.
   d. consumers pressure farmers to lower prices.
10. Because the demand curve for oil is _____ elastic in the long run, OPEC's reduction in the supply of oil had a _____ impact on the price in the long run than it did in the short run.
    a. less; smaller
    b. less; larger
    c. more; smaller
    d. more; larger

Over the next eight years, so many people moved to California searching for gold that miners' average earnings fell to $3 a day, minus expenses—barely more than they could have made if they had been cutting stones in New York.

What killed the gold rush wasn't the lack of gold—production tripled over that time. It was the entry of so many competing miners that drove average earnings down so low that most of them barely made enough to stay in business.

And so it is with ride-share drivers today. Another study, by a New York University professor and two Uber employees, found the same dynamic: Higher prices increased driver incomes, but only for a few weeks.

As new drivers entered the market, attracted by higher wages, the average driver had to spend more time waiting for fares. Average pay returned to the level economists refer to as "the outside option"—the pay level of whatever else the drivers could be doing if they weren't driving for Uber or Lyft.

If for many ride-share drivers the next best option is delivering for an outfit like Domino's Pizza, or working at a fast-food restaurant, then average pay for the drivers will likely end up around minimum wage, too.

Some of this is just educated guesswork. It's not as easy to measure drivers' average wages—and, therefore, their price sensitivity—as you might think. Since drivers pay their own fuel and depreciation expenses, we need to subtract those costs from their earnings, but we don't have good data.

Still, a major survey of drivers done last year by the analysis firm Ridester showed average raw earnings for UberX drivers of about $15 per hour, before projected deductions of about $8 per hour.

The bad news for the drivers, then, is that average pay will be low. Also, even if they convinced ride-share companies to raise the share of revenue the drivers keep and to increase benefits, the earnings boost would likely be ephemeral. Thousands of new drivers would enter the more lucrative market, bringing average earnings back down to the market rate.

The good news for drivers, though, is that it won't be easy for ride-share companies to cut wages much lower. Many drivers will simply stop driving if wages fall.

One of the most important studies of driver behavior (conducted by professors from Yale and U.C.L.A. and, again, one Uber employee), confirms the sensitivity of drivers to earnings changes. On average, they increase their hours by 20 percent in response to a 10 percent increase in wages. That is about four times larger than the response by passengers to changes in the price of a ride.

Economics says that the likelihood that a person will bear the burden of an increase in profit margins is inversely proportional to their price sensitivity. In other words, because drivers are four times more price sensitive than riders, a reasonable guess is that 80 percent of the price burden will fall on passengers, 20 percent on drivers. Uber and Lyft are still building their networks and market share, which complicates matters, and may delay price increases.

Nonetheless, I think that as a passenger, you should take your Uber and Lyft rides now, while they're still relatively cheap. And, if you're an investor counting on wage cuts and robots to carry Uber and Lyft to profit nirvana, you may want to buy a certain bridge first. ■

Questions to Discuss

1. If the price of ride shares rose by 10 percent, how much would your use of them decline? What is your price elasticity of demand?
2. Why do you think that supply is more elastic than demand in this market?

Mr. Goolsbee is a professor of economics at the University of Chicago's Booth School of Business.

Source: *New York Times*, June 2, 2019.

11. Over time, technological advances increase consumers' incomes and reduce the price of smartphones. Each of these forces increases the amount consumers spend on smartphones if the income elasticity of demand is greater than _____ and the price elasticity of demand is greater than _____.
    a. zero; zero
    b. zero; one
    c. one; zero
    d. one; one

Answers are at the end of the chapter.

5-4 Conclusion

Even a parrot can become an economist simply by learning to say "supply and demand." That's an old joke, but these last two chapters should have convinced you that there is much truth to it. The tools of supply and demand are useful for analyzing the events and policies that shape the economy. You are now well on your way to becoming an economist (or at least a well-educated parrot).

Chapter in a Nutshell

- The price elasticity of demand measures how much the quantity demanded responds to price changes. Demand tends to be more elastic if close substitutes are available, if the good is a luxury rather than a necessity, if the market is narrowly defined, or if buyers have substantial time to react to a price change.
- The price elasticity of demand is calculated as the percentage change in quantity demanded divided by the percentage change in price. If the quantity demanded moves proportionately less than the price, then the elasticity is less than one, and demand is inelastic. If the quantity demanded moves proportionately more than the price, then the elasticity is greater than one, and demand is elastic.
- Total revenue, the total amount paid for a good, equals the price times the quantity sold. For inelastic demand curves, total revenue moves in the same direction as the price. For elastic demand curves, total revenue moves in the opposite direction.
- The income elasticity of demand measures how much the quantity demanded responds to changes in consumers' income. The cross-price elasticity of demand measures how much the quantity demanded of one good responds to changes in the price of another.
- The price elasticity of supply measures how much the quantity supplied responds to changes in the price. This elasticity often depends on the time horizon: In most markets, supply is more elastic in the long run than in the short run.
- The price elasticity of supply is calculated as the percentage change in quantity supplied divided by the percentage change in price. If the quantity supplied moves proportionately less than the price, then the elasticity is less than one, and supply is inelastic. If the quantity supplied moves proportionately more than the price, then the elasticity is greater than one, and supply is elastic.
- The tools of supply and demand can be applied to many different markets. This chapter uses them to analyze the market for wheat, the market for oil, and the market for illegal drugs.

Key Concepts

elasticity, p. 88
price elasticity of demand, p. 88
total revenue, p. 91
income elasticity of demand, p. 96
cross-price elasticity of demand, p. 96
price elasticity of supply, p. 97

Questions for Review

1. Define the price elasticity of demand and the income elasticity of demand.
2. List and explain the four determinants of the price elasticity of demand discussed in the chapter.
3. If the elasticity is greater than one, is demand elastic or inelastic? If the elasticity equals zero, is demand perfectly elastic or perfectly inelastic?
4. On a supply-and-demand diagram, show the equilibrium price, equilibrium quantity, and total revenue received by producers.
5. If demand is elastic, how will an increase in the price affect total revenue? Explain.
6. What do we call a good with an income elasticity less than zero?
7. How is the price elasticity of supply calculated? Explain what it measures.
8. If a fixed quantity of a good is available, and no more can be made, what is the price elasticity of supply?
9. A storm destroys half the fava bean crop. Is this event more likely to hurt fava bean farmers if the demand for fava beans is very elastic or very inelastic? Explain.

Problems and Applications

1. For each of the following pairs of goods, which good would you expect to have more elastic demand and why?
   a. required textbooks or mystery novels
   b. Billie Eilish recordings or pop music recordings in general
   c. subway rides during the next six months or subway rides during the next five years
   d. root beer or water

2. Suppose that business travelers and vacationers have the following demand for airline tickets from Chicago to Miami:

Price	Quantity Demanded (business travelers)	Quantity Demanded (vacationers)
$150	2,100 tickets	1,000 tickets
200	2,000	800
250	1,900	600
300	1,800	400

   a. As the price of tickets rises from $200 to $250, what is the price elasticity of demand for (i) business travelers and (ii) vacationers? (Use the midpoint method in your calculations.)
   b. Why might vacationers and business travelers have different elasticities?

3. Suppose the price elasticity of demand for heating oil is 0.2 in the short run and 0.7 in the long run.
   a. If the price of heating oil rises from $1.80 to $2.20 per gallon, what happens to the quantity of heating oil demanded in the short run? In the long run? (Use the midpoint method in your calculations.)
   b. Why might this elasticity depend on the time horizon?

4. A price change causes the quantity demanded of a good to decrease by 30 percent, while the total revenue of that good increases by 15 percent. Is the demand curve elastic or inelastic? Explain.

5. Cups of coffee and donuts are complements. Both have inelastic demand. A hurricane destroys half the coffee bean crop. Use appropriately labeled diagrams to answer the following questions.
   a. What happens to the price of coffee beans?
   b. What happens to the price of a cup of coffee? What happens to total expenditure on cups of coffee?
   c. What happens to the price of donuts? What happens to total expenditure on donuts?

6. The price of aspirin rose sharply last month, while the quantity sold remained the same. Five people suggest various diagnoses of the phenomenon:

   Meredith: Demand increased, but supply was perfectly inelastic.

   Alex: Demand increased, but it was perfectly inelastic.

   Miranda: Demand increased, but supply decreased at the same time.

   Richard: Supply decreased, but demand was unit elastic.

   Owen: Supply decreased, but demand was perfectly inelastic.

   Who could possibly be right? Use graphs to explain your answer.

7. Suppose that your demand schedule for pizza is as follows:

Price	Quantity Demanded (income = $20,000)	Quantity Demanded (income = $24,000)
$8	40 pizzas	50 pizzas
10	32	45
12	24	30
14	16	20
16	8	12

   a. Use the midpoint method to calculate your price elasticity of demand as the price of pizza increases from $8 to $10 if (i) your income is $20,000 and (ii) your income is $24,000.
   b. Calculate your income elasticity of demand as your income increases from $20,000 to $24,000 if (i) the price is $12 and (ii) the price is $16.

8. The *New York Times* reported (Feb. 17, 1996) that subway ridership declined after a fare increase: "There were nearly four million fewer riders in December 1995, the first full month after the price of a token increased 25 cents to $1.50, than in the previous December, a 4.3 percent decline."
   a. Use these data to estimate the price elasticity of demand for subway rides.
   b. According to your estimate, what happened to the Transit Authority's revenue when the fare rose?
   c. Why might your estimate of the elasticity be unreliable?

9. Two drivers, Thelma and Louise, each drive up to a gas station. Before looking at the price, each places an order. Thelma says, "I'd like 5 gallons of gas." Louise says, "I'd like $20 worth of gas." What is each driver's price elasticity of demand?

10. Consider public policy aimed at smoking.
    a. Studies indicate that the price elasticity of demand for cigarettes is about 0.4. If a pack of cigarettes currently costs $5 and the government wants to reduce smoking by 20 percent, by how much should it increase the price?
    b. If the government permanently increases the price of cigarettes, will the policy have a larger effect on smoking one year from now or five years from now?
    c. Studies also find that teenagers have a higher price elasticity of demand than adults. Why might this be true?

11. You are the curator of a museum. It is running short of funds, so you would like to increase revenue. Should you increase or decrease the price of admission? Explain.

12. Explain why the following might be true: A drought around the world raises the total revenue that farmers receive from the sale of grain, but a drought only in Kansas reduces the total revenue that Kansas farmers receive.

Quick**Quiz Answers**

1. **a** 2. **d** 3. **d** 4. **c** 5. **b** 6. **c** 7. **c** 8. **c** 9. **a** 10. **c** 11. **b**

Chapter

6

Supply, Demand, and Government Policies

Economists have many roles. As scientists, they develop and test theories to explain the world around them. As policy analysts and advisers, they try to use these theories to change the world. The focus of the preceding two chapters has been scientific. The theory of supply and demand explains the relationships between the prices of goods and the quantities sold. When various events shift supply and demand, the equilibrium price and quantity change. The concept of elasticity helps to gauge the size of these changes. This theory is the foundation for much of economics.

This chapter is about policy. Here, we analyze several types of government policy using the tools of supply and demand, with some surprising insights. Policies often have effects that their architects did not anticipate.

Efforts to control prices are worthy of close consideration. In this category, we examine rent-control laws, which set a maximum fee that landlords may charge tenants, and minimum-wage laws, which set a pay threshold below which employers must not go. Policymakers often enact price controls when they believe that the market price of a good or service is too high or too low. Yet these policies can generate problems of their own.

ROMAN SAMBORSKYI/SHUTTERSTOCK.COM, ARNO VAN ENGELAND/EYEEM/GETTY IMAGES

After price controls, we consider the impact of taxes. Policymakers use taxes to raise revenue and to influence market outcomes. The prevalence of taxes in the economy is obvious, but their effects are not. For example, when the government levies a tax on the amount that firms pay their workers, do the firms or workers bear the burden of the tax? The answer is not clear—until we apply the powerful tools of supply and demand.

6-1 The Surprising Effects of Price Controls

To see how price controls affect market outcomes, let's return to the market for ice cream. As we saw in Chapter 4, if ice cream is sold in a competitive market, the price normally adjusts to balance supply and demand: At the equilibrium price, the quantity of ice cream that buyers want to buy exactly equals the quantity that sellers want to sell. To be concrete, suppose that the equilibrium price is $3 per cone.

Some people may not like this outcome. The American Association of Ice-Cream Eaters complains that the $3 price is too high for everyone to enjoy a cone a day (their recommended daily allowance). Meanwhile, the National Organization of Ice-Cream Makers complains that the $3 price—the result of "cutthroat competition"—is so low that it is depressing the incomes of its members. Each group lobbies the government to alter the market outcome by passing laws that control the price of an ice-cream cone.

Because buyers usually want a lower price while sellers want a higher one, the interests of the two groups conflict. If the Ice-Cream Eaters are successful in their lobbying, the government imposes a legal maximum on the price at which ice-cream cones can be sold. Because the price is not allowed to rise above this level, the legislated maximum is called a **price ceiling**. By contrast, if the Ice-Cream Makers are successful, the government imposes a legal minimum on the price. Because the price cannot fall below this level, the legislated minimum is called a **price floor**.

price ceiling
a legal maximum on the price at which a good can be sold

price floor
a legal minimum on the price at which a good can be sold

6-1a How Price Ceilings Affect Market Outcomes

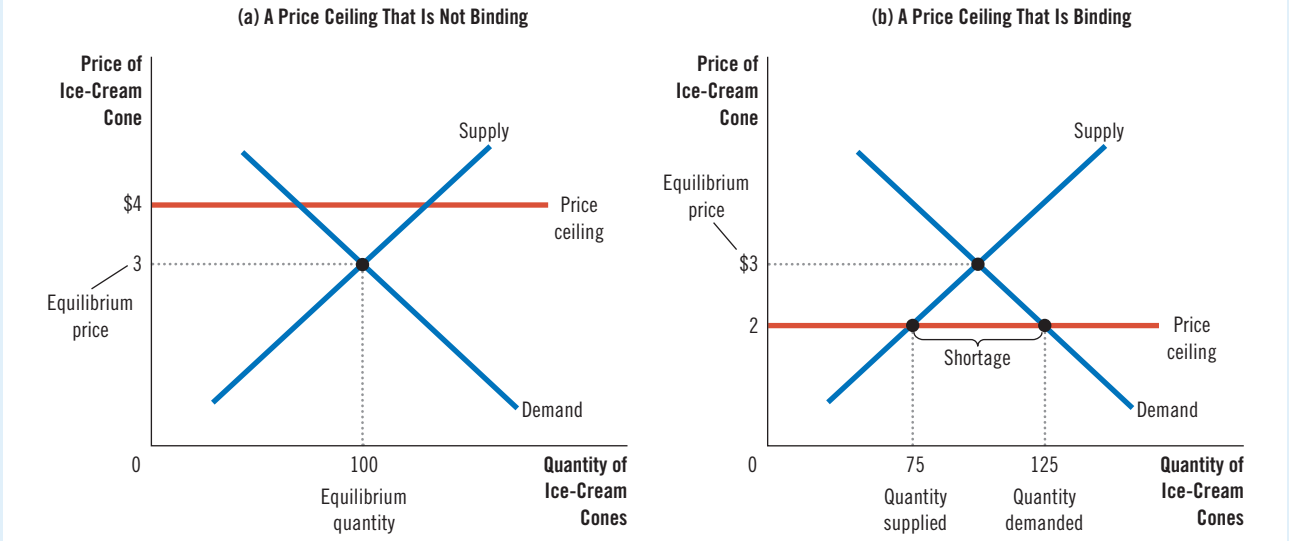
When the government, moved by the complaints and campaign contributions of the Ice-Cream Eaters, imposes a price ceiling in the market for ice cream, two outcomes are possible. In panel (a) of Figure 1, the government imposes a price ceiling of $4 per cone. In this case, because the price that balances supply and demand ($3) is below the ceiling, the price ceiling is **not binding**. Market forces move the economy to the equilibrium, and the ceiling has no effect on the price or on the quantity sold.

Panel (b) of Figure 1 shows another, more interesting possibility. In this case, the government imposes a price ceiling of $2 per cone. Because the equilibrium price of $3 is above the price ceiling, the ceiling is a **binding constraint** on the market. The forces of supply and demand tend to move the price toward equilibrium, but the ceiling prevents the market price from reaching it. Instead, the market price must be the price ceiling. At this price, the quantity of ice cream demanded (125 cones in the figure) exceeds the quantity supplied (75 cones). With excess demand of 50 cones, some people who want ice cream at the going price can't buy it. The price ceiling has created an ice-cream shortage.

Figure 1

A Market with a Price Ceiling

In panel (a), the government imposes a price ceiling of $4. Because it is above the equilibrium price of $3, the ceiling has no effect, and the market can reach the equilibrium of supply and demand. At this point, quantity supplied and quantity demanded both equal 100 cones. In panel (b), the government imposes a price ceiling of $2. Because the ceiling is below the equilibrium price of $3, the market price is $2. At this price, 125 cones are demanded while only 75 are supplied, so there is a shortage of 50 cones.



In response to the shortage, a mechanism for rationing ice cream will naturally develop. It could be long lines: Buyers who arrive early and wait in line (or pay others to do so) get a cone, while those who can't or won't do this must go without. Another possibility is that sellers ration ice-cream cones according to their own personal biases, selling only to friends, relatives, members of their own racial or ethnic group, or those who provide favors in return. Clearly, even though the price ceiling was intended to help buyers of ice cream, not all buyers benefit from the policy. Some buyers pay a lower price, though they may have to wait in line to do so, but others cannot get any ice cream at all.

This illustrates a general result: **When the government imposes a binding price ceiling on a competitive market, a shortage arises, and sellers must ration scarce goods among potential buyers.** The rationing mechanisms that develop under price ceilings are rarely desirable. Long lines are inefficient because they waste buyers' time. Relying on the biases of sellers is both inefficient (because the good may not go to the buyer who values it most) and unfair. By contrast, the rationing mechanism in a free, competitive market is straightforward. When the market reaches its equilibrium, anyone who wants to pay the market price can buy the good. This may seem unfair to some buyers when prices are high, but it is efficient and impersonal. You don't need to be the ice-cream maker's friend or relative to buy a cone. You just need to be able and willing to pay $3.

Case Study

How to Create Long Lines at the Gas Pump

Chapter 5 discussed how, in 1973, the Organization of Petroleum Exporting Countries (OPEC) reduced production of crude oil and increased its price. Because crude oil is used to make gasoline, the higher oil prices reduced the supply of gasoline. Long lines at gas stations became common, with motorists often waiting for hours to buy a few gallons of gas.

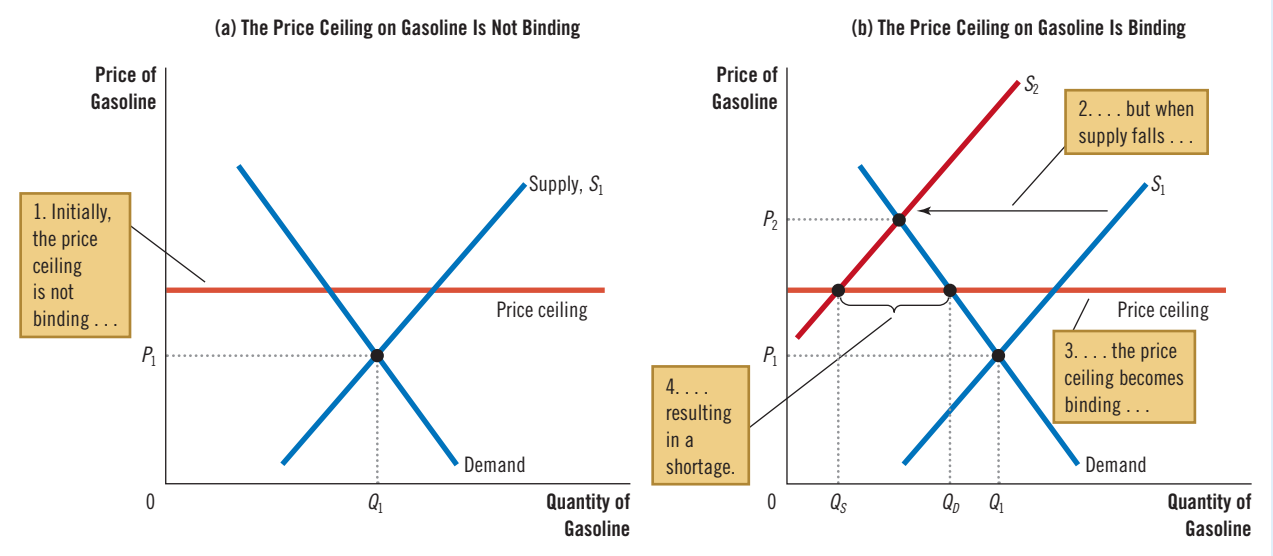
What caused the long gas lines? Most people blamed OPEC. To be sure, if it had not reduced production of crude oil, the gasoline shortage would not have occurred. Yet economists found another culprit: U.S. government regulations that set a ceiling on the price of gasoline.

Figure 2 reveals what happened. As panel (a) shows, before OPEC raised the price of crude oil, the equilibrium price of gasoline, P_1, was below the price ceiling. The price regulation, therefore, had no effect. When the price of crude oil rose, however, the situation changed. The increase in the price of crude oil raised the cost of producing gasoline and thereby reduced the supply of gasoline. As panel (b) shows, the supply curve shifted to the left from S_1 to S_2. In an unregulated market, this shift in supply would have raised the equilibrium price of gasoline from P_1 to P_2, and no shortage would have occurred. Instead, the price ceiling prevented the price from rising to the equilibrium level. At the price ceiling, producers were willing to sell Q_S, but consumers were willing to buy Q_D. The supply shift caused a severe shortage at the regulated price.

Figure 2

The Market for Gasoline with a Price Ceiling

Panel (a) shows the gasoline market when the price ceiling is not binding because the equilibrium price, P_1, is below the ceiling. Panel (b) shows the gasoline market after an increase in the price of crude oil (an input into making gasoline) shifts the supply curve to the left from S_1 to S_2. In an unregulated market, the price would have risen from P_1 to P_2. The price ceiling, however, prevents this from happening. At the binding price ceiling, consumers are willing to buy Q_D, but producers of gasoline are willing to sell only Q_S. The difference between quantity demanded and quantity supplied, $Q_D - Q_S$, measures the gasoline shortage.



Eventually, the laws regulating the price of gasoline were repealed. Lawmakers came to understand that they were partly responsible for the many hours Americans lost waiting in line to buy gasoline. Today, when the price of crude oil changes, the price of gasoline adjusts freely to bring supply and demand into equilibrium. ●

Case Study

Why Rent Control Causes Housing Shortages, Especially in the Long Run

In many cities, the local government places a ceiling on rents that landlords may charge their tenants. This is rent control, a policy aimed at helping the poor by keeping housing costs low. Yet economists often criticize rent control, saying that it is a highly inefficient way to help the poor. One economist went so far as to call rent control "the best way to destroy a city, other than bombing."

The adverse effects of rent control may not be apparent because these effects occur over many years. In the short run, landlords have a fixed number of apartments to rent, and they cannot adjust this number quickly as market conditions change. Moreover, the number of people looking for apartments may not be highly responsive to rents in the short run because people take time to adjust their housing arrangements. In other words, the short-run supply and demand for housing are both relatively inelastic.

Panel (a) of Figure 3 shows the short-run effects of rent control on the housing market. As with any binding price ceiling, rent control causes a shortage. But

Figure 3

Rent Control in the Short Run and in the Long Run

Panel (a) shows the short-run effects of rent control: Because the supply and demand curves for apartments are relatively inelastic, the price ceiling imposed by a rent-control law causes only a small shortage of housing. Panel (b) shows the long-run effects of rent control: Because the supply and demand curves for apartments are more elastic, rent control causes a larger shortage.

because supply and demand are inelastic in the short run, the initial shortage is small. The primary result in the short run is popular among tenants: a reduction in rents.

The long-run story is very different because the buyers and sellers of rental housing respond more to market conditions as time passes. On the supply side, landlords respond to low rents by not building new apartments and by failing to maintain existing ones. On the demand side, low rents encourage people to find their own apartments (rather than live with roommates or their parents) and to move into the city. Therefore, both supply and demand are more elastic in the long run.

Panel (b) of Figure 3 illustrates the housing market in the long run. When rent control depresses rents below the equilibrium level, the quantity of apartments supplied falls substantially, and the quantity of apartments demanded rises substantially. The result is a large shortage of housing.

In cities with rent control, landlords and building superintendents use various mechanisms to ration housing. Some keep long waiting lists. Others give preference to tenants without children. Still others discriminate based on race. Sometimes, apartments are allocated to those willing to offer under-the-table payments; these bribes bring the total price of an apartment closer to the equilibrium price.

Recall one of the **Ten Principles of Economics** from Chapter 1: People respond to incentives. In well-functioning markets, landlords can command higher prices if they keep their buildings clean and safe. But when rent control creates shortages and waiting lists, landlords lose that incentive. Why spend money to maintain and improve the property when people are waiting to move in as it is? In the end, rent control reduces what tenants have to pay, but it also lowers the quantity and quality of a city's housing stock.

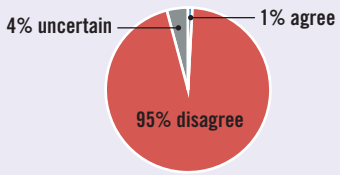
When these adverse effects become evident, policymakers often react by imposing additional regulations. For example, various laws make racial discrimination in housing illegal and require landlords to provide minimally adequate living conditions. These laws, however, are difficult and costly to enforce. By contrast, without rent control, such laws are less necessary because the market for housing is regulated by the forces of competition. If the price of housing were permitted to increase to the equilibrium level, the shortages that give rise to undesirable landlord behavior would be largely eliminated. ●

Ask the Experts

Rent Control

"Local ordinances that limit rent increases for some rental housing units, such as in New York and San Francisco, have had a positive impact over the past three decades on the amount and quality of broadly affordable rental housing in cities that have used them."

What do economists say?



Source: IGM Economic Experts Panel, February 7, 2012.

6-1b How Price Floors Affect Market Outcomes

To examine the effects of another kind of government price control, let's return to the market for ice cream. Imagine now that the National Organization of Ice-Cream Makers persuades the government that the $3 equilibrium price is too low. In this case, the government might institute a price floor. Price floors, like price ceilings, are an attempt by the government to maintain prices at other than equilibrium levels. While a price ceiling places a legal maximum on prices, a price floor places a legal minimum.

When the government imposes a price floor on ice cream, two outcomes are possible. If the floor is $2 per cone but the equilibrium price is $3, nothing happens. Because the equilibrium price is above the floor, the price floor is not binding. Market forces move the economy to the equilibrium, and the price floor has no effect. Panel (a) of Figure 4 shows this outcome.

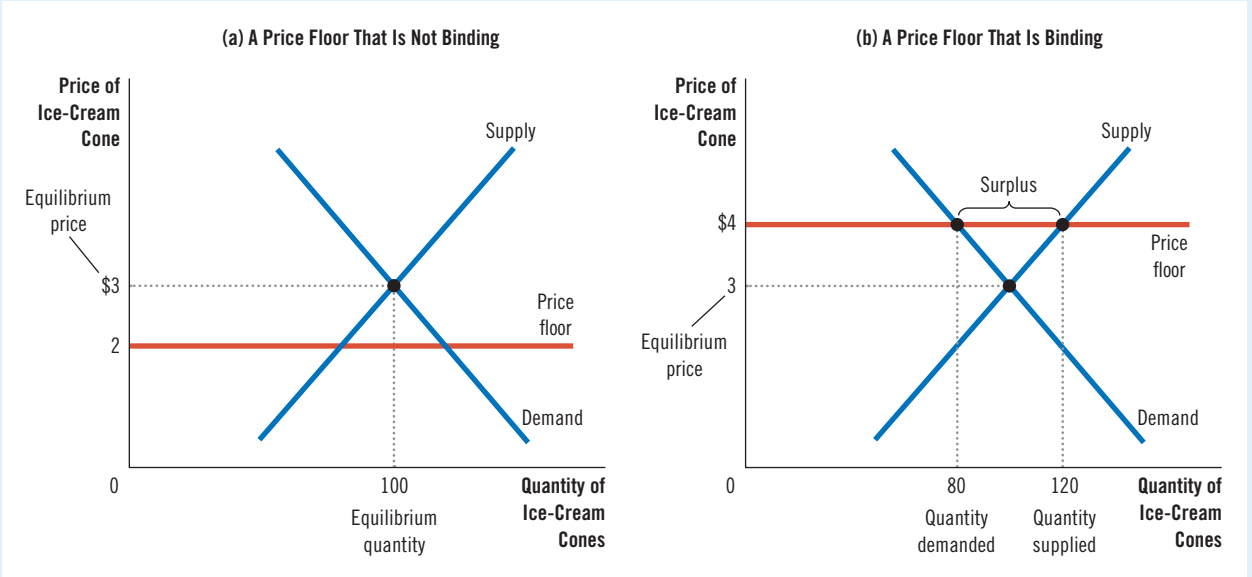
Panel (b) of Figure 4 shows what happens when the government imposes a price floor of $4 per cone, which is higher than the equilibrium price of $3. In this case, the price floor is a binding constraint on the market. The forces of supply and demand tend to move the price toward the equilibrium price, but the price can't go below the floor. As a result, the price floor becomes the market price. At this level, the quantity of ice cream supplied (120 cones) exceeds the quantity demanded (80 cones). There is an excess supply of 40 cones. In other words, some people who want to sell ice cream at the going price have no buyers: **A binding price floor causes a surplus.**

Just as shortages caused by price ceilings can lead to undesirable rationing mechanisms, so can the surpluses resulting from price floors. The sellers who appeal to the buyers' personal biases may be better able to sell their goods than those who do not. By contrast, in a free market, the price is the rationing mechanism. Sellers may not be happy about how much they are paid at the equilibrium price, but they can sell all they want.

Figure 4

A Market with a Price Floor

In panel (a), the government imposes a price floor of $2. Because it is below the equilibrium price of $3, the floor has no effect, and the market can reach the equilibrium of supply and demand. At this point, quantity supplied and demanded both equal 100 cones. In panel (b), the government imposes a price floor of $4. Because the floor is above the equilibrium price of $3, the market price is $4. At this price, 120 cones are supplied while only 80 are demanded, so there is a surplus of 40 cones.



Case Study

Controversies over the Minimum Wage

The minimum wage is an important and contentious example of a price floor. Minimum-wage laws set the lowest price for labor that any employer may pay. The U.S. Congress first instituted a minimum wage with the Fair Labor Standards Act of 1938 to ensure workers a minimally adequate standard of living.

In 2021, the minimum wage according to federal law was $7.25 per hour. In addition, many states and cities mandate minimum wages above the federal level. The minimum wage in Seattle, for instance, was $16.69 per hour for large employers in 2021. Most European nations also have laws that establish a minimum wage, often much higher than in the United States. For example, even though the average income in France is almost 30 percent lower than it is in the United States, the French minimum wage is more than 50 percent higher.

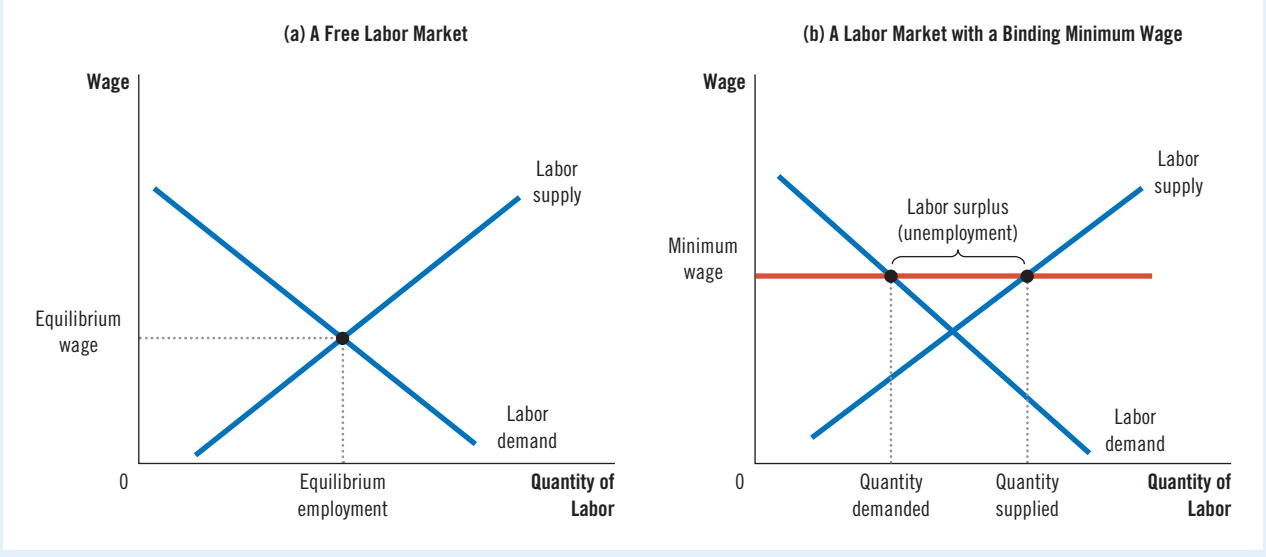
To see what the theory of supply and demand predicts for the effects of a minimum wage, consider the market for labor. Panel (a) of Figure 5 shows a competitive labor market, which, like all competitive markets, is subject to the forces of supply and demand. Workers supply labor, and firms demand labor. If the government doesn't intervene, the wage adjusts to balance labor supply and labor demand.

Panel (b) of Figure 5 shows the labor market with a minimum wage. If the minimum wage is above the equilibrium level, as it is here, the quantity of labor supplied exceeds the quantity demanded. The result is a surplus of labor, or unemployment. While the minimum wage raises the incomes of those workers who have jobs, it lowers the incomes of would-be workers who now cannot find jobs.

Figure 5

How the Minimum Wage Affects a Competitive Labor Market

Panel (a) shows a labor market in which the wage adjusts to balance labor supply and labor demand. Panel (b) shows the impact of a binding minimum wage. Because the minimum wage is a price floor, it causes a surplus: The quantity of labor supplied exceeds the quantity demanded. The result is unemployment.



To fully understand the minimum wage, keep in mind that the economy contains not a single labor market but many labor markets for different types of workers. The impact of the minimum wage depends on the skill and experience of the worker. Highly skilled and experienced workers are not affected because their equilibrium wages are well above the minimum. For these workers, the minimum wage is not binding.

The minimum wage has its greatest impact on the market for teenage labor. The equilibrium wages of teenagers are low because teenagers are among the least skilled and least experienced members of the labor force. In addition, teenagers are often willing to accept a lower wage in exchange for on-the-job training. (Some teenagers are willing to work as interns for no pay at all. Because many internships pay nothing, minimum-wage laws often do not apply to them. If they did, some of these internship opportunities might not exist.) As a result, the minimum wage is binding more often for teenagers than for other members of the labor force.

Many economists have studied how minimum-wage laws affect the teenage labor market. These researchers compare the changes in the minimum wage over time with the changes in teenage employment. Although there is some debate about the effects of minimum wages, the typical study finds that a 10 percent increase in the minimum wage depresses teenage employment by 1 to 3 percent.

One drawback of most minimum-wage studies is that they focus on the effects over short periods. For example, they might compare employment the year before and the year after a change in the minimum wage. The longer-term effects on employment are harder to estimate reliably, but they are more relevant for evaluating the policy. Because it takes time for firms to reorganize the workplace, the long-run decline in employment from a higher minimum wage may be larger than the estimated short-run decline.

In addition to altering the quantity of labor demanded, the minimum wage alters the quantity supplied. Because the minimum wage raises the wage that teenagers can earn, it increases the number of teenagers who choose to look for jobs. Some studies have found that a higher minimum wage also influences which teenagers are employed. When the minimum wage rises, some teenagers who are still attending high school choose to drop out and take jobs. With more people vying for the available jobs, some of these new dropouts displace other teenagers who had already dropped out of school, and these displaced teenagers become unemployed.

The minimum wage is a frequent topic of debate. Advocates of a higher minimum wage view the policy as a humane way to raise the income of the working poor. They correctly point out that workers who earn the minimum wage can afford only a meager standard of living. In 2021, for instance, when the minimum wage was $7.25 per hour, two adults working 40 hours a week for every week of the year at minimum-wage jobs had a joint annual income of only $30,160. This amount was only about 40 percent of the median family income in the United States. Some proponents of a higher minimum wage contend that labor markets are not well explained using the theory of supply and demand in competitive markets, so they doubt the theory's predictions regarding unemployment. Others acknowledge that the policy has some adverse effects, including job loss, but say these effects are small and that, all things considered, a higher minimum wage makes the poor better off.

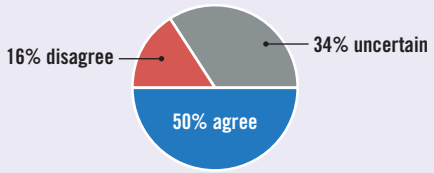
Opponents of raising the minimum wage contend that it is not the best way to combat poverty. They say that a high minimum wage causes unemployment,

Ask the Experts

The Minimum Wage

"The current US federal minimum wage is $7.25 per hour. States can choose whether to have a higher minimum—and many do. A federal minimum wage of $15 per hour would lower employment for low-wage workers in many states."

What do economists say?



Source: IGM Economic Experts Panel, February 2, 2021.

encourages teenagers to drop out of school, and results in some unskilled workers not getting on-the-job training. Moreover, opponents of raising the minimum wage note that it is a poorly targeted policy. Less than a third of minimum-wage earners are in families with incomes below the poverty line. Many are teenagers from middle-class homes working at part-time jobs for extra spending money.

In 2021, President Biden proposed increasing the minimum wage to $15 per hour by 2025. "No one should work 40 hours a week and live in poverty," he said. In February 2021, the Congressional Budget Office, a government agency staffed by nonpartisan policy analysts, released a study of the proposal. They estimated that it would increase the wages of 17 million people, lift 900,000 out of poverty, and put 1.4 million out of work. As this book went to press, Congress had not yet enacted the Biden proposal. ●

6-1c Evaluating Price Controls

One of the **Ten Principles of Economics** in Chapter 1 is that markets are usually a good way to organize economic activity. It is why economists often oppose price ceilings and price floors. To economists, prices are not the outcome of some haphazard process. Prices, they say, are the result of millions of business and consumer decisions that lie behind the supply and demand curves. Prices have the crucial job of balancing supply and demand and, thereby, coordinating economic activity. Government price-setting obscures the signals that would otherwise guide the allocation of society's resources.

That's just one side of the story. Another of the **Ten Principles of Economics** is that governments can sometimes improve market outcomes. Indeed, policymakers are often motivated to control prices because they view the market's outcome as unfair. Price controls are frequently aimed at helping the poor. For instance, rent-control laws try to make housing affordable for everyone, and minimum-wage laws try to help people escape poverty.

Yet price controls can hurt some people they are intended to help. Rent control keeps rents low, but it also discourages landlords from maintaining their buildings and makes housing hard to find. Minimum-wage laws raise the incomes of some workers, but they can also lead to job losses for others.

Helping those in need can be accomplished in ways other than controlling prices. For instance, the government can make housing more affordable by paying a fraction of the rent for poor families or by giving them cash transfers so they can pay the rent themselves. Unlike rent control, such subsidies do not reduce the quantity of housing supplied and, therefore, do not lead to housing shortages. Similarly, wage subsidies raise the living standards of the working poor without discouraging firms from hiring them. An example of a wage subsidy is the **earned income tax credit**, a government program that supplements the incomes of low-wage workers.

These alternative policies are often better than price controls, but they are not perfect. Applying for rent or wage subsidies can be a burden for poor people. In addition, rent and wage subsidies cost the government money and, therefore, require higher taxes. As the next section shows, taxation has costs of its own.

QuickQuiz

1. When the government imposes a binding price floor, it causes
   a. the supply curve to shift to the left.
   b. the demand curve to shift to the right.
   c. a shortage of the good to develop.
   d. a surplus of the good to develop.
2. In a market with a binding price ceiling, increasing the ceiling price will
   a. increase the surplus.
   b. increase the shortage.
   c. decrease the surplus.
   d. decrease the shortage.
3. Rent control causes larger shortages in the _____ run because over that time horizon, supply and demand are _____ elastic.
   a. long; more
   b. long; less
   c. short; more
   d. short; less
4. An increase in the minimum wage reduces the total amount paid to the affected workers if the price elasticity of _____ is _____ than one.
   a. supply; greater
   b. supply; less
   c. demand; greater
   d. demand; less

Answers are at the end of the chapter.

6-2 The Surprising Study of Tax Incidence

All governments—from national governments around the world to local governments in small towns—use taxes to raise revenue for public projects, such as roads, schools, and national defense. Because taxes are such an important policy instrument and affect our lives in many ways, they appear throughout this book. This section begins our study of how taxes affect the economy.

To set the stage for the analysis, imagine that a local government decides to hold an annual ice-cream celebration—with a parade, fireworks, and speeches by town bigwigs. To raise revenue for the event, the town will place a $0.50 tax on each sale of an ice-cream cone. When the plan is announced, our two lobbying groups swing into action. The American Association of Ice-Cream Eaters claims that consumers of ice cream are having trouble making ends meet, and it argues that **sellers** of ice cream should pay the tax. The National Organization of Ice-Cream Makers claims that its members are struggling to survive in a competitive market, and it argues that **buyers** of ice cream should pay the tax. The mayor, hoping for a compromise, suggests that buyers and sellers each pay half the tax.

To assess the proposals, ask a simple but subtle question: When the government levies a tax on a good, who actually bears the burden of the tax? The people buying the good? The people selling it? Or, if buyers and sellers share the tax burden, what determines how it is divided? Can the government make that decision, as the mayor suggests, or do market forces intervene? These issues involve **tax incidence**, the study of how the burden of a tax is distributed among the various people in the economy. The tools of supply and demand will reveal some surprising lessons about tax incidence.

tax incidence
the manner in which the burden of a tax is shared among participants in a market

6-2a How Taxes on Sellers Affect Market Outcomes

Let's begin with a tax levied on sellers. Suppose sellers of ice-cream cones are required to send the local government $0.50 for every cone they sell. How does this law affect the buyers and sellers of ice cream? To answer this question, follow the three steps in Chapter 4 for analyzing supply and demand: (1) Decide whether

In the News

Should the Minimum Wage Be $15 an Hour?

In 2021, President Biden proposed a minimum wage of $15 an hour, an idea that was controversial among both politicians and economists.

Raising the Minimum Wage Will Definitely Cost Jobs

By David Neumark

A recent Congressional Budget Office report estimated that 1.4 million jobs would be lost if a new $15 federal minimum wage is signed into law. Advocates were quick to dismiss the CBO's conclusion. "It is not a stretch to say that a new consensus has emerged among economists that minimum wage increases have raised wages without substantial job loss," said Heidi Shierholz of the Economic Policy Institute, which has also circulated a letter signed by economics Nobel laureates and others making the same claim.

As I show in a recent extensive survey of research on minimum wages and job loss in the U.S., this is simply not true. Most studies find that a minimum wage reduces employment of low-skilled workers, especially the lowest earners most directly affected by raising the minimum wage.

There are conflicting individual studies of the effects of minimum wages on employment. That there is disagreement shouldn't be surprising. Economics is a social science, not a natural one. Studies of minimum wages and job loss are not laboratory experiments. They can't be replicated and so can't be expected to yield exactly the same results.

What's surprising, though, is that summaries of the research literature make contradictory claims about what the overall body of evidence says. Distinguished economists like Angus Deaton and Peter Diamond signed the EPI letter asserting that the research shows little or no evidence of job loss, whereas others look at the research and conclude that it points to job loss. How can that be? Who is right?

Most economists have a strong stance on the minimum wage one way or the other. Perhaps this colors how they look at and interpret the evidence. Or perhaps there are so many studies of the employment effects of minimum wages that it is difficult to keep a "scorecard" of what the overall body of evidence says.

To provide an accurate reading of the research, Peter Shirley and I surveyed the authors of nearly all U.S. studies estimating the effects of minimum wages on employment published in the past 30 years. We asked them to report to us their best estimate of the employment effect, measured as the "elasticity," or the percent change in employment for each 1-percent change in the minimum wage. Most authors responded, and in the few cases in which they did not, we pulled this estimate from their study.

The results are stark. Across all studies, 79 percent report that minimum wages reduced employment. In 46 percent of studies the negative effect was statistically significant. In contrast, only 21 percent of studies found small positive effects of minimum wages on employment, and in only a minuscule percentage (4 percent) was the evidence statistically significant. A simplistic but useful calculation shows that the odds of nearly 80 percent of studies finding negative employment effects if the true effect is zero is less than one in a million.

Across all the studies, the average employment elasticity is about minus 0.15, which means, for example, that a 10-percent

the law affects the supply curve or the demand curve. (2) Decide which way the curve shifts. (3) Examine how the shift affects the equilibrium price and quantity.

Step One The immediate impact of the tax is on the sellers. Because the tax is not imposed on buyers, the quantity demanded at any price remains the same; thus, the demand curve does not change. By contrast, the tax on sellers makes the ice-cream business less profitable at any price, so it shifts the supply curve.

Step Two Because the tax on sellers raises the cost of producing and selling ice cream, it reduces the quantity supplied at every price. The supply curve shifts to the left (or, equivalently, upward).

Let's be precise about the size of the shift. For any market price of ice cream, the effective price to sellers—the amount they keep after paying the tax—is $0.50 lower. For example, if the market price of a cone happened to be $2.00, the effective price received by sellers would be $1.50. Whatever the market price, the effective price

increase in the minimum wage reduces employment of the low-skilled by 1.5 percent. Extrapolating this to a $15 minimum wage, this 107-percent increase in the states where the federal minimum wage of $7.25 now prevails would imply a 16-percent decline in low-skilled employment (broadly consistent with the recent CBO study). That sounds like a substantial job loss.

It's true that some workers would experience higher incomes, and that, on net, incomes of low-wage workers would probably rise. But that doesn't mean the minimum wage is the best way to help low-wage workers or low-income families, as research clearly demonstrates that a large share of income gains from a higher minimum wage flows to families with higher incomes. An alternative policy—the Earned Income Tax Credit—targets benefits to lower-income families far more effectively, is proven to reduce poverty, and creates rather than destroys jobs.

Our survey finds other important results. First, contrary to what is sometimes claimed, there is no tendency for the most recent research to provide less evidence of job loss. Second, the sharper a study's focus on workers directly affected by the minimum wage, the stronger the evidence of job loss. For example, the average employment elasticity for those with at most a high school education is minus 0.24, implying that a 10-percent increase in the minimum wage reduces their employment by 2.4 percent. The only studies that produce more mixed evidence are studies of low-wage industries, like retail or restaurants. Notably, in these studies the job loss among those most affected by the minimum wage may be masked by employers substituting from lower-skilled to higher-skilled workers.

True, some studies don't find evidence of job loss. But advocates for a higher minimum wage can claim support from the overall body of research evidence only if they discard most of that evidence. The consensus of economic research on the effects of minimum wages points clearly to job loss, and policy makers should consider this job loss in weighing the potential costs and benefits of a sharp increase in the minimum wage. ■

AP IMAGES/LYNNE SLADKY

Source: *The Wall Street Journal*, March 19, 2021.

Questions to Discuss

1. Suppose you are an economist in charge of designing policy to help low-wage workers. Would you prefer a minimum wage or an earned income tax credit? Why?
2. Suppose now you are a politician running for office. Would it be easier to campaign on a platform of a higher minimum wage or a more generous earned income tax credit? Why?

Mr. Neumark is a professor of economics at the University of California, Irvine.

for sellers is $0.50 less, and sellers supply a quantity of ice cream that is appropriate for that lower price. In other words, to induce sellers to supply any given quantity, the market price must now be $0.50 higher to compensate for the effect of the tax. As Figure 6 shows, the supply curve shifts **upward** from S_1 to S_2 by the exact size of the tax ($0.50).

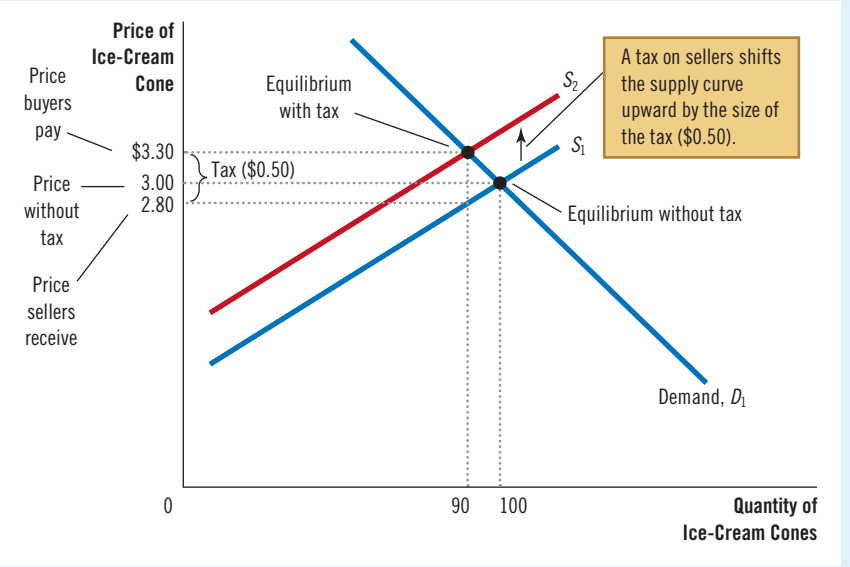
Step Three Having determined how the supply curve shifts, let's now compare the initial and the new equilibria. Figure 6 shows that the equilibrium price of ice cream rises from $3.00 to $3.30, and the equilibrium quantity falls from 100 to 90 cones. Because sellers now sell less and buyers buy less, the tax reduces the size of the ice-cream market.

Implications Now consider the question of tax incidence: Who pays the tax? Although sellers, not buyers, send the money to the government, buyers and sellers share the burden. Because the tax increases the market price from $3.00 to $3.30,

Figure 6

A Tax on Sellers

When a tax of $0.50 is levied on sellers, the supply curve shifts up by $0.50 from S_1 to S_2. The equilibrium quantity falls from 100 to 90 cones. The price that buyers pay rises from $3.00 to $3.30. The price that sellers receive (after paying the tax) falls from $3.00 to $2.80. Even though sellers are legally responsible for paying the tax, buyers and sellers share the burden.



buyers pay $0.30 more for ice-cream cones. Sellers get a higher price ($3.30), but after paying the tax, they only keep $2.80 ($3.30 − $0.50 = $2.80), which is $0.20 less than they did before. The tax makes both buyers and sellers worse off.

To sum up, this analysis yields two lessons:

- Taxes discourage market activity. When a good is taxed, the quantity sold is smaller in the new equilibrium.
- Buyers and sellers share the tax burden. In the new equilibrium, buyers pay more, and sellers receive less.

6-2b How Taxes on Buyers Affect Market Outcomes

Now consider a tax levied on buyers. Suppose that ice-cream lovers are required to send $0.50 to the local government for each cone they buy. What are the law's effects? Let's again turn to our three steps.

Step One The immediate impact is on the demand for ice cream. The supply curve doesn't change because, for any price, sellers have the same incentive to provide ice cream to the market. But buyers now have to pay a tax to the government (on top of the price to the sellers), so the tax shifts the demand curve for ice cream.

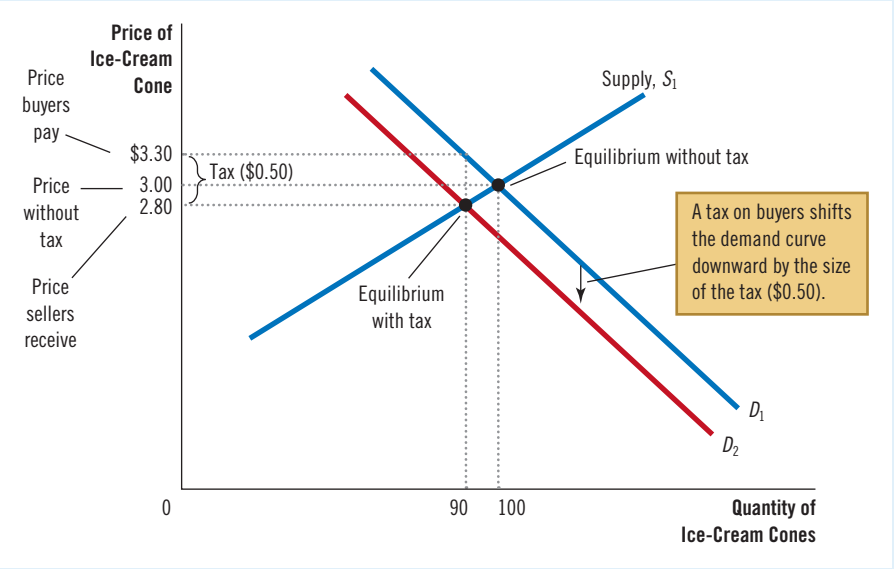
Step Two Next, determine the direction of the shift. Because the tax makes buying ice cream less attractive, buyers demand a smaller quantity of ice cream at every price. The demand curve shifts to the left (or, equivalently, downward), as shown in Figure 7.

Once again, let's be precise about the size of the shift. Because of the $0.50 tax on buyers, their effective price is now $0.50 higher than whatever the market price happens to be. For example, if the market price of a cone were $2.00, buyers would face an effective price for buyers of $2.50. Because buyers look at their total cost,

Figure 7

A Tax on Buyers

When a tax of $0.50 is imposed on buyers, the demand curve shifts down by $0.50 from D_1 to D_2. The equilibrium quantity falls from 100 to 90 cones. The price that sellers receive falls from $3.00 to $2.80. The price that buyers pay (including the tax) rises from $3.00 to $3.30. Even though buyers are legally responsible for paying the tax, buyers and sellers share the burden.



including the tax, they demand a quantity of ice cream as if the market price were $0.50 higher than it actually is. In other words, to induce buyers to demand any given quantity, the market price must now be $0.50 lower to make up for the effect of the tax. The tax shifts the demand curve **downward** from D_1 to D_2 by the exact size of the tax ($0.50).

Step Three Let's now evaluate the effect of the tax by comparing the initial equilibrium with the new one. In Figure 7, the equilibrium price of ice cream falls from $3.00 to $2.80, and the equilibrium quantity drops from 100 to 90 cones. Once again, the tax reduces the size of the ice-cream market. And once again, buyers and sellers share the burden. Sellers get a lower price for their product; buyers pay a lower market price to sellers than they previously did, but the effective price (including the tax) rises from $3.00 to $3.30.

Implications If you compare Figures 6 and 7, you will notice a surprising conclusion: **Taxes on sellers and taxes on buyers are equivalent.** In both cases, the tax inserts a wedge between the price that buyers pay and the price that sellers receive. Regardless of whether the tax is levied on buyers or sellers, the wedge remains the same. In either case, it shifts the relative position of the supply and demand curves. In the new equilibrium, buyers and sellers share the tax burden. The only difference between a tax on sellers and a tax on buyers is who sends the money to the government.

To better understand the equivalence of these two taxes, imagine the government collects the $0.50 ice-cream tax in a bowl on the counter of each ice-cream store. When the tax is imposed on sellers, the sellers are required to place $0.50 in the bowl each time they sell a cone. When the tax is imposed on buyers, the buyers must place $0.50 in the bowl whenever they buy a cone. Whether the $0.50 goes directly from the buyer's pocket into the bowl, or indirectly from the buyer's pocket into the seller's hand and then into the bowl, does not matter. Once the market reaches its new equilibrium, buyers and sellers share the burden, regardless of how the tax is levied.

Case Study

Can Congress Distribute the Burden of a Payroll Tax?

If you have ever received a paycheck, you probably noticed that taxes were deducted from the amount you earned. One of these taxes is called FICA, an acronym for the Federal Insurance Contributions Act. The federal government uses the revenue from the FICA tax to pay for Social Security and Medicare, the income support and healthcare programs for older Americans. FICA is a **payroll tax**, which is a tax on the wages that firms pay their workers. In 2021, the total FICA tax for the typical worker was 15.3 percent of earnings.

Who bears the burden of this payroll tax—firms or workers? When Congress passed this legislation, it tried to divide the tax burden. According to the law, half of the tax is paid by firms and half by workers. That is, half of the tax is paid out of firms' revenues, and half is deducted from workers' paychecks. The amount that shows up as a deduction on your pay stub is the worker contribution. (Self-employed people generally pay the whole tax themselves.)

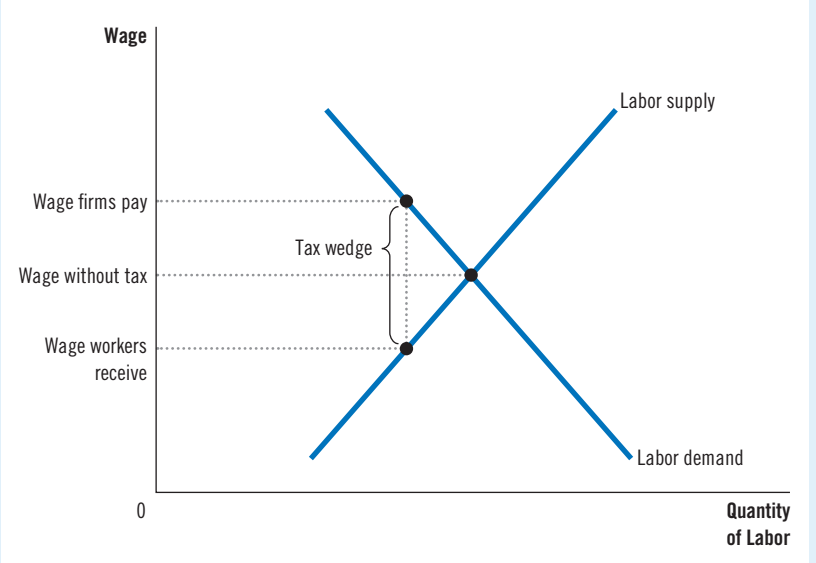
Our analysis of tax incidence, however, shows that lawmakers cannot dictate the distribution of a tax burden so easily. A payroll tax is analyzed much the same as a tax on a good like ice cream. In this case, the good is labor, and the price is the wage. Again, the tax inserts a wedge—here, between the wage that firms pay and the wage that workers receive (AKA "take-home pay"). Figure 8 shows the outcome. When a payroll tax is enacted, the wage received by workers falls, and the wage paid by firms rises. In the end, workers and firms share the burden, much as the legislation requires. Yet this economic division has nothing to do with the legislated one: The division of the tax burden in Figure 8 is not necessarily 50–50, and the same outcome would prevail if the law imposed the entire tax on either workers or firms.

This example highlights an often overlooked lesson. Lawmakers can decide whether a tax comes from the buyer's pocket or from the seller's, but they cannot legislate the true burden of a tax. Rather, tax incidence depends on the forces of supply and demand. ●

Figure 8

A Payroll Tax

A payroll tax places a wedge between what firms pay and what workers receive. Comparing wages with and without the tax makes it clear that workers and firms share the tax burden. This division does not depend on whether the government imposes the tax entirely on workers, imposes it entirely on firms, or divides it equally between the two groups.



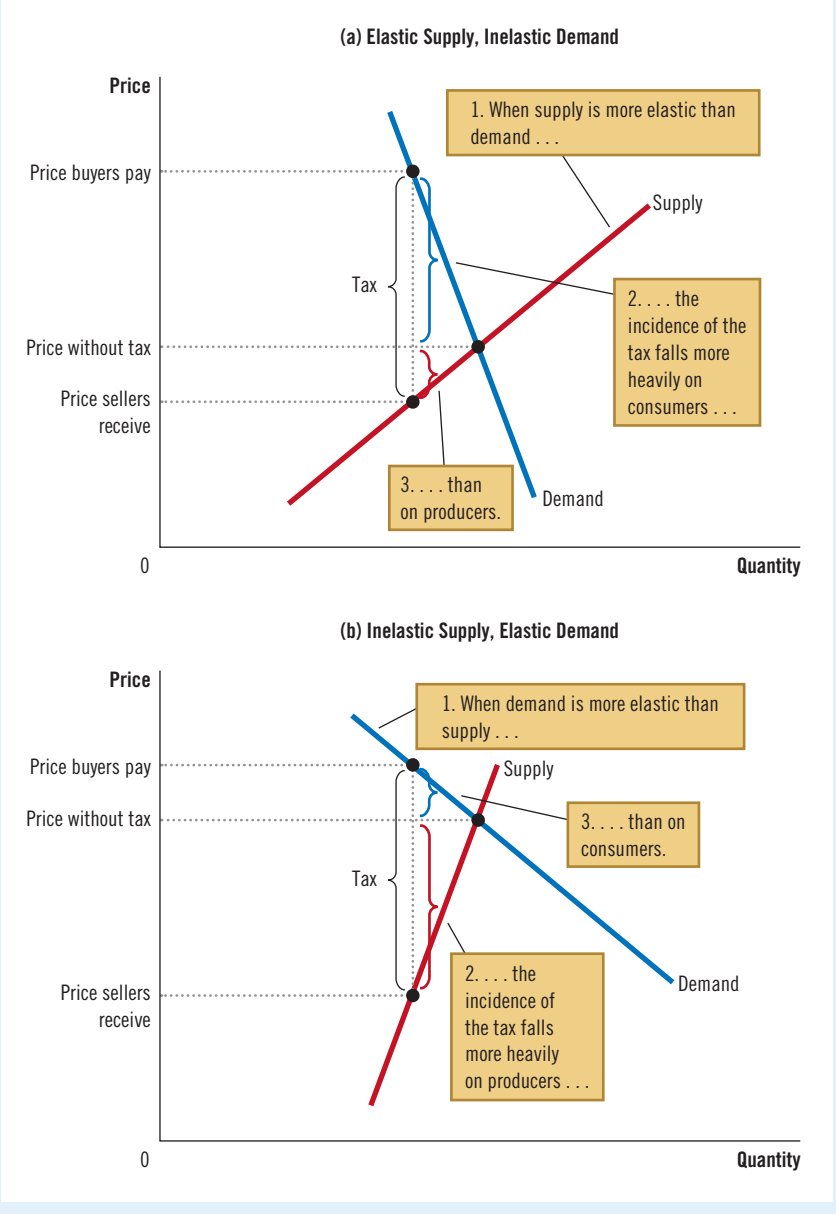
6-2c Elasticity and Tax Incidence

How exactly is the tax burden divided between buyers and sellers? Only rarely will it be shared equally. To see this, consider the impact of taxation on the two markets in Figure 9. In both cases, the figure shows the initial supply and demand curves and a tax that drives a wedge between what buyers pay and sellers receive. (Not drawn in either panel is the new supply or demand curve. Which curve shifts depends on whether the tax is levied on buyers or sellers, a fact that is irrelevant for determining the tax's incidence.) The difference between the two panels is the relative elasticity of supply and demand.

Figure 9

How a Tax Burden Is Divided

In panel (a), the supply curve is elastic, and the demand curve is inelastic. In this case, the price received by sellers falls only slightly, while the price paid by buyers rises substantially. This means that buyers bear most of the tax burden. In panel (b), the situation is reversed: The supply curve is inelastic, and the demand curve is elastic. In this case, the price received by sellers falls substantially, while the price paid by buyers rises only slightly. Here, sellers bear most of the burden.



Panel (a) of Figure 9 shows a tax in a market with very elastic supply and relatively inelastic demand. That means that sellers are very responsive to changes in the price (so the supply curve is relatively flat), while buyers are not very responsive (so the demand curve is relatively steep). When a tax is imposed on a market like this one, the price received by sellers does not fall much, so sellers bear only a small burden. But the price paid by buyers rises substantially, indicating that they bear most of the tax burden.

Panel (b) of Figure 9 shows a tax in a market with fairly inelastic supply and very elastic demand. In this case, sellers are not very responsive to changes in the price (so the supply curve is steeper), while buyers are very responsive (so the demand curve is flatter). When a tax is imposed, the price paid by buyers doesn't rise much, but the price received by sellers falls substantially. Thus, sellers bear most of the tax burden.

Together, the two panels show a general lesson: **A tax burden falls more heavily on the side of the market that is less elastic.** Why is this true? In essence, elasticity measures the willingness of buyers or sellers to leave the market when conditions worsen. A small elasticity of demand means that buyers do not have good alternatives to consuming this particular good. A small elasticity of supply means that sellers do not have good alternatives to producing this particular good. When the good is taxed, the side of the market with fewer good alternatives is less willing to leave the market and bears more of the burden of the tax.

This logic applies to the payroll tax discussed in the previous case study. Because economists have generally found that labor supply is less elastic than labor demand, workers, rather than firms, bear most of the burden of the payroll tax. In other words, the distribution of the tax burden is far from the 50–50 split that lawmakers intended.

Case Study

Who Pays the Luxury Tax?

In 1990, Congress adopted a luxury tax on items such as yachts, private airplanes, furs, jewelry, and high-end cars. The goal was to raise revenue from those who could most easily afford to pay. Because only the rich could afford such extravagances, taxing luxuries seemed like a logical way of doing that.

Yet, when the forces of supply and demand took over, the outcome was not what Congress intended. Consider the market for yachts, for example. The demand in this market is quite elastic. A billionaire can easily not buy a yacht; the money can be used to buy an island, take a more luxurious vacation, or leave a larger bequest to heirs. By contrast, the supply of yachts is relatively inelastic, at least in the short run. The shipyards that produce yachts are not easily converted to alternative uses, and the workers who work there are not eager to change jobs in response to changing market conditions.

VALENTINRUSSANOV/E+/GETTY IMAGES

"If this boat were any more expensive, we'd be playing golf."

Our analysis makes a clear prediction. With elastic demand and inelastic supply, the tax burden falls largely on the suppliers. In this case, a tax on yachts places the burden largely on the businesses and workers who build yachts because they end up getting significantly less for their vessels. The workers are not wealthy, even if some of the business owners are. In the end, the burden of a luxury tax can fall more on the middle-class workers than on the rich customers.

The mistaken assumptions about the incidence of the luxury tax quickly became apparent after it went into effect. Suppliers of luxuries made their elected representatives well aware of their problems, and Congress repealed most of the luxury tax in 1993. ●

QuickQuiz

5. A $1 per unit tax levied on consumers of a good is equivalent to
   a. a $1 per unit tax levied on producers of the good.
   b. a $1 per unit subsidy paid to producers of the good.
   c. a price floor that raises the good's price by $1 per unit.
   d. a price ceiling that raises the good's price by $1 per unit.
6. When a good is taxed, the burden falls mainly on consumers if
   a. the tax is levied on consumers.
   b. the tax is levied on producers.
   c. supply is inelastic and demand is elastic.
   d. supply is elastic and demand is inelastic.
7. Which of the following increases the quantity supplied, decreases the quantity demanded, and increases the price that consumers pay?
   a. the passage of a tax on a good
   b. the repeal of a tax on a good
   c. the imposition of a binding price floor
   d. the removal of a binding price floor
8. Which of the following increases the quantity supplied, increases the quantity demanded, and decreases the price that consumers pay?
   a. the passage of a tax on a good
   b. the repeal of a tax on a good
   c. the imposition of a binding price floor
   d. the removal of a binding price floor

Answers are at the end of the chapter.

6-3 Conclusion

The economy is governed by two kinds of laws: the laws of supply and demand and the laws enacted by governments. In this chapter, we have begun to see how these laws interact. Price controls and taxes are common in various markets, and their effects are frequently debated. Even a little bit of economic knowledge can go a long way toward understanding and evaluating these policies.

Subsequent chapters analyze government policies in greater detail. We will examine the effects of taxation more fully and consider a broader range of policies. Yet the basic lessons will not change: When analyzing government policies, supply and demand are the first and most useful tools of analysis.

Chapter in a Nutshell

- A price ceiling is a legal maximum on the price of a good or service. Rent control is an example. If the ceiling is below the equilibrium price, then it is binding, and the quantity demanded exceeds the quantity supplied. Because of the resulting shortage, sellers must somehow ration the good or service among buyers.
- A price floor is a legal minimum on the price of a good or service. The minimum wage is an example. If the floor is above the equilibrium price, then it is binding, and the quantity supplied exceeds the quantity demanded. Because of the resulting surplus, buyers' demands for the good or service must somehow be rationed among sellers.
- When the government levies a tax on a good, the equilibrium quantity of the good falls. That is, a tax on a market shrinks the market's size.
- A tax on a good places a wedge between the price paid by buyers and the price received by sellers. When the market moves to the new equilibrium, buyers pay more for the good, and sellers receive less for it. In this sense, buyers and sellers share the tax burden. The incidence of a tax (that is, the division of the tax burden) does not depend on whether the tax is levied on buyers or sellers.
- The incidence of a tax depends on the price elasticities of supply and demand. Most of the burden falls on the side of the market that is less elastic because it cannot respond as easily to the tax by changing the quantity bought or sold.

Key Concepts

price ceiling, p. 112 price floor, p. 112 tax incidence, p. 121

Questions for Review

1. Give an example of a price ceiling and an example of a price floor.
2. Which causes a shortage of a good—a price ceiling or a price floor? Justify your answer with a graph.
3. What mechanisms allocate resources when the price of a good is not allowed to bring supply and demand into equilibrium?
4. Explain why economists frequently oppose price controls.
5. Suppose the government removes a tax on buyers of a good and levies a tax of the same size on sellers. How does this policy change affect the price that buyers pay sellers for this good, the amount buyers are out of pocket (including any tax payments they make), the amount sellers receive (net of any tax payments they make), and the quantity of the good sold?
6. How does a tax on a good affect the price paid by buyers, the price received by sellers, and the quantity sold?
7. What determines how the burden of a tax is divided between buyers and sellers? Why?

Problems and Applications

1. Lovers of comedy persuade Congress to impose a price ceiling of $50 per ticket for live comedy performances. As a result of this policy, do more or fewer people attend comedy performances? Explain.
2. The government has decided that the free-market price of cheese is too low.
   a. Suppose the government imposes a binding price floor in the cheese market. Draw a supply-and-demand diagram to show the effect of this policy on the price and quantity of cheese sold. Is there a shortage or surplus of cheese?
   b. Producers of cheese complain that the price floor has reduced their total revenue. Is this possible? Explain.
   c. In response to cheese producers' complaints, the government agrees to purchase all the surplus cheese at the price floor. Compared to the basic price floor, who benefits from this new policy? Who loses?
3. A recent study found that the demand-and-supply schedules for Frisbees are as follows:

Price per Frisbee	Quantity Demanded	Quantity Supplied
$11	1 million Frisbees	15 million Frisbees
10	2	12
9	4	9
8	6	6
7	8	3
6	10	1

   a. What are the equilibrium price and quantity of Frisbees?
   b. Frisbee manufacturers persuade the government that Frisbee production improves scientists' understanding of aerodynamics and is thus important for national security. A concerned Congress votes to impose a price floor $2 above the equilibrium price. What is the new market price? How many Frisbees are sold?
   c. Irate college students march on Washington and demand a reduction in the price of Frisbees. An even more concerned Congress votes to repeal the price floor and impose a price ceiling $1 below the former price floor. What is the new market price? How many Frisbees are sold?
4. Suppose the federal government requires beer drinkers to pay a $2 tax on each case of beer purchased. (In fact, both the federal and state governments impose beer taxes of some sort.)
   a. Draw a supply-and-demand diagram of the market for beer without the tax. Show the price paid by consumers, the price received by producers, and the quantity of beer sold. What is the difference between the price paid by consumers and the price received by producers?
   b. Now draw a supply-and-demand diagram for the beer market with the tax. Show the price paid by consumers, the price received by producers, and the quantity of beer sold. What is the difference between the price paid by consumers and the

price received by producers? Has the quantity of beer sold increased or decreased?

5. A senator wants to raise tax revenue and make workers better off. A staff member proposes raising the payroll tax paid by firms and using part of the extra revenue to reduce the payroll tax paid by workers. Would this accomplish the senator's goal? Explain.
6. If the government places a $500 tax on luxury cars, will the price paid by consumers rise by more than $500, less than $500, or exactly $500? Explain.
7. Congress and the president decide that the United States should reduce air pollution by reducing its use of gasoline. They impose a $0.50 tax on each gallon of gasoline sold.
   a. Should they impose this tax on producers or consumers? Explain carefully using a supply-and-demand diagram.
   b. If the demand for gasoline were more elastic, would this tax be more effective or less effective in reducing the quantity of gasoline consumed? Explain with both words and a diagram.
   c. Are consumers of gasoline helped or hurt by this tax? Why?
   d. Are workers in the oil industry helped or hurt by this tax? Why?
8. A case study in this chapter discusses the federal minimum-wage law.
   a. Suppose the minimum wage is above the equilibrium wage in the market for unskilled labor. Using a supply-and-demand diagram of the market for unskilled labor, show the market wage, the number of workers who are employed, and the number of workers who are unemployed. Also, show the total wage payments to unskilled workers.
   b. Now suppose the Secretary of Labor proposes an increase in the minimum wage. What effect would this increase have on employment? Does the change in employment depend on the elasticity of demand, the elasticity of supply, both elasticities, or neither?
   c. What effect would this increase in the minimum wage have on unemployment? Does the change in unemployment depend on the elasticity of demand, the elasticity of supply, both elasticities, or neither?
   d. If the demand for unskilled labor were inelastic, would the proposed increase in the minimum wage raise or lower total wage payments to unskilled workers? Would your answer change if the demand for unskilled labor were elastic?
9. At Fenway Park, home of the Boston Red Sox, seating is limited to about 38,000. Hence, the number of tickets issued is fixed at that figure. Seeing a golden opportunity to raise revenue, the City of Boston levies a per ticket tax of $5 to be paid by the ticket buyer. Boston sports fans, a famously civic-minded lot, dutifully send in the $5 per ticket. Draw a well-labeled graph showing the impact of the tax. On whom does the tax burden fall—the team's owners, the fans, or both? Why?
10. A market is described by the following supply and demand curves:

$$Q^S = 2P$$
$$Q^D = 300 - P.$$

   a. Solve for the equilibrium price and quantity.
   b. If the government imposes a price ceiling of $90, does a shortage or surplus (or neither) develop? What are the price, quantity supplied, quantity demanded, and size of the shortage or surplus?
   c. If the government imposes a price floor of $90, does a shortage or surplus (or neither) develop? What are the price, quantity supplied, quantity demanded, and size of the shortage or surplus?
   d. Instead of a price control, the government levies a $30 tax on producers. As a result, the new supply curve is:

$$Q^S = 2(P - 30).$$

   Does a shortage or surplus (or neither) develop? What are the price, quantity supplied, quantity demanded, and size of the shortage or surplus?

QuickQuiz Answers

1. **d** 2. **d** 3. **a** 4. **c** 5. **a** 6. **d** 7. **c** 8. **b**

Chapter

7

Consumers, Producers, and the Efficiency of Markets

When consumers go to their local farmers' market, they may be delighted to find juicy, red tomatoes but appalled by their high price. At the same time, when farmers bring to market the tomatoes they have raised, they probably wish the price were even higher. These views are not surprising: Other things being equal, buyers usually want to pay less, and sellers usually want to be paid more. But is there a "right price" for tomatoes from the standpoint of society as a whole?

Previous chapters showed how, in competitive markets, the forces of supply and demand determine the prices of goods and services and the quantities sold. So far, however, we have described how markets allocate scarce resources without considering whether these market allocations are desirable. We know that the price of tomatoes adjusts to ensure that the quantity of tomatoes supplied equals the quantity of tomatoes demanded. But at this equilibrium, is the quantity of tomatoes produced and consumed too small, too large, or just right?

GAUDILAB/SHUTTERSTOCK.COM, LECHATNOIR/E+/GETTY IMAGES

welfare economics
the study of how the allocation of resources affects economic well-being

This chapter takes up the topic of **welfare economics**, the study of how the allocation of resources affects economic well-being. We begin by examining the benefits that buyers and sellers receive from engaging in market transactions. We then examine how society can make these benefits as large as possible. This analysis leads to a profound conclusion: The equilibrium of supply and demand in competitive markets maximizes the total benefits received by all buyers and sellers combined.

As you may recall from Chapter 1, one of the **Ten Principles of Economics** is that markets are usually a good way to organize economic activity. The study of welfare economics explains this principle more fully. It also answers our question about the right price of tomatoes: The price that balances the supply and demand for tomatoes is, in a particular sense, the best one because it maximizes the total welfare of consumers and producers. No consumer or producer of tomatoes needs to aim for this goal, but their joint action, directed by market prices, moves them toward a welfare-maximizing outcome, as if led by an invisible hand.

7-1 Consumer Surplus

We begin our study of welfare economics by looking at the benefits buyers receive from participating in a market.

7-1a Willingness to Pay

Imagine that you inherit from your great aunt a rare, mint-condition recording of Elvis Presley's first album. Because you don't care for Elvis's music, you decide to sell the album by auctioning it off.

Four Elvis fans come to your auction: Whitney, Ella, Mariah, and Karen. They all want the album, but they have set limits on the amount they will pay for it. Table 1 shows the maximum price that each of the four possible buyers would pay. A buyer's maximum is called her **willingness to pay**, and it measures how much she values the good. Each buyer would be eager to buy the album at a price less than her willingness to pay, and each would refuse to buy the album at a price greater than her willingness to pay. At a price equal to her willingness to pay, the buyer would be indifferent about buying the good: If the price is exactly the same as the value she places on the album, she would be equally happy buying it or keeping her money.

willingness to pay
the maximum amount that a buyer will pay for a good

To sell your album, you begin the bidding process at a low price, say, $100. Because all four buyers are willing to pay much more, the price quickly rises. The bidding stops when Whitney bids $800 (or slightly more). At this point, Ella, Mariah, and Karen have dropped out of the bidding because they are unwilling to offer more than $800. Whitney pays $800 and gets the album. Note that the album goes to the buyer who values it most.

Table 1

The Willingness to Pay of Four Possible Buyers

Buyer	Willingness to Pay
Whitney	$1,000
Ella	800
Mariah	700
Karen	500

What benefit does Whitney derive from buying the Elvis Presley album? In a sense, she has found a bargain: She is willing to pay $1,000 for the album but hands over only $800. Economists say that Whitney receives **consumer surplus** of $200. **Consumer surplus** is the amount a buyer is willing to pay for a good minus the amount the buyer actually pays for it.

consumer surplus the amount a buyer is willing to pay for a good minus the amount the buyer actually pays for it

Consumer surplus measures the benefit buyers receive from participating in a market. In this example, Whitney gets a $200 benefit because she pays only $800 for something she values at $1,000. Ella, Mariah, and Karen get no consumer surplus because they leave the auction without the album and without paying anything.

Now consider a somewhat different example. Suppose that you have two identical Elvis Presley albums to sell. Again, you auction them off to the four possible buyers. To keep things simple, assume that both copies are to be sold for the same price and that no one wants more than one album. Therefore, the price rises until two buyers are left.

In this case, the bidding stops when Whitney and Ella each bid $700 (or slightly higher). At this price, Whitney and Ella are happy to buy an album, and Mariah and Karen are not willing to bid any higher. Whitney and Ella each receive consumer surplus equal to her willingness to pay minus the price. Whitney's consumer surplus is $300, and Ella's is $100. Whitney's consumer surplus is higher than in the previous example because she pays less for the same album. The total consumer surplus in the market is $400.

7-1b Using the Demand Curve to Measure Consumer Surplus

Consumer surplus is closely related to the demand curve for a product. To see how, consider the demand curve for this rare Elvis Presley album.

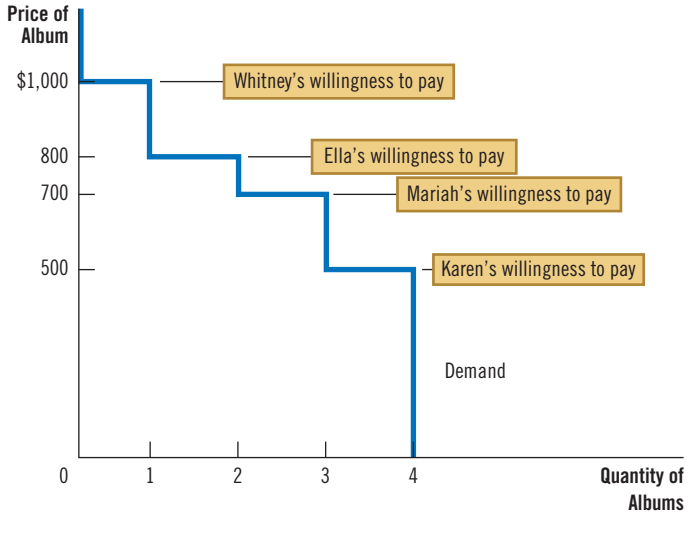
Begin by using the willingness to pay of the four possible buyers to find the album's market demand schedule. The table in Figure 1 shows the demand schedule that corresponds to the valuations in Table 1. If the price is above $1,000, the

Figure 1

The Demand Schedule and the Demand Curve

The table shows the demand schedule for the buyers (listed in Table 1) of the mint-condition copy of Elvis Presley's first album. The graph shows the corresponding demand curve. The height of the demand curve reflects the buyers' willingness to pay.

Price	Buyers	Quantity Demanded
More than $1,000	None	0
$800 to $1,000	Whitney	1
$700 to $800	Whitney, Ella	2
$500 to $700	Whitney, Ella, Mariah	3
$500 or less	Whitney, Ella, Mariah, Karen	4



quantity demanded is 0 because no buyer is willing to pay that much. If the price is between $800 and $1,000, the quantity demanded is 1: Only Whitney is willing to pay such a high price. If the price is between $700 and $800, the quantity demanded is 2 because both Whitney and Ella are willing to pay the price. You can continue this analysis for other prices as well. In this way, the demand schedule is derived from the willingness to pay of the four possible buyers.

The graph in Figure 1 shows the demand curve that corresponds to this demand schedule. Note the relationship between the height of the curve and the buyers' willingness to pay. At any quantity, the price given by the demand curve shows the willingness to pay of the **marginal buyer**, the buyer who would leave the market first if the price were any higher. At a quantity of 4 albums, for instance, the demand curve has a height of $500, the price that Karen (the marginal buyer) is willing to pay. At a quantity of 3 albums, the demand curve has a height of $700, the price that Mariah (now the marginal buyer) is willing to pay.

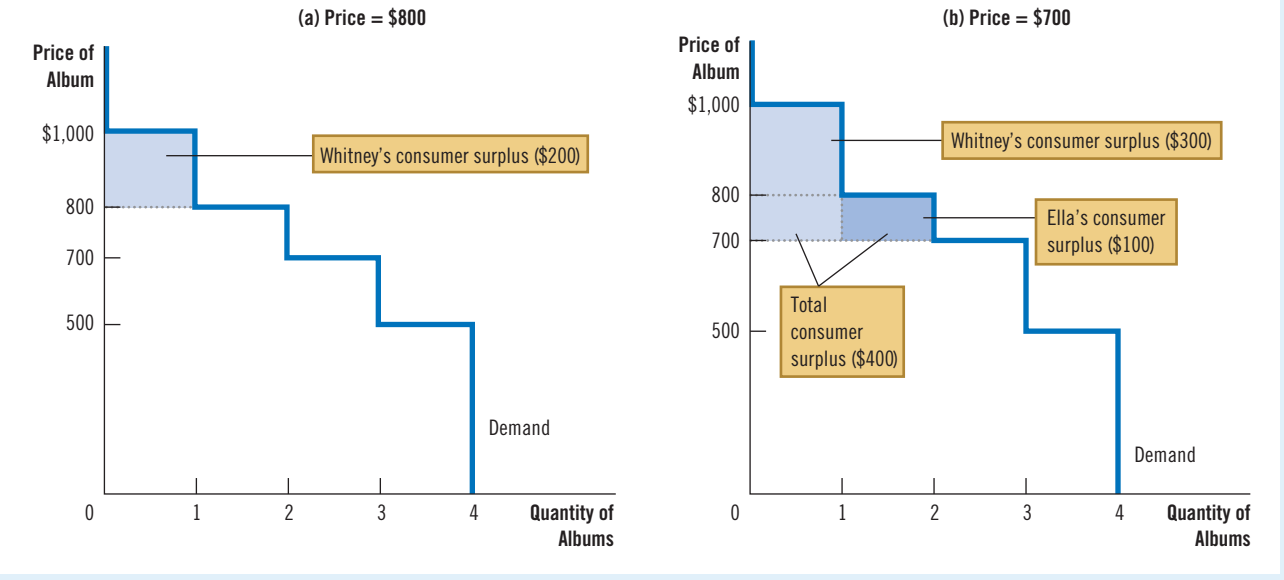
Because the demand curve reflects buyers' willingness to pay, it can also be used to measure consumer surplus. Figure 2 does this for our two examples. In panel (a), the price is $800 (or slightly above), and the quantity demanded is 1. Note that the area above the price and below the demand curve equals $200. This amount is the consumer surplus calculated earlier when only 1 album is sold.

Panel (b) of Figure 2 shows consumer surplus when the price is $700 (or slightly above). In this case, the area above the price and below the demand curve equals the total area of the two rectangles: Whitney's consumer surplus at this price is $300, and Ella's is $100. This area equals a total of $400. Once again, this amount is the consumer surplus calculated earlier.

Figure 2

Measuring Consumer Surplus with the Demand Curve

In panel (a), the price of the good is $800, and consumer surplus is $200. In panel (b), the price is $700, and consumer surplus is $400.



The lesson from this example holds for all demand curves: **The area below the demand curve and above the price measures the consumer surplus in a market.** This is true because the height of the demand curve represents the value buyers place on the good, measured by their willingness to pay for it. The difference between this willingness to pay and the market price is each buyer's consumer surplus. The area between the demand curve and the price line is the sum of the consumer surplus of all buyers in the market for a good or service.

7-1c How a Lower Price Raises Consumer Surplus

Because buyers want to pay less for the goods they buy, a lower price makes buyers of a good better off. But how much does buyers' well-being increase in response to a lower price? The concept of consumer surplus offers a precise answer.

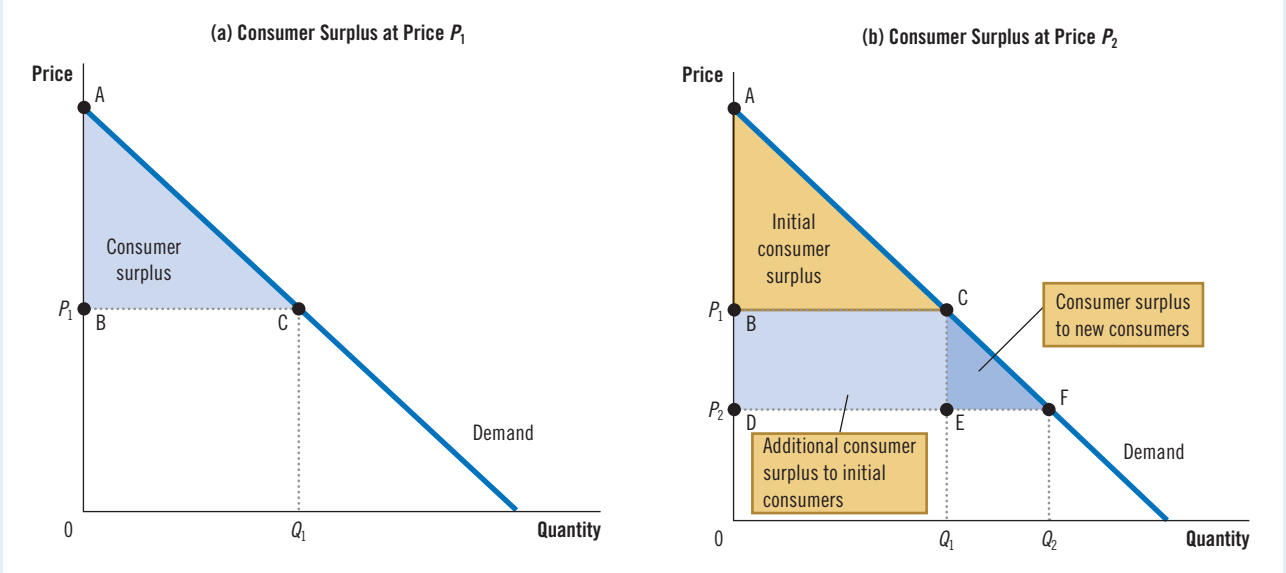
Figure 3 shows a typical demand curve. You may notice that it gradually slopes downward instead of taking discrete steps as in the previous two figures. In a market with many buyers, the resulting steps from each buyer dropping out are so small that they form a smooth demand curve. Although this curve has a different shape, the ideas we have just developed still apply: Consumer surplus is the area above the price line and below the demand curve. In panel (a), the price is P_1, and consumer surplus is the area of triangle ABC.

Now suppose that the price falls from P_1 to P_2, as shown in panel (b). Consumer surplus equals area ADF. The increase in consumer surplus from the price cut is the area BCFD.

Figure 3

How Price Affects Consumer Surplus

In panel (a), the price is P_1, the quantity demanded is Q_1, and consumer surplus equals the area of the triangle ABC. When the price falls from P_1 to P_2, as in panel (b), the quantity demanded rises from Q_1 to Q_2, and consumer surplus rises to the area of the triangle ADF. The increase in consumer surplus (area BCFD) occurs in part because existing consumers pay less (area BCED) and in part because new consumers enter the market at the lower price (area CEF).



This increase in consumer surplus has two parts. First, buyers who were already purchasing Q_1 of the good at the higher price P_1 are better off because now they pay less. The increase in consumer surplus of existing buyers is the reduction in the amount they pay. It equals the area of the rectangle BCED. Second, some new buyers enter the market because they are willing to buy the good at the lower price, which increases the quantity demanded from Q_1 to Q_2. The consumer surplus for the newcomers is the area of the triangle CEF.

7-1d What Does Consumer Surplus Measure?

The concept of consumer surplus can be helpful in making judgments about the desirability of market outcomes. Having seen what consumer surplus is, let's consider whether it is a good measure of economic well-being.

Imagine that you are a policymaker designing an economic system. Would you care about consumer surplus? Because it is the amount that buyers are willing to pay for a good minus the amount they actually pay for it, consumer surplus measures the benefit that buyers derive from a market **as the buyers themselves perceive it**. Consumer surplus is a good measure of economic well-being if policymakers want to satisfy buyers' preferences.

In some circumstances, policymakers might choose to disregard consumer surplus because they do not respect the preferences that drive buyer behavior. For example, addicts are willing to pay a high price for their drug of choice. Yet policymakers would be unlikely to say that addicts get a large benefit from being able to buy drugs at a low price (even though addicts might say they do). From the standpoint of society, willingness to pay in this instance is not a good measure of the buyers' benefit, and consumer surplus is not a good measure of economic well-being, because addicts are not looking after their own best interests.

In most markets, however, consumer surplus reflects economic well-being. Economists generally assume that buyers are rational when they make decisions. Rational people do the best they can to achieve their objectives, given their opportunities. Economists also normally assume that people's preferences should be respected. In this case, consumers are the best judges of how much benefit they get from the goods they buy.

QuickQuiz

1. Alexis, Bruno, and Camila each want an ice-cream cone. Alexis is willing to pay $12, Bruno is willing to pay $8, and Camila is willing to pay $4. The market price is $6. Their consumer surplus is
   a. $6.
   b. $8.
   c. $14.
   d. $18.
2. If the price of an ice-cream cone falls to $3, the consumer surplus of Alexis, Bruno, and Camila increases by
   a. $6.
   b. $7.
   c. $8.
   d. $9.
3. The demand curve for cookies slopes downward. When the price is $3 per cookie, the quantity demanded is 100. If the price falls to $2, what happens to consumer surplus?
   a. It falls by less than $100.
   b. It falls by more than $100.
   c. It rises by less than $100.
   d. It rises by more than $100.

Answers are at the end of the chapter.

7-2 Producer Surplus

Let's now turn to the other side of the market and consider the benefits sellers receive from participating in a market. The analysis of sellers' welfare is parallel to the analysis of buyers' welfare.

7-2a Cost and the Willingness to Sell

Imagine that you are a homeowner and want to get your house painted. You turn to four sellers of painting services: Vincent, Claude, Pablo, and Andy. Each painter is willing to do the work for you if the price is right. You set up an auction and take bids from the four painters.

Each painter is willing to take the job if the price exceeds his cost of doing the work. Here the term **cost** should be interpreted as the painter's opportunity cost: It includes his out-of-pocket expenses (for paint, brushes, and so on) and, most importantly, the value that he places on his time. Table 2 shows each painter's cost. Because a painter's cost is the lowest price he would accept for his work, it measures his willingness to sell his services. Each painter would be eager to sell his services at a price greater than his cost and would refuse to sell his services at a price less than his cost. At a price exactly equal to his cost, he would be indifferent about selling his services: He would be equally satisfied getting the job or using his time and energy elsewhere.

cost
the value of everything a seller must give up to produce a good

When you take bids, the price might start high, but it quickly falls as the painters compete for the job. Once Andy has bid $2,400 (or slightly less), he is the sole remaining bidder. Andy wants to do the job for this price because his cost is only $2,000. Vincent, Claude, and Pablo are unwilling to do it for less than $2,400. Note that the job goes to the painter who can do the work at the lowest cost.

What benefit does Andy derive from getting the job? Because he is willing to do the work for $2,000 but is paid $2,400, economists say that he receives **producer surplus** of $400. **Producer surplus** is the amount a seller is paid minus his cost of production. It measures how much a seller benefits from participating in a market.

producer surplus
the amount a seller is paid for a good minus the seller's cost of providing it

Now consider a different example. Suppose two houses need painting. Again, you auction off the jobs to the four painters. To keep things simple, assume that no painter can paint both houses and that you will pay the same amount to paint each house. Therefore, the price falls until two painters are left.

In this case, the auction stops when Andy and Pablo each bid $3,200 (or slightly less). They are willing to do the work at this price, while Vincent and Claude won't go lower. At $3,200, Andy's producer surplus is $1,200, and Pablo's is $800. The total producer surplus in the market is $2,000.

Table 2

The Costs of Four Possible Sellers

Seller	Cost
Vincent	$3,600
Claude	3,200
Pablo	2,400
Andy	2,000

7-2b Using the Supply Curve to Measure Producer Surplus

Just as consumer surplus is closely related to the demand curve, producer surplus is closely related to the supply curve. To see how, let's derive the supply curve for painting services.

Begin by using the costs of the four painters to find the supply schedule. The table in Figure 4 shows the schedule that corresponds to the costs in Table 2. If the price is below $2,000, no painter will do the job, so the quantity supplied is zero. If the price is between $2,000 and $2,400, only Andy will do it, so the quantity supplied is 1. If the price is between $2,400 and $3,200, Andy and Pablo will do the work, so the quantity supplied is 2, and so on. Thus, the supply schedule is derived from the costs of the four painters.

The graph in Figure 4 shows the supply curve that corresponds to this supply schedule. Note that the height of the supply curve is related to the sellers' costs. At any quantity, the price given by the supply curve shows the cost of the **marginal seller**, the seller who would leave the market first if the price were lower. At a quantity of 4 houses, for instance, the supply curve has a height of $3,600, the cost that Vincent (the marginal seller) incurs to provide his painting services. At a quantity of 3 houses, the supply curve has a height of $3,200, the cost that Claude (who is now the marginal seller) incurs.

Because the supply curve reflects sellers' costs, it can be used to measure producer surplus. Figure 5 uses the supply curve to compute producer surplus in the two examples. In panel (a), the price is $2,400 (or slightly less), and the quantity supplied is 1. The area below the price and above the supply curve equals $400. This is the producer surplus calculated earlier for Andy.

Panel (b) of Figure 5 shows producer surplus when the price is $3,200 (or slightly less). In this case, the area below the price and above the supply curve equals the total area of the two rectangles. This area equals $2,000, the producer surplus computed earlier for Pablo and Andy when two houses needed painting.

Figure 4

The Supply Schedule and the Supply Curve

The table shows the supply schedule for the sellers (listed in Table 2) of painting services. The graph shows the corresponding supply curve. The height of the supply curve reflects the sellers' costs.

Price	Sellers	Quantity Supplied
$3,600 or more	Vincent, Claude, Pablo, Andy	4
$3,200 to $3,600	Claude, Pablo, Andy	3
$2,400 to $3,200	Pablo, Andy	2
$2,000 to $2,400	Andy	1
Less than $2,000	None	0

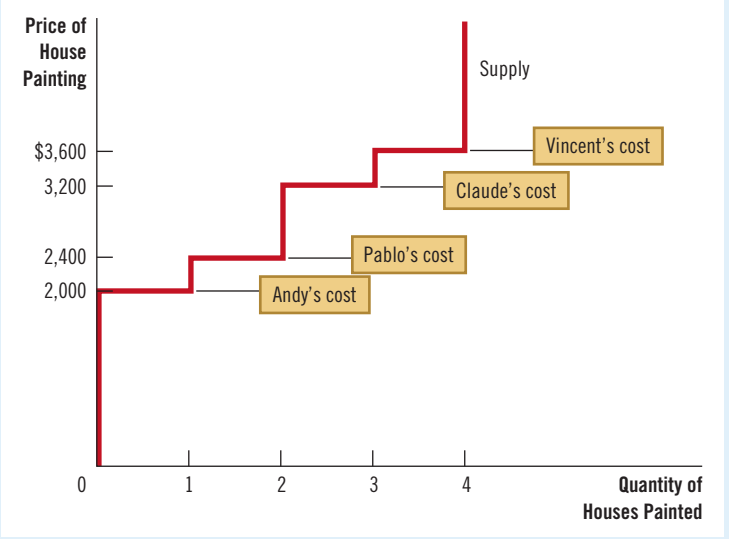
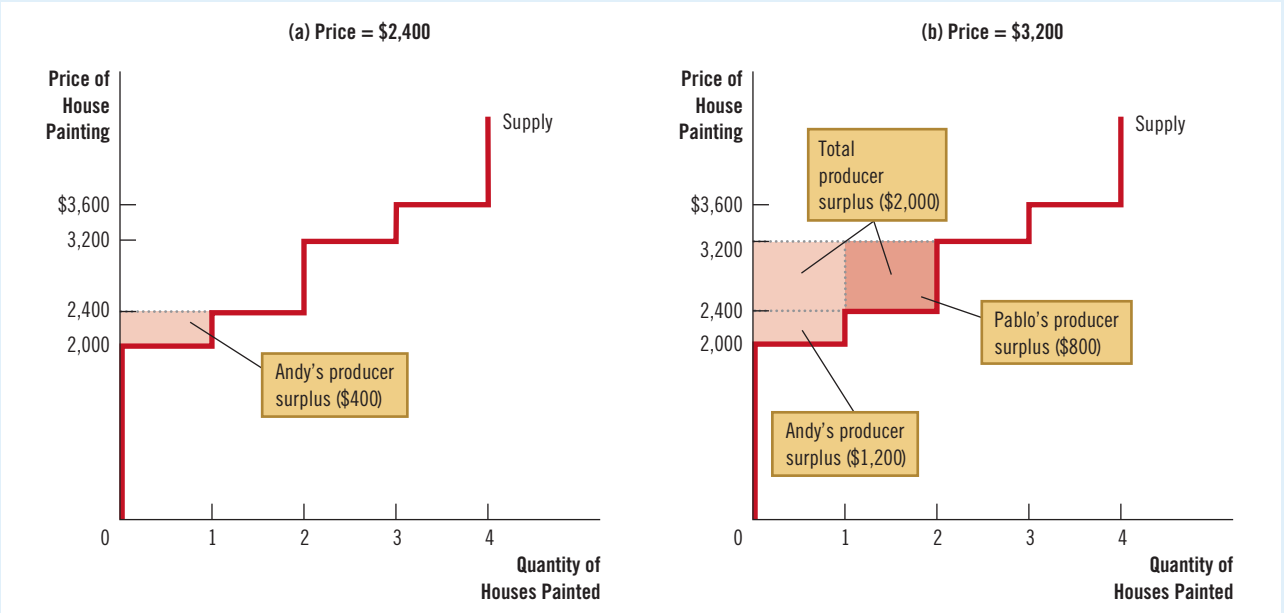


Figure 5

Measuring Producer Surplus with the Supply Curve

In panel (a), the price of the good is $2,400, and producer surplus is $400. In panel (b), the price is $3,200, and producer surplus is $2,000.



The lesson from this example applies to all supply curves: **The area below the price and above the supply curve measures the producer surplus in a market.** The logic is straightforward: The height of the supply curve measures sellers' costs, and the difference between the price and the cost of production is each seller's producer surplus. Thus, the area between the price line and the supply curve is the sum of all sellers' producer surplus.

7-2c How a Higher Price Raises Producer Surplus

You will not be surprised to hear that sellers generally prefer a higher price for the goods they sell. But how much does sellers' well-being increase in response to a higher price? The concept of producer surplus offers an answer.

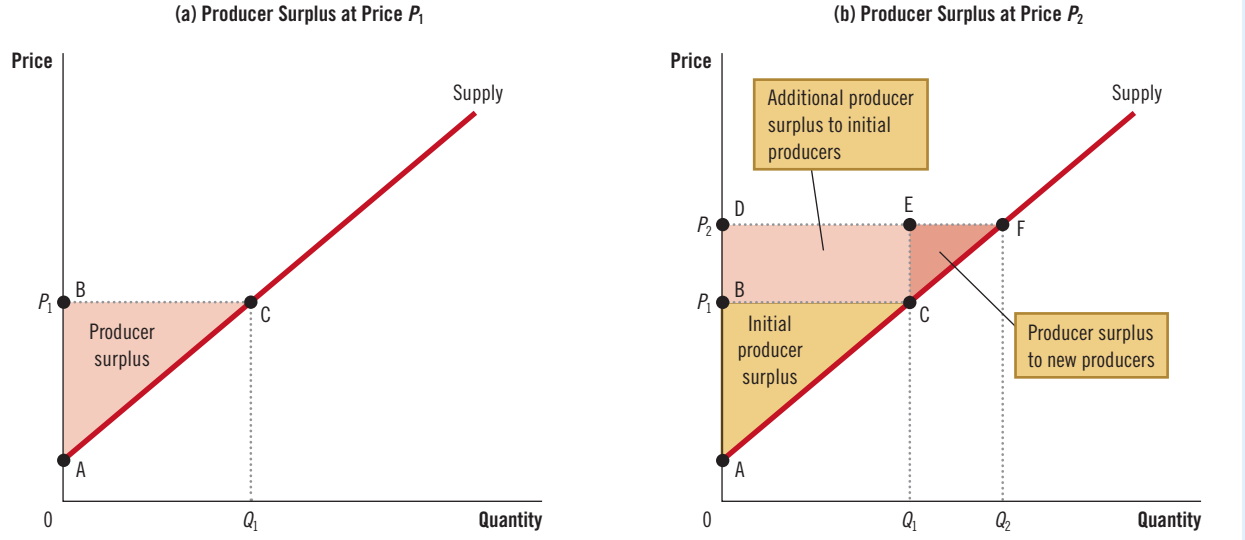
Figure 6 shows a typical upward-sloping supply curve that would arise in a market with many sellers. Although this supply curve differs in shape from the previous one, producer surplus is measured in the same way: Producer surplus is the area below the price and above the supply curve. In panel (a), the price is P_1, and producer surplus is the area of triangle ABC.

Panel (b) shows what happens when the price rises from P_1 to P_2. Producer surplus now equals area ADF. This increase in producer surplus has two parts. First, those sellers who were already selling Q_1 of the good at the lower price of P_1 get more for what they sell. The increase in producer surplus for these existing sellers equals the area of the rectangle BCED. Second, new sellers enter the market at the higher price, so the quantity supplied increases from Q_1 to Q_2. The producer surplus of these newcomers is the area of the triangle CEF.

Figure 6

How Price Affects Producer Surplus

In panel (a), the price is P_1, the quantity supplied is Q_1, and producer surplus equals the area of triangle ABC. When the price rises from P_1 to P_2, as in panel (b), the quantity supplied rises from Q_1 to Q_2, and producer surplus increases to the area of the triangle ADF. The increase in producer surplus (area BCFD) occurs in part because existing producers receive more at the higher price (area BCED) and in part because the higher price induces new producers to enter the market (area CEF).



As this analysis shows, producer surplus measures the well-being of sellers in much the same way that consumer surplus measures the well-being of buyers. Because these two measures of economic welfare are so similar, it is natural to consider them together, as we do in the next section.

QuickQuiz

4. Diego, Emi, and Finn are available to work as tutors for the semester. The opportunity cost of tutoring is \$400 for Diego, \$200 for Emi, and \$100 for Finn. The university is hiring tutors at a price of \$300. Producer surplus in this market is
   a. \$100.
   b. \$200.
   c. \$300.
   d. \$400.

5. Gavin has been working full-time as a gardener for \$300 a week. When the market price rises to \$400, Hector becomes a gardener as well. How much does producer surplus rise because of this price increase?
   a. by less than \$100
   b. between \$100 and \$200
   c. between \$200 and \$300
   d. by more than \$300

6. The supply curve for a product is $Q^S = 2P$, and the market price is \$10. What is producer surplus in this market? (Hint: Graph the supply curve and recall the formula for the area of a triangle.)
   a. \$5
   b. \$20
   c. \$100
   d. \$200

Answers are at the end of the chapter.

7-3 Market Efficiency

Consumer surplus and producer surplus are the basic tools that economists use to study the welfare of buyers and sellers in a market. These tools can help address a fundamental question: Do competitive markets reach a desirable allocation of resources?

7-3a Benevolent Social Planners

To evaluate market outcomes, we introduce a new, hypothetical group: the committee of the benevolent social planners. The benevolent social planners are all-powerful, all-knowing, and well-intentioned. They want to maximize the economic well-being of everyone in society. What should they do? Just let buyers and sellers find an equilibrium on their own? Or can the planners enhance well-being by somehow altering the market outcome?

To answer this question, the planners must first decide how to gauge the well-being of society. One input into measuring societal well-being is the sum of consumer surplus and producer surplus, which we call **total surplus**. Consumer surplus is the benefit that buyers receive from participating in a market, and producer surplus is the benefit that sellers receive. Total surplus is, therefore, a natural variable for the social planners to consider when judging a market's allocation of resources.

To better understand this measure, recall the definitions of consumer and producer surplus. Consumer surplus is:

$$\text{Consumer surplus} = \text{Value to buyers} - \text{Amount paid by buyers}.$$

Similarly, producer surplus is:

$$\text{Producer surplus} = \text{Amount received by sellers} - \text{Cost to sellers}.$$

Adding consumer and producer surplus together, we obtain:

$$\begin{aligned}\text{Total surplus} = {} & (\text{Value to buyers} - \text{Amount paid by buyers}) \\ & + (\text{Amount received by sellers} - \text{Cost to sellers}).\end{aligned}$$

Here, the amount paid by buyers equals the amount received by sellers, so the middle two terms cancel each other. As a result:

$$\text{Total surplus} = \text{Value to buyers} - \text{Cost to sellers}.$$

The total surplus in a market is the total value to buyers of the goods, measured by their willingness to pay, minus the total cost to sellers of providing those goods.

If an allocation of resources maximizes total surplus, economists say that the allocation exhibits **efficiency**. If an allocation is not efficient, some of the potential gains from trade among buyers and sellers are not being realized. For example, an allocation is inefficient if a good is not being produced by the sellers with the lowest costs. In this case, moving production from a high-cost producer to a lower-cost producer would reduce the total cost to sellers and raise total surplus. Similarly, an allocation is inefficient if a good is not being consumed by the buyers who are willing to pay the most for it. In this case, moving consumption of the good from a buyer with a low valuation to a buyer with a higher valuation would raise total surplus.

efficiency
the property regarding a resource allocation of maximizing the total surplus received by all members of society

equality
the property of distributing economic prosperity uniformly among the members of society

In addition to efficiency, the social planners might also care about **equality**—that is, whether the various buyers and sellers in the market have similar levels of economic well-being. In essence, the gains from trade in a market are like a pie to be shared among the market participants. The question of efficiency concerns whether the pie is as big as possible. The question of equality concerns how the pie is sliced and distributed among members of society. This chapter focuses on efficiency as the social planners' criterion. Keep in mind, however, that real policymakers often care about equality as well.

7-3b Evaluating the Market Equilibrium

Figure 7 shows the welfare measures when a market reaches the equilibrium of supply and demand. Recall that consumer surplus equals the area above the price and under the demand curve, and producer surplus equals the area below the price and above the supply curve. The area between the supply and demand curves up to the point of equilibrium represents the total surplus in this market.

Is this equilibrium allocation of resources efficient? That is, does it maximize total surplus? Recall that when a market is in equilibrium, the price determines which of the possible buyers and sellers participate in the market. Buyers who value the good more than the price (represented by the segment AE on the demand curve) choose to buy it; buyers who value it less than the price (represented by the segment EB) do not. Similarly, sellers whose costs are less than the price (represented by the segment CE on the supply curve) choose to produce and sell the good; sellers whose costs are greater than the price (represented by the segment ED) do not.

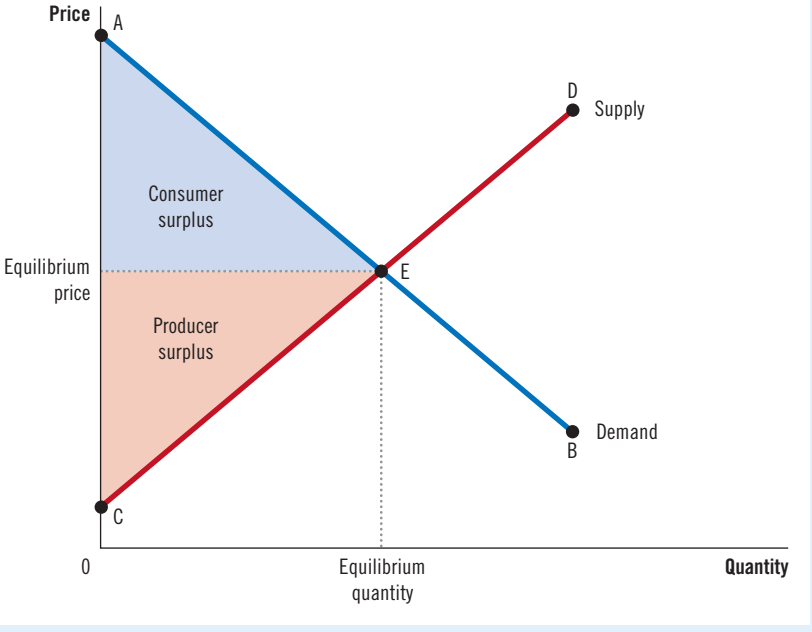
These observations lead to two insights about market outcomes:

1. Competitive markets allocate the supply of goods to the buyers who value them most, as measured by their willingness to pay.
2. Competitive markets allocate the demand for goods to the sellers who can produce them at the lowest cost.

Figure 7

Consumer and Producer Surplus in the Market Equilibrium

Total surplus—the sum of consumer and producer surplus—is the area between the supply and demand curves up to the equilibrium quantity.



Thus, given the quantity produced and sold in a market equilibrium, the social planners cannot increase economic well-being by changing the allocation of consumption among buyers or the allocation of production among sellers.

But can the social planners raise well-being by increasing or decreasing the quantity of the good? The answer is no, as stated in this third insight about market outcomes:

3. Competitive markets produce the quantity of goods that maximizes the sum of consumer and producer surplus.

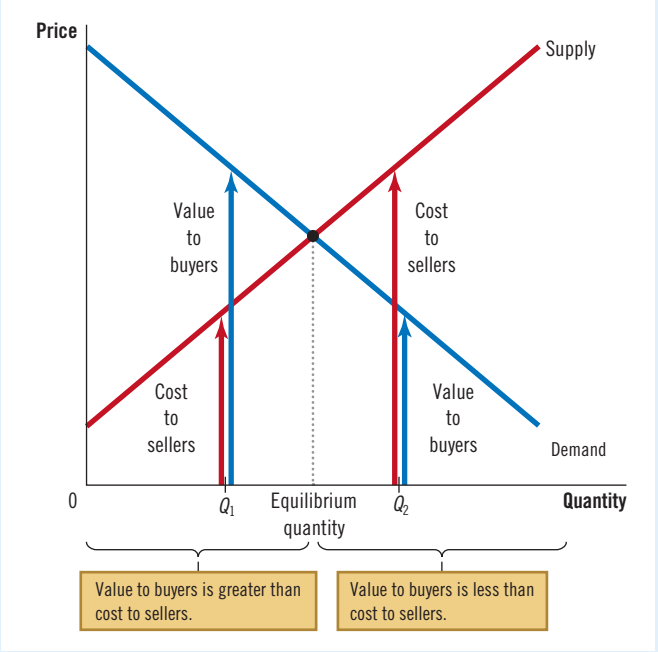
Figure 8 shows why this is true. To interpret this figure, recall that the demand curve reflects the value to buyers and the supply curve reflects the cost to sellers. At any quantity below the equilibrium level, such as Q_1, the value to the marginal buyer exceeds the cost to the marginal seller. As a result, increasing the quantity produced and consumed raises total surplus. This continues to be true until the quantity reaches the equilibrium level. Conversely, at any quantity beyond the equilibrium level, such as Q_2, the value to the marginal buyer is less than the cost to the marginal seller. In this case, decreasing the quantity raises total surplus, and this continues to be true until the quantity falls to the equilibrium level. To maximize total surplus, the social planners would choose the quantity at which the supply and demand curves intersect.

Together, these three insights tell us that the market outcome maximizes the sum of consumer and producer surplus. In other words, the equilibrium outcome is an efficient allocation of resources. Social planners concerned about efficiency can, therefore, leave the market outcome just as they find it. This policy of leaving well enough alone goes by the French expression **laissez-faire**, which literally translates to "leave to do" but is more broadly interpreted as "let people do as they will."

Figure 8

The Efficiency of the Equilibrium Quantity

At quantities less than the equilibrium quantity, such as Q_1, the value to buyers exceeds the cost to sellers. At quantities greater than the equilibrium quantity, such as Q_2, the cost to sellers exceeds the value to buyers. Therefore, the market equilibrium maximizes the sum of producer and consumer surplus.



Society is lucky that the social planners don't need to intervene. Imagining what all-powerful, all-knowing, and well-intentioned planners would do has been a useful exercise, but let's face it: Such characters are hard to come by. Some dictators may be close to all powerful, but they are rarely benevolent. Even if we found some people so virtuous, they would lack crucial information.

Suppose the social planners tried to choose an efficient allocation of resources on their own instead of relying on market forces. To do so, they would need to know the value of a particular good to every potential consumer and the cost for every potential producer. And they would need this information not only for this market but for every one of the many thousands of markets in the economy. This task is practically impossible, which explains why economies with substantial central planning are rife with inefficiencies.

The planners' job becomes easy, however, once they take on a partner: Adam Smith's invisible hand of the marketplace. The invisible hand takes all the information about buyers and sellers into account and guides everyone in the market to the best outcome as judged by the standard of economic efficiency. It is a remarkable feat. That is why economists so often advocate unfettered, competitive markets as the best way to organize economic activity.

Case Study

Should There Be a Market for Organs?

Some years ago, *The Boston Globe* ran an article "How a Mother's Love Helped Save Two Lives." It told the story of Susan Stephens, a woman whose son needed a kidney transplant. When the doctor learned that the mother's kidney was not compatible, he proposed a novel solution: If Stephens donated one of her kidneys to a stranger, her son would move to the top of the kidney waiting list. The mother accepted the deal, and soon two patients had the transplants they were waiting for.

The ingenuity of the doctor's proposal and the nobility of the mother's act cannot be doubted. But the story raises intriguing questions. If the mother could trade a kidney for a kidney, would the hospital allow her to trade a kidney for an expensive, experimental cancer treatment that she could not otherwise afford? Should she be allowed to exchange her kidney for free tuition for her son at the hospital's medical school? Should she be able to sell her kidney and use the cash to trade in her old Chevy for a new Lexus?

As a matter of public policy, it is illegal for people to sell their organs. Many people view the very notion of buying and selling human organs as repugnant, perhaps because it violates cultural and religious norms about the sanctity of life. But put aside that reaction for a moment and think about kidneys as a good subject to market forces. In essence, in the market for kidneys, the government has imposed a price ceiling of zero. The result, as with any binding price ceiling, is a shortage. The deal in the Stephens case did not fall under this prohibition because no cash changed hands. In a legal sense, it was not a market transaction.

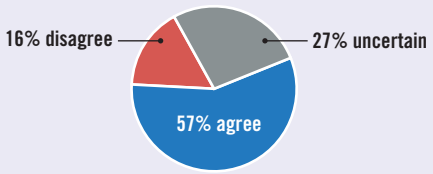
Yet many economists say that repealing this prohibition and allowing an open market for organs would yield large benefits. People are born with two kidneys, but they usually need only one. Meanwhile, some people suffer from illnesses that leave

Ask the Experts

Supplying Kidneys

"A market that allows payment for human kidneys should be established on a trial basis to help extend the lives of patients with kidney disease."

What do economists say?



Source: IGM Economic Experts Panel, March 11, 2014.

them without any working kidney. Despite the obvious gains from trade, the current situation is dire: The typical patient has to wait several years for a kidney transplant, and every year thousands of people die because a compatible kidney cannot be found. Proponents of a market solution say that if those needing a kidney could buy one from those who have two, the price would rise to balance supply and demand. Sellers would be better off with the extra cash in their pockets. Buyers would be better off with the organ they need to save their lives. The shortage of kidneys would disappear.

Such a market would lead to an efficient allocation of resources, but this plan nonetheless has its critics. In addition to the issue of repugnance, some worry about fairness. A market for organs, they argue, would benefit the rich at the expense of the poor because organs would be allocated to those most willing and able to pay. But you can also question the fairness of the current system. Now, most of us walk around with an extra organ that we don't really need, while some of our fellow citizens are dying to get one. Is that fair? ●

QuickQuiz

7. Isabelle values her time at $60 an hour. She spends 2 hours giving Jayla a massage. Jayla was willing to pay as much as $300 for the massage, but they negotiated a price of $200. In this transaction,
   a. consumer surplus is $20 larger than producer surplus.
   b. consumer surplus is $40 larger than producer surplus.
   c. producer surplus is $20 larger than consumer surplus.
   d. producer surplus is $40 larger than consumer surplus.
8. An efficient allocation of resources maximizes
   a. consumer surplus.
   b. producer surplus.
   c. consumer surplus plus producer surplus.
   d. consumer surplus minus producer surplus.
9. When a market is in equilibrium, the buyers are those with the _____ willingness to pay, and the sellers are those with the _____ costs.
   a. highest; highest
   b. highest; lowest
   c. lowest; highest
   d. lowest; lowest
10. Producing a quantity larger than the equilibrium of supply and demand is inefficient because the marginal buyer's willingness to pay is
    a. negative.
    b. zero.
    c. positive but less than the marginal seller's cost.
    d. positive and greater than the marginal seller's cost.

Answers are at the end of the chapter.

7-4 Conclusion: Market Efficiency and Market Failure

This chapter introduced the basic tools of welfare economics—consumer and producer surplus—and used them to evaluate the efficiency of market outcomes. We showed that the forces of supply and demand allocate resources efficiently. Even though buyers and sellers in a market are each concerned only about their own welfare, they are guided by an invisible hand to an equilibrium that maximizes the total benefits to buyers and sellers.

In the News

How Ticket Resellers Help Allocate Scarce Resources

Is ticket reselling a scourge or a way to make markets more efficient?

Scalping Isn't Scamming

By Tracy C. Miller

The cost of tickets to the Broadway musical "Hamilton" skyrocketed at one point over the summer as scalpers charged $1,000 or more for tickets to the show, when the average ticket's face value was $189. In response, Sen. Chuck Schumer (D-NY) is proposing federal legislation that would prohibit the use of software to facilitate ticket scalping. Do we really need legislation to curb this practice?

Scalpers are using bots to buy up a large share of tickets online before the public gets a chance to purchase them. Then they resell those tickets for much higher prices. This is a modern twist on a practice that has long been demonized by the public and legislators.

Scalping certainly results in some consumers paying higher prices than they otherwise would. But in exchange for high prices, consumers can get the tickets they want, when they want them, without waiting in line or competing to be among the first to buy them online at a given time. Opponents mistakenly conclude that high prices are the fault of scalpers, when in fact prices are high because of a large demand and a limited supply.

At present, no federal laws limit scalping, but 15 states have laws that prohibit scalping in at least some circumstances. Another seven states require a seller to have a license to broker a ticket, and some limit how much ticket brokers can mark up the price of tickets. Some states don't allow scalping within a specified distance of the venue where an event is held. Others allow reselling tickets purchased for personal use, while prohibiting anyone not registered as a broker from buying and selling tickets for a profit.

Scalping benefits the scalper and the buyer, by getting tickets to whomever values them most highly. If someone decides at the last minute to attend a play, a concert or a game, they can find tickets at some price. Without scalpers, some people who value the event highly would be unable to buy tickets for seats of the quality they desire.

Scalping can also benefit ticket producers—the sports teams or performing artists who supply tickets—in two ways. First, it enables them to earn ticket revenue through face-value prices long before an event, while scalpers bear the risk that demand and prices might fall below the price they paid. Second, because of scalpers, the initial demand for tickets may be higher than it would otherwise be, enabling ticket producers to charge more.

Ticket producers incur expenses long before an event, such as the cost of renting an arena. They can keep their selling costs down by selling all or most tickets quickly rather than over an extended period of time. By buying tickets

A few words of warning are in order. To conclude that markets are efficient, we made several assumptions about how markets work. When these assumptions do not hold, the conclusion that the market equilibrium is efficient may no longer be true. As we close this chapter, let's briefly consider two of the most important assumptions we made.

First, our analysis assumed that markets are perfectly competitive. In actual economies, however, competition is sometimes far from perfect. In some markets, a single buyer or seller (or a small group of them) may be able to control market prices. This ability to influence prices is called **market power**. Market power can make markets inefficient by keeping the price and quantity away from the levels that equilibrate supply and demand.

Second, our analysis assumed that the market outcome matters only to the buyers and sellers who participate in the market. Yet sometimes, the decisions of buyers and sellers affect bystanders. Pollution is the classic example. The use of agricultural pesticides, for instance, affects not only the manufacturers who make them and the farmers who use them but also many others who breathe the air or drink the water contaminated by these pesticides. When a market exhibits such side effects, called **externalities**, the welfare implications of market activity depend on more than

when they first become available and holding an inventory to sell at times that are most convenient to consumers, scalpers connect buyers with sellers and benefit both. They act as brokers, and the difference between the price they pay and the price they receive is their reward for doing this. The more scalpers compete to buy and resell tickets, the lower the markup that each will earn.

If scalpers are few in number and skilled at assessing each consumer's demand for tickets, they can charge each consumer a price close to the maximum he or she is willing to pay. The higher the average price they can charge per ticket, the more they can pay to the team or performing artists who produce the tickets.

Scalping does alienate some consumers who pay higher prices to buy from scalpers who got to the ticket site before they did. As a result, these consumers may be less willing to attend future events. If performing artists or sports teams want to avoid alienating their loyal customers, they can choose their method of distributing tickets to accomplish that goal, such as by setting aside a percentage of tickets to sell at what they consider a reasonable price to those customers. In many cases, though, ticket producers may prefer to lock-in a high price and sell all their tickets quickly, which may mean selling a large percentage of their tickets to scalpers.

THEO WARGO/WIREIMAGE/GETTY IMAGES

Lin-Manuel Miranda as Hamilton

Laws to prevent scalping are unnecessary and prevent mutually beneficial transactions. Scalping only occurs when original ticket sellers charge a price that's lower than some consumers are willing to pay. If scalpers use software that's efficient at buying and selling tickets, it will save time and effort and each party involved in the process benefits. In one way or another, the ticket producer, the scalper and the people who attend the event will each be better off. ■

Questions to Discuss

1. Why do you think the producers of *Hamilton* charge much less for tickets than the ticket resellers charge?
2. Do you think there should be laws against reselling tickets above their face value? Why or why not?

Mr. Miller is an economist at the Mercatus Center at George Mason University.

Source: *U.S. News and World Report*, October 4, 2016.

just the value realized by buyers and the cost incurred by sellers. Because buyers and sellers may ignore these externalities when deciding how much to consume and produce, the equilibrium in a market can be inefficient from the standpoint of society as a whole.

Market power and externalities are examples of a general phenomenon called **market failure**—the inability of some unregulated markets to allocate resources efficiently. When markets fail, public policy can potentially remedy the problem and enhance economic efficiency. Microeconomists devote much effort to studying when market failures are likely and how they are best corrected. As you continue your study of economics, you will see that the tools of welfare economics developed here are readily adapted to that endeavor.

Despite the possibility of market failure, the invisible hand of the marketplace is extraordinarily important. In many markets, the assumptions made in this chapter work well, and the conclusion of market efficiency applies directly. Moreover, we can use our analysis of welfare economics and market efficiency to shed light on the effects of various government policies. The next two chapters apply the tools we have just developed to two important policy issues—the welfare effects of taxation and of international trade.

Chapter in a Nutshell

- Consumer surplus equals buyers' willingness to pay for a good minus the amount they actually pay, and it measures the benefit buyers get from participating in a market. Consumer surplus can be found by computing the area below the demand curve and above the price.
- Producer surplus equals the amount sellers are paid for their goods minus their costs of production, and it measures the benefit sellers get from participating in a market. Producer surplus can be found by computing the area below the price and above the supply curve.
- An allocation of resources that maximizes total surplus (the sum of consumer and producer surplus) is said to be efficient. Policymakers are often concerned with the efficiency, as well as the equality, of economic outcomes.
- Under normal conditions, the equilibrium of supply and demand maximizes total surplus. That is, the invisible hand of the marketplace usually leads buyers and sellers in competitive markets to allocate resources efficiently.
- Markets do not allocate resources efficiently in the presence of market failures such as market power or externalities.

Key Concepts

welfare economics, p. 134
willingness to pay, p. 134
consumer surplus, p. 135
cost, p. 139
producer surplus, p. 139
efficiency, p. 143
equality, p. 144

Questions for Review

1. Explain how buyers' willingness to pay, consumer surplus, and the demand curve are related.
2. Explain how sellers' costs, producer surplus, and the supply curve are related.
3. In a supply-and-demand diagram, show producer surplus and consumer surplus at the market equilibrium.
4. What is efficiency? Is it the only goal of economic policymakers?
5. Name two types of market failure. Explain why each may cause market outcomes to be inefficient.

Problems and Applications

1. Kyra buys an iPhone for $360 and gets consumer surplus of $240.
   a. What is her willingness to pay?
   b. If she had bought the iPhone on sale for $270, what would her consumer surplus have been?
   c. If the price of an iPhone were $750, what would her consumer surplus have been?
2. An early freeze in California sours the lemon crop. Explain what happens to consumer surplus in the market for lemons. Explain what happens to consumer surplus in the market for lemonade. Illustrate your answers with diagrams.
3. Suppose the demand for French bread rises. Explain what happens to producer surplus in the market for French bread. Explain what happens to producer surplus in the market for flour. Illustrate your answers with diagrams.
4. It is a hot day, and Bert is thirsty. Here is the value he places on each bottle of water:

Value of first bottle	$7
Value of second bottle	$5
Value of third bottle	$3
Value of fourth bottle	$1

a. From this information, derive Bert's demand schedule. Graph his demand curve for bottled water.
b. If the price of a bottle of water is $4, how many bottles does Bert buy? How much consumer surplus does Bert get from his purchases? Show Bert's consumer surplus in your graph.
c. If the price falls to $2, how does the quantity demanded change? How does Bert's consumer surplus change? Show these changes in your graph.

5. Ernie owns a water pump. Because pumping large amounts of water is harder than pumping small amounts, the cost of producing a bottle of water rises as he pumps more. Here is the cost he incurs to produce each bottle of water:

Cost of first bottle	$1
Cost of second bottle	$3
Cost of third bottle	$5
Cost of fourth bottle	$7

a. From this information, derive Ernie's supply schedule. Graph his supply curve for bottled water.
b. If the price of a bottle of water is $4, how many bottles does Ernie produce and sell? How much producer surplus does Ernie get from these sales? Show Ernie's producer surplus in your graph.
c. If the price rises to $6, how does the quantity supplied change? How does Ernie's producer surplus change? Show these changes in your graph.

6. Consider a market in which Bert from problem 4 is the buyer and Ernie from problem 5 is the seller.
a. Use Ernie's supply schedule and Bert's demand schedule to find the quantity supplied and quantity demanded at prices of $2, $4, and $6. Which of these prices brings supply and demand into equilibrium?
b. What are consumer surplus, producer surplus, and total surplus in this equilibrium?
c. If Ernie produced and Bert consumed one fewer bottle of water, what would happen to total surplus?
d. If Ernie produced and Bert consumed one additional bottle of water, what would happen to total surplus?

7. The cost of producing flat-screen TVs has fallen over the past decade. Let's consider some implications of this change.
a. Draw a supply-and-demand diagram to show the effect of falling production costs on the price and quantity of flat-screen TVs sold.
b. In your diagram, show what happens to consumer surplus and producer surplus.
c. Suppose the supply of flat-screen TVs is very elastic. Who benefits most from falling production costs—consumers or producers of these TVs?

8. Four consumers are willing to pay these amounts for haircuts:

Gloria: $35	Jay: $10	Claire: $40	Phil: $25

Four haircutting businesses have these costs:

Firm A: $15	Firm B: $30	Firm C: $20	Firm D: $10

Each firm can give, at most, one haircut. To achieve efficiency, how many haircuts should be given? Which businesses should cut hair and which consumers should have their hair cut? How large is the maximum possible total surplus?

9. One of the largest changes in the economy over the past several decades is that technological advances have reduced the cost of making computers.
a. Draw a supply-and-demand diagram to show what happened to the price, quantity, consumer surplus, and producer surplus in the computer market.
b. Forty years ago, students used typewriters to prepare papers for their classes; today, they use computers. Does that make computers and typewriters complements or substitutes? Use a supply-and-demand diagram to show what happened to the price, quantity, consumer surplus, and producer surplus in the market for typewriters. Should typewriter producers have been happy or sad about the technological advance in computers?
c. Are computers and software complements or substitutes? Draw a supply-and-demand diagram to show what happened to the price, quantity, consumer surplus, and producer surplus in the market for software. Should software producers have been happy or sad about the technological advance in computers?
d. Does this analysis help explain why software producer Bill Gates became one of the world's richest people?

10. A friend of yours is considering two movie-streaming services. Provider A charges $120 per year regardless

of the number of movies streamed. Provider B does not have a fixed service fee but instead charges $1 per movie. Your friend's annual demand for movies is given by the equation $Q^D = 150 - 50P$, where P is the price per movie.

a. With each provider, what is the cost to your friend of an extra movie?
b. In light of your answer to (a), how many movies with each provider would your friend watch?
c. How much would she end up paying each provider every year?
d. How much consumer surplus would she obtain with each provider? (Hint: Graph the demand curve and recall the formula for the area of a triangle.)
e. Which provider would you recommend that your friend choose? Why?

QuickQuiz Answers

1. **b** 2. **b** 3. **d** 4. **c** 5. **b** 6. **c** 7. **a** 8. **c** 9. **b** 10. **c**

Chapter

8

Application: The Costs of Taxation

Taxes are often a source of heated political debate. In 1776, the anger of the American colonists over British taxes sparked the American Revolution. More than two centuries later, Americans still debate the proper size and shape of the tax system. Yet few would deny that some taxation is necessary. As the jurist Oliver Wendell Holmes Jr. once said, "Taxes are what we pay for civilized society."

Because taxation has a large impact on the modern economy, the topic appears repeatedly throughout this book as we expand the tools at our disposal. Chapter 6 used supply, demand, and elasticity to show how a tax on a good affects its price and quantity and how the tax burden is split between buyers and sellers. This chapter extends the analysis to examine how taxes affect welfare, defined as the economic well-being of market participants.

The effects of taxes on welfare might seem obvious. The government enacts taxes to raise revenue, and that revenue must come from someone's wallet. As Chapter 6 showed, both buyers and sellers of a good are worse off when it is taxed: Buyers pay more, and sellers receive less. Yet to fully understand how taxes affect well-being, we must compare the losses of buyers and sellers with the amount of revenue the government collects. The tools of consumer

BLACKCAT/E+/GETTY IMAGES, GOWITHSTOCK/SHUTTERSTOCK.COM

and producer surplus enable this comparison. The analysis will show that the cost of taxes to buyers and sellers typically exceeds the revenue raised by the government.

That is not to say that taxation is always undesirable. Tax revenue is needed to fund government programs, and these can be valuable. But judging alternative policies requires an understanding of how high the price of civilized society can be.

8-1 The Deadweight Loss of Taxation

Let's begin by recalling a lesson from Chapter 6: The ultimate impact of a tax on a market is the same whether the tax is levied on buyers or sellers. A tax on buyers shifts the demand curve downward by the size of the tax; a tax on sellers shifts the supply curve upward by that amount. In either case, the tax raises the price paid by buyers and reduces the price received by sellers. As a result, how the tax burden is distributed between producers and consumers depends not on how the tax is levied but on the elasticities of supply and demand.

Figure 1 illustrates the effect of a tax. To keep things simple, this figure does not show a shift in the supply or demand curve, though one curve must shift, depending on whom the tax is levied. This chapter keeps the analysis general and the graphs less cluttered by not showing the shift. The important point is that the tax places a wedge between the price buyers pay and the price sellers receive. Because of this wedge, the quantity sold is less than it would be if there were no tax. In other words, a tax on a good shrinks the size of the market for it, as Chapter 6 showed.

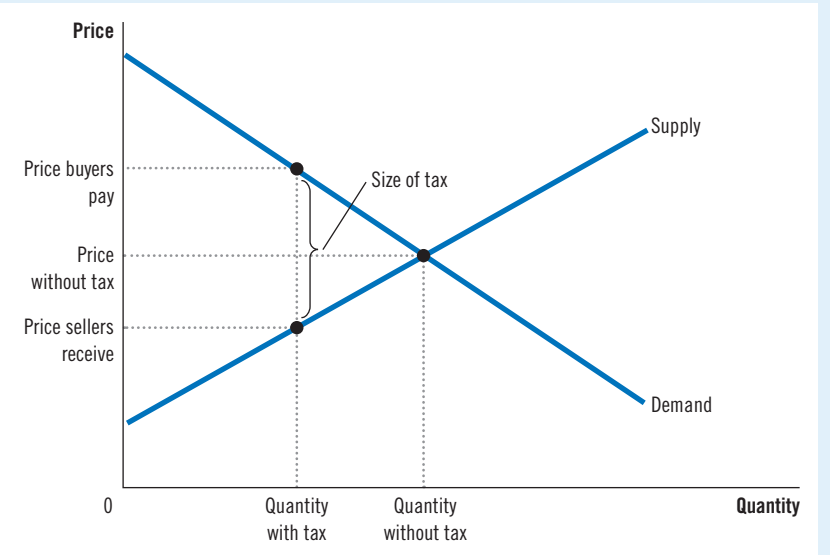
8-1a How a Tax Affects Market Participants

Let's now use the tools of welfare economics to measure the gains and losses from a tax on a good. To do this, we must consider how the tax affects buyers, sellers, and the government. The welfare of buyers is measured by consumer surplus—the amount buyers are willing to pay for the good minus the amount they actually pay

Figure 1

The Effects of a Tax

A tax on a good places a wedge between the price that buyers pay and the price that sellers receive. The quantity of the good sold declines.



for it. The welfare of sellers is measured by producer surplus—the amount sellers receive for the good minus their costs of producing it. Chapter 7 introduced these measures of economic well-being.

What about the third interested party, the government? If T is the size of the tax and Q is the quantity of the good sold, then the government takes in tax revenue of $T \times Q$. It can use this revenue to provide government services, such as roads, police, and public education, or to fund transfer programs for, say, lower-income households. When analyzing how taxes affect economic well-being, we use the government's tax revenue to measure the public benefit from the tax. This benefit, however, actually accrues not to the government but to those on whom the revenue is spent.

In Figure 2, the government's tax revenue is represented by the rectangle between the supply and demand curves. The height of this rectangle is the size of the tax, T, and its width is the quantity of the good sold, Q. Because a rectangle's area is its height multiplied by its width, this rectangle's area is $T \times Q$, which equals the tax revenue.

J.B. HANDELSMAN/THE NEW YORKER COLLECTION/THE CARTOON BANK

"You know, the idea of taxation with representation doesn't appeal to me very much, either."

Welfare without a Tax To see how a tax affects welfare, let's begin by considering the situation before a tax is imposed. Figure 3 shows the supply-and-demand diagram with the key areas marked by the letters A through F.

Without a tax, the equilibrium price and quantity are found at the intersection of the supply and demand curves. The price is P_1, and the quantity sold is Q_1. Because the demand curve reflects buyers' willingness to pay, consumer surplus is the area between the demand curve and the price, A + B + C. Similarly, because the supply curve reflects sellers' costs, producer surplus is the area between the supply curve and the price, D + E + F. Because there is no tax, tax revenue is zero.

Total surplus, the sum of consumer and producer surplus, equals the area A + B + C + D + E + F. In other words, as in Chapter 7, it is the area between the supply and demand curves up to the equilibrium quantity. The first column of the table in Figure 3 summarizes these results.

Figure 2

Tax Revenue

The tax revenue that the government collects equals $T \times Q$, the size of the tax, T, times the quantity sold, Q. Thus, tax revenue equals the area of the rectangle between the supply and demand curves.

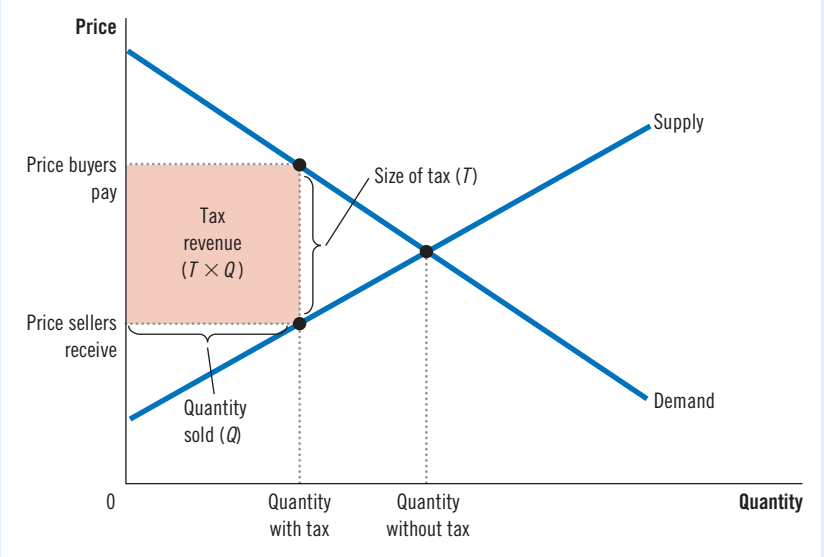


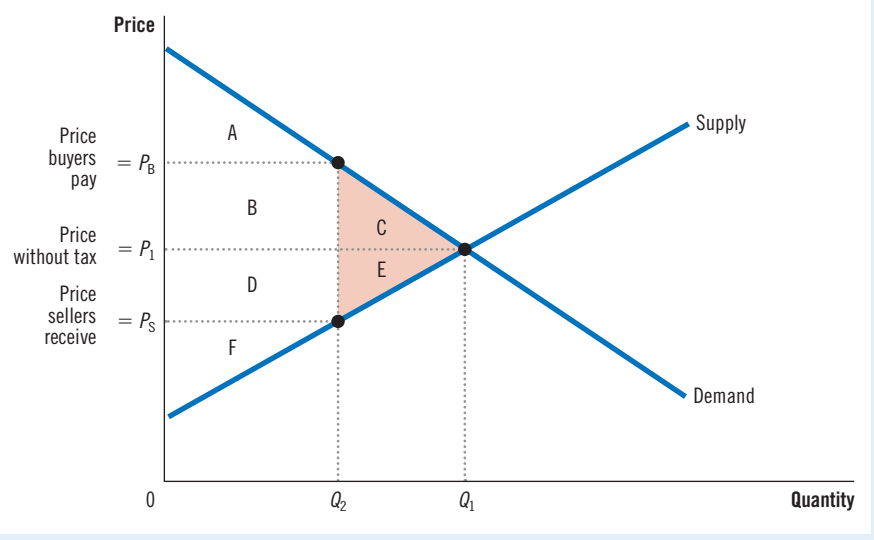
Figure 3

How a Tax Affects Welfare

A tax on a good reduces consumer surplus (by the area B + C) and producer surplus (by the area D + E). Because the fall in producer and consumer surplus exceeds the tax revenue (area B + D), the tax is said to impose a deadweight loss (area C + E).

	Without Tax	With Tax	Change
Consumer Surplus	A + B + C	A	− (B + C)
Producer Surplus	D + E + F	F	− (D + E)
Tax Revenue	None	B + D	+ (B + D)
Total Surplus	A + B + C + D + E + F	A + B + D + F	− (C + E)

The area C + E shows the fall in total surplus and is the deadweight loss of the tax.



Welfare with a Tax Now consider welfare with a tax. The price paid by buyers rises from P_1 to P_B, so consumer surplus equals only area A (the area below the demand curve and above the buyers' price, P_B). The price received by sellers falls from P_1 to P_S, so producer surplus equals only area F (the area above the supply curve and below the sellers' price P_S). The quantity sold falls from Q_1 to Q_2, and the government collects tax revenue equal to the area B + D.

To find total surplus with the tax, add consumer surplus, producer surplus, and tax revenue. Thus, total surplus is area A + B + D + F. The table's second column summarizes these results.

Changes in Welfare We can now see the effects of the tax by comparing welfare before and after the tax is enacted. The table's third column shows the changes. Consumer surplus falls by the area B + C, and producer surplus falls by the area D + E. Tax revenue rises by the area B + D. Not surprisingly, with the tax, the buyers and sellers are worse off, and the government has more revenue.

The change in total welfare includes the change in consumer surplus (which is negative), the change in producer surplus (which is also negative), and the change

in tax revenue (which is positive). When we add these three pieces together, we find that total surplus in the market falls by the area C + E. **The losses to buyers and sellers from a tax exceed the revenue raised by the government.** The fall in total surplus that results when a tax (or some other policy) distorts the outcome in an otherwise efficient market is called a **deadweight loss**. The area C + E measures the size of the loss.

deadweight loss
the fall in total surplus that results from a market distortion

To understand why taxes cause deadweight losses, recall one of the **Ten Principles of Economics** from Chapter 1: People respond to incentives. Chapter 7 showed that competitive markets typically allocate scarce resources efficiently. That is, in the absence of a tax, the equilibrium of supply and demand maximizes the total surplus of buyers and sellers in a market. When the government imposes a tax, it raises the price buyers pay and lowers the price sellers receive, giving incentives to buyers to consume less and sellers to produce less. As a result, the market shrinks below its optimum (as shown in the figure by the movement from Q_1 to Q_2). Thus, because taxes distort incentives, they cause markets to allocate resources inefficiently.

8-1b Deadweight Losses and the Gains from Trade

To better understand why taxes cause deadweight losses, consider an example. Imagine that Malik cleans Mei's house each week for $100. The opportunity cost of Malik's time is $80, and the value of a clean house to Mei is $120, so they each receive a $20 benefit from their deal. The total surplus of $40 measures the gains from trade in this transaction.

Now suppose that the government levies a $50 tax on providers of cleaning services. There is now no price that Mei can pay Malik that will leave both better off. The most Mei would be willing to pay is $120, but then Malik would be left with only $70 after paying the tax, which is less than his $80 opportunity cost. Conversely, for Malik to cover his opportunity cost of $80, Mei would need to pay $130, which is above the $120 value she places on the cleaning service. As a result, Mei and Malik cancel their arrangement. Malik loses the income, and Mei cleans her own house.

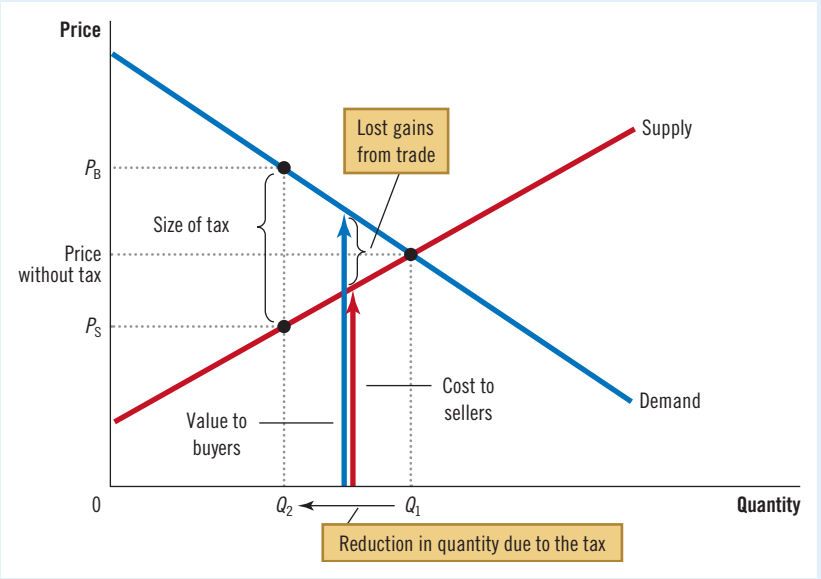
The tax has made Malik and Mei worse off by a total of $40 because they have each lost $20 of surplus. But the government collects no revenue from Malik and Mei because they have canceled their arrangement. The $40 is pure deadweight loss: It is a loss to buyers and sellers in a market that is not offset by an increase in government revenue. This example shows the ultimate source of deadweight losses: **Taxes cause deadweight losses because they prevent buyers and sellers from realizing some of the gains from trade.**

The area of the triangle between the supply and demand curves created by the tax wedge (area C + E in Figure 3) measures these losses. This conclusion can be seen more easily in Figure 4 by recalling that the demand curve reflects the value of the good to consumers and that the supply curve reflects the costs of producers. When the tax raises the price buyers pay to P_B and lowers the price sellers receive to P_S, the marginal buyers and sellers leave the market, so the quantity sold falls from Q_1 to Q_2. Yet as the figure shows, the value of the good to these buyers still exceeds the cost to these sellers. At every quantity between Q_1 and Q_2, the situation is the same as in our example with Malik and Mei. The gains from trade—the difference between buyers' value and sellers' cost—are less than the tax. As a result, these trades are not made once the tax is imposed. The deadweight loss is the surplus that is lost because the tax discourages these mutually advantageous trades.

Figure 4

The Source of a Deadweight Loss

When the government imposes a tax on a good, the quantity sold falls from Q_1 to Q_2. At every quantity between Q_1 and Q_2, the potential gains from trade among buyers and sellers are not realized. These lost gains from trade make up the deadweight loss.



QuickQuiz

1. A tax on a good has a deadweight loss if
   a. the reduction in consumer and producer surplus is greater than the tax revenue.
   b. the tax revenue is greater than the reduction in consumer and producer surplus.
   c. the reduction in consumer surplus is greater than the reduction in producer surplus.
   d. the reduction in producer surplus is greater than the reduction in consumer surplus.
2. Donna runs an inn and charges $300 a night for a room, which equals her cost. Sam, Harry, and Bill are three potential customers willing to pay $500, $325, and $250, respectively. When the government levies a tax on innkeepers of $50 per night of occupancy, Donna raises her price to $350. The deadweight loss of the tax is
   a. $25.
   b. $50.
   c. $100.
   d. $150.
3. Sophie pays Sky $50 to mow her lawn every week. When the government levies a mowing tax of $10 on Sky, he raises his price to $60. Sophie continues to hire him at the higher price. What is the change in producer surplus, the change in consumer surplus, and the deadweight loss?
   a. $0, $0, $10
   b. $0, −$10, $0
   c. +$10, −$10, $10
   d. +$10, −$10, $0

Answers are at the end of the chapter.

8-2 The Determinants of the Deadweight Loss

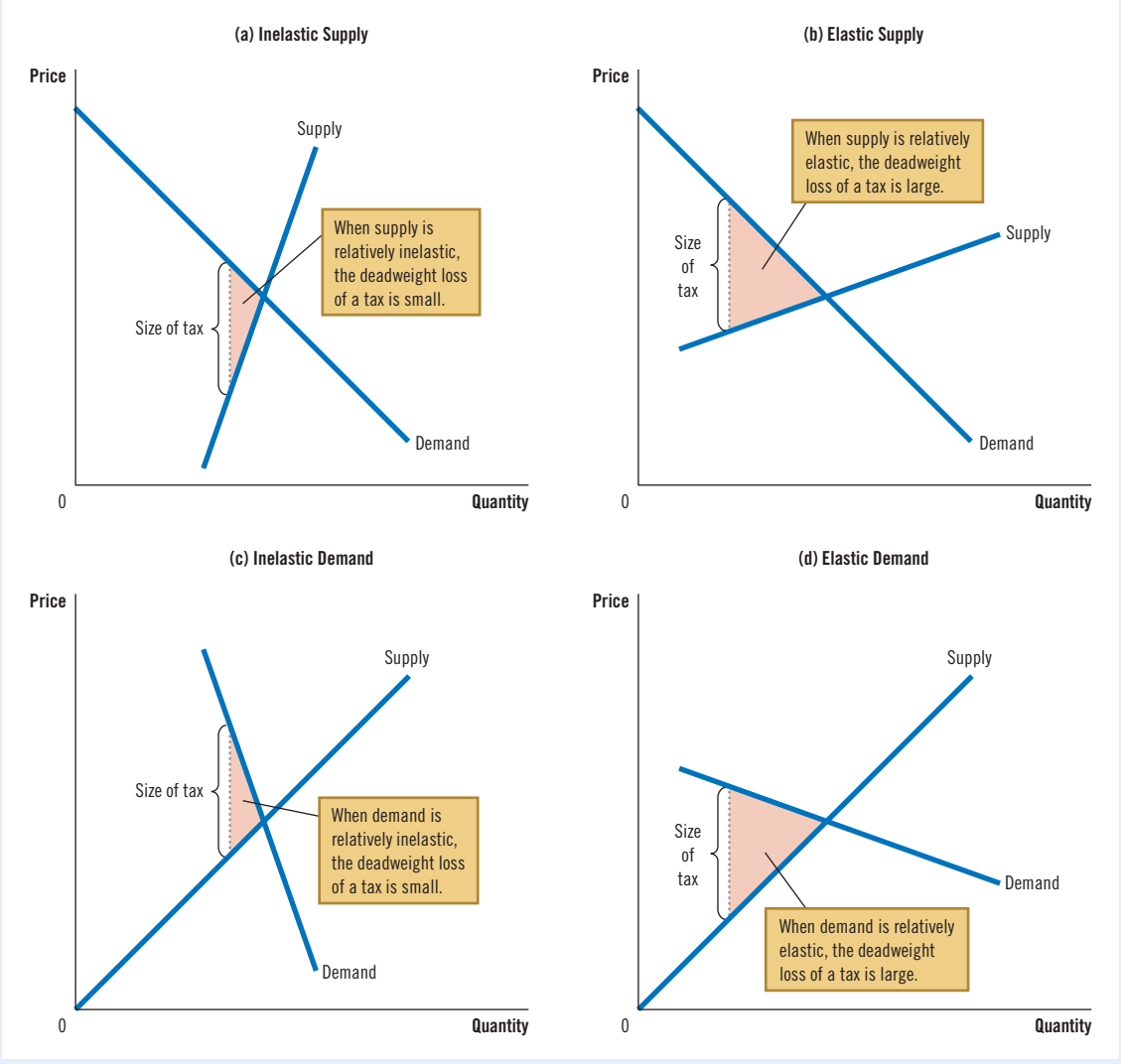
What determines whether the deadweight loss from a tax is large or small? The answer is to be found in the price elasticities of supply and demand, which measure how much the quantity supplied and quantity demanded respond to changes in the price.

Consider first how the elasticity of supply affects the size of the deadweight loss. In the top two panels of Figure 5, the demand curve and the size of the tax are the same. The only difference is the elasticity of the supply curve. In panel (a), the supply curve is relatively inelastic: The quantity supplied responds only slightly to changes in the price. In panel (b), the supply curve is relatively elastic: The quantity supplied responds substantially to changes in the price. Notice that the deadweight loss, the area of the triangle between the supply and demand curves, is larger when the supply curve is more elastic.

Figure 5

Tax Distortions and Elasticities

In panels (a) and (b), the demand curve and the size of the tax are the same, but the price elasticity of supply is different. Notice that the more elastic the supply curve, the larger the deadweight loss of the tax. In panels (c) and (d), the supply curve and the size of the tax are the same, but the price elasticity of demand is different. The more elastic the demand curve, the larger the deadweight loss of the tax.



Similarly, the bottom two panels of Figure 5 show how the elasticity of demand affects the size of the deadweight loss. Here, the supply curve and the size of the tax are held constant. In panel (c), the demand curve is relatively inelastic, and the deadweight loss is small. In panel (d), the demand curve is more elastic, and the deadweight loss from the tax is larger.

The lesson from this figure is apparent. A tax has a deadweight loss because it induces buyers and sellers to change their behavior. The tax raises the price paid by buyers, so they consume less. At the same time, the tax lowers the price received by sellers, so they produce less. Because of these changes in behavior, the equilibrium quantity in the market shrinks below the optimal quantity for economic efficiency. The more responsive buyers and sellers are to changes in the price, the more the equilibrium quantity shrinks. **Hence, the greater the elasticities of supply and demand, the larger the deadweight loss of a tax.**

The Deadweight Loss Debate

Supply, demand, elasticity, deadweight loss—all this economic theory is enough to make your head spin. But these ideas are part of a profound political question: How big should the government be? These concepts are important because the larger the deadweight loss of taxation, the larger the cost of government programs. If taxation entails large deadweight losses, then these losses are an argument for a leaner government that does less and taxes less. But if taxes impose small deadweight losses, then government programs are less costly than they otherwise might be, which in turn argues for a more expansive government. Of course, the phenomenon of deadweight loss is not the only reason to embrace large or small government: Another key part of the debate is the value of the government programs that would be funded with tax revenue.

So how big are the deadweight losses of taxation? Economists disagree on the answer. To see the nature of this disagreement, consider the most important tax in the U.S. economy: the tax on labor. The Social Security tax, the Medicare tax, and much of the federal income tax are labor taxes. Many state governments also tax labor earnings through state income taxes. A labor tax places a wedge between the wage that firms pay and the wage that workers receive. For a typical worker, if all forms of labor taxes are added together, the **marginal tax rate** on labor income—the tax on the last dollar of earnings—is about 40 percent.

The size of the labor tax is easy to determine, but calculating the deadweight loss of this tax is less straightforward. Economists disagree about whether this 40 percent labor tax has a small or a large deadweight loss. This disagreement arises because economists hold different views about the elasticity of labor supply.

Economists who say labor taxes do not greatly distort market outcomes hold that labor supply is fairly inelastic. Most people, they claim, would work full-time regardless of the wage. If so, the labor supply curve is almost vertical, and a tax on labor has a small deadweight loss. Some evidence suggests that this may be the case for workers in their prime working years who are the main breadwinners for their families.

Economists who say labor taxes are highly distortionary hold that labor supply is more elastic. While noting that some groups of workers may not change the quantity of labor they supply by very much in response to changes in labor taxes, these economists claim that other groups respond more to incentives. Here are some examples:

- Some people can adjust the number of hours they work—for instance, by working overtime. The higher their wage, the more hours they choose to work.
- Many families have second earners—say, married women with children—with some discretion over whether to do unpaid work at home or paid work in the marketplace. When deciding whether to take a job, these second earners compare the benefits of being at home (including savings on the cost of childcare) with the wages they could earn.
- Many people can choose when to retire, and their decisions are partly based on wages. Once they stop full-time work, wage levels determine their incentive to work part-time.
- Some people evade taxes by working at jobs that pay "under the table" or by engaging in illegal economic activity, such as the sale of prohibited drugs. Economists call this the **underground economy**. In deciding whether to work in the underground economy or at a legitimate job, potential criminals compare what they can earn by breaking the law with the wage they can earn legally.

MORRY GASH/AP PHOTO/BLOOMBERG/GETTY IMAGES

"What's your position on the elasticity of labor supply?"

In each of these cases, the quantity of labor supplied depends on the after-tax wage, so taxes on labor earnings affect people's decisions. These taxes encourage workers to work fewer hours, second earners to stay at home, the elderly to retire early, and the unscrupulous to enter the underground economy.

The debate over the distortionary effects of labor taxation persists to this day. Indeed, whenever two political candidates disagree about whether the government should provide more services or reduce the tax burden, part of the disagreement may rest on different views about the elasticity of labor supply and the deadweight loss of taxation. ●

QuickQuiz

4. If policymakers want to raise revenue by taxing goods while minimizing the deadweight losses, they should look for goods with _____ elasticities of demand and _____ elasticities of supply.
   a. small; small
   b. small; large
   c. large; small
   d. large; large

5. In the economy of Agricola, tenant farmers rent the land they use. If the supply of land is perfectly inelastic, then a tax on land would have _____ deadweight losses, and the burden of the tax would fall entirely on the _____.
   a. sizable; farmers
   b. sizable; landowners
   c. no; farmers
   d. no; landowners

6. Suppose the demand for grape jelly is perfectly elastic (because strawberry jelly is a good substitute), while the supply is unit elastic. A tax on grape jelly would have _____ deadweight losses, and the burden of the tax would fall entirely on the _____ of grape jelly.
   a. sizable; consumers
   b. sizable; producers
   c. no; consumers
   d. no; producers

Answers are at the end of the chapter.

8-3 Deadweight Loss and Tax Revenue as Taxes Vary

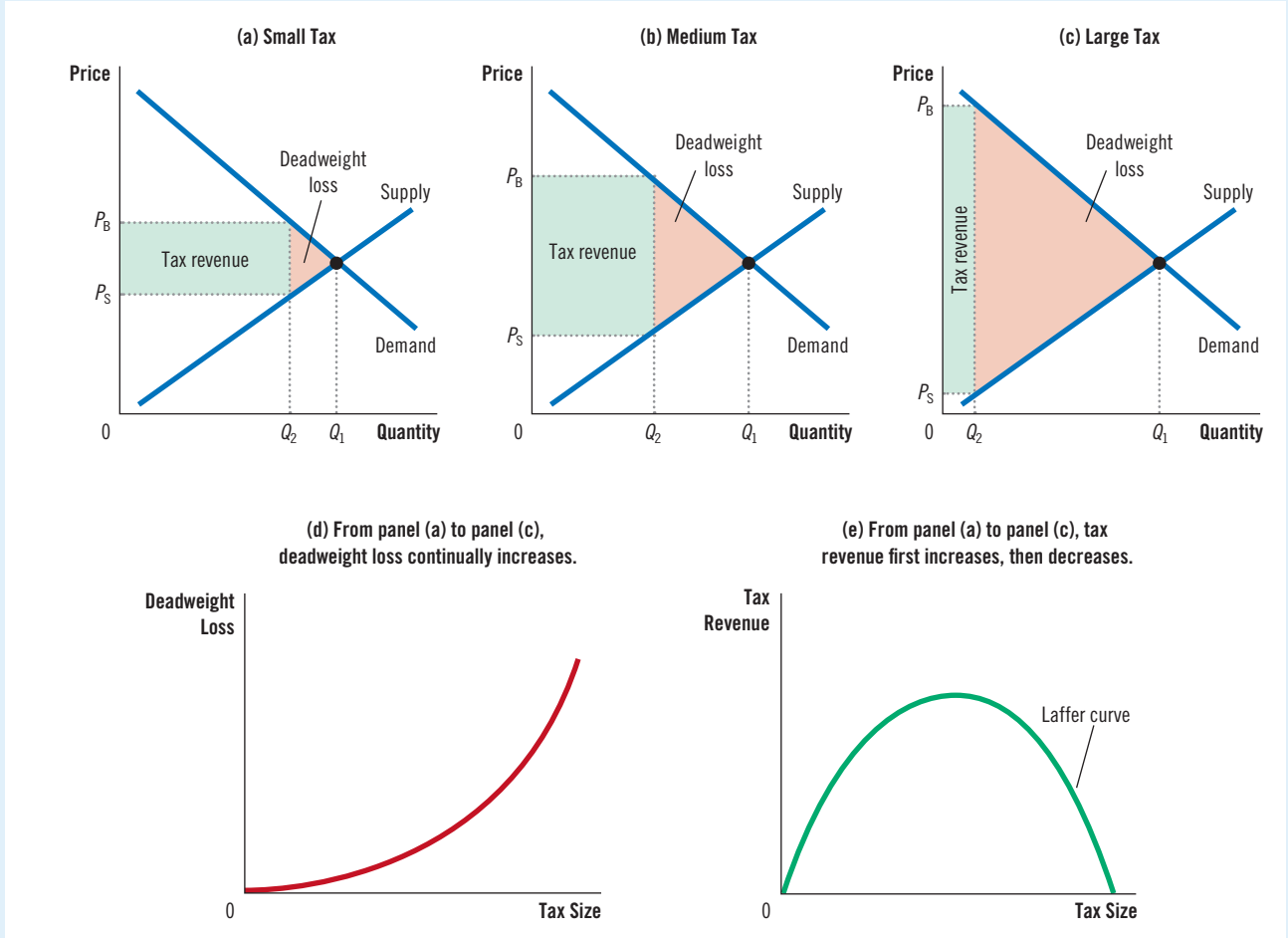
Taxes rarely stay the same for long. Policymakers are always considering raising one tax or lowering another. Let's consider what happens to the deadweight loss and tax revenue when the size of a tax changes.

Figure 6 shows the effects of a small, medium, and large tax, holding constant the market's supply and demand curves. The deadweight loss—the reduction

Figure 6

How Deadweight Loss and Tax Revenue Vary with the Size of a Tax

The deadweight loss is the reduction in total surplus resulting from the tax. Tax revenue is the size of the tax multiplied by the amount of the good sold. In panel (a), a small tax has a small deadweight loss and raises a small amount of revenue. In panel (b), a somewhat larger tax has a larger deadweight loss and raises more revenue. In panel (c), a very large tax has a very large deadweight loss, but because it reduces the size of the market so much, the tax raises only a small amount of revenue. Panels (d) and (e) summarize these conclusions. Panel (d) shows that as the size of a tax grows larger, the deadweight loss grows larger. Panel (e) shows that tax revenue first rises and then falls. This relationship is called the Laffer curve.



in total surplus that results when the tax reduces a market's size below the optimum—equals the area of the triangle between the supply and demand curves. For the small tax in panel (a), the area of the deadweight loss triangle is quite small. But as the size of the tax rises in panels (b) and (c), the deadweight loss grows larger and larger.

Indeed, the deadweight loss of a tax rises even more rapidly than the size of the tax. This occurs because the deadweight loss is the area of a triangle, and the area of a triangle depends on the **square** of its size. If we double the size of a tax, for instance, the base and height of the triangle double, so the deadweight loss rises by a factor of four. If we triple the size of a tax, the base and height triple, so the deadweight loss rises by a factor of nine.

The government's tax revenue is the size of the tax times the amount of the good sold. As the first three panels of Figure 6 show, tax revenue equals the area of the rectangle between the supply and demand curves. For the small tax in panel (a), tax revenue is small. As the size of the tax increases from panel (a) to panel (b), tax revenue grows. But as the size of the tax increases further from panel (b) to panel (c), tax revenue falls because the higher tax drastically reduces the size of the market. For a very large tax, no revenue would be raised because people would stop buying and selling the good altogether.

The last two panels of Figure 6 summarize these results. In panel (d), we see that as the size of a tax increases, its deadweight loss quickly gets larger. By contrast, panel (e) shows that tax revenue first rises with the size of the tax, but as the tax increases further, the market shrinks so much that tax revenue starts to fall.

The Laffer Curve and Supply-Side Economics

One day in 1974, the economist Arthur Laffer sat in a Washington restaurant with some prominent journalists and politicians. He took out a napkin and drew a figure on it to show how tax rates affect tax revenue. It looked much like panel (e) of Figure 6. Laffer then suggested that the United States was on the downward-sloping side of this curve. Tax rates were so high, he argued, that reducing them might actually increase tax revenue.

Most economists were skeptical of Laffer's suggestion. They accepted the idea that a cut in tax rates could increase tax revenue as a matter of theory, but they doubted whether it would do so in practice. There was scant evidence for Laffer's view that U.S. tax rates had, in fact, reached such extreme levels.

Nonetheless, the **Laffer curve** (as it became known) captured the imagination of Ronald Reagan. David Stockman, budget director in the first Reagan administration, offers the following story:

> [Reagan] had once been on the Laffer curve himself. "I came into the Big Money making pictures during World War II," he would always say. At that time, the wartime income surtax hit 90 percent. "You could only make four pictures, and then you were in the top bracket," he would continue. "So we all quit working after four pictures and went off to the country." High tax rates caused less work. Low tax rates caused more. His experience proved it.

When Reagan ran for president in 1980, he made cutting taxes part of his platform. Reagan argued that taxes were so high that they were discouraging hard work

and thereby depressing incomes. He argued that lower taxes would give people more incentive to work, which in turn would raise economic well-being. He suggested that incomes could rise by so much that tax revenue might increase, despite the lower tax rates. Because the cut in tax rates was intended to encourage people to increase the quantity of labor they supplied, the views of Laffer and Reagan became known as **supply-side economics**.

Economists continue to debate Laffer's argument. Many believe that subsequent history refuted Laffer's conjecture that lower tax rates would raise tax revenue. Yet because history is open to alternative interpretations, others view the events of the 1980s as more favorable to the supply siders. To evaluate Laffer's hypothesis definitively, we would need to rerun history without the Reagan tax cuts and see if tax revenues would have been higher or lower. But that experiment is impossible.

Some economists take an intermediate position. They believe that while an overall cut in tax rates normally reduces revenue, some taxpayers may occasionally find themselves on the wrong side of the Laffer curve. Other things being equal, a tax cut is more likely to raise tax revenue if the cut applies to those taxpayers facing the highest tax rates. In addition, Laffer's argument may be more compelling for countries with much higher tax rates than the United States. In Sweden in the early 1980s, for instance, the typical worker faced a marginal tax rate of about 80 percent. Such a high tax rate provides a substantial disincentive to work. Studies have suggested that Sweden would have indeed raised more tax revenue with lower tax rates.

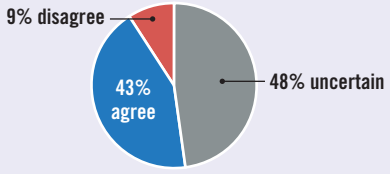
Economists disagree about these issues in part because there is no consensus about the size of the relevant elasticities. The more elastic supply and demand are in any market, the more taxes distort behavior, and the more likely it is that a tax cut will increase tax revenue. There is, however, agreement about the general lesson: How much revenue the government gains or loses from a tax change cannot be computed just by looking at tax rates. It also depends on how the tax change affects people's behavior.

An update to this story: Arthur Laffer rose to prominence again during the 2016 presidential campaign, when he was an adviser to Donald Trump. As recounted in his book with Stephen Moore, *Trumponomics*, he encouraged the candidate to propose a large tax cut. Laffer's argument was like the one he made years earlier: Why settle for the 2 percent growth that most economists were projecting? Wouldn't all our problems be easier to handle with a more rapidly expanding economy? The book quotes Trump as saying, when announcing his tax plan, that it would not increase the government's budget deficit (the shortfall of tax revenue from government spending) because it would raise growth rates to "3, or 4, 5, or even 6 percent." Most economists, however, were skeptical. And they were right to be. In the two years after the tax cut went into effect, the economy grew at 2.4 percent, and the budget deficit swelled. ●

Ask the Experts — The Laffer Curve

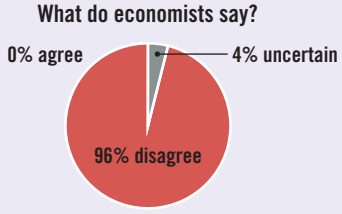
"A cut in federal income tax rates in the United States right now [2012] would lead to higher national income within five years than without the tax cut."

What do economists say?



"A cut in federal income tax rates in the United States right now would raise taxable income enough so that the annual total tax revenue would be higher within five years than without the tax cut."

What do economists say?



Source: IGM Economic Experts Panel, June 26, 2012.

QuickQuiz

7. The Laffer curve shows that, in some circumstances, the government can reduce a tax on a good and increase the
   a. price paid by consumers.
   b. equilibrium quantity.
   c. deadweight loss.
   d. government's tax revenue.
8. Eggs have a supply curve that is linear and upward-sloping and a demand curve that is linear and downward-sloping. If a 2 cent per egg tax is increased to 3 cents, the deadweight loss of the tax
   a. increases by less than 50 percent and may even decline.
   b. increases by exactly 50 percent.
   c. increases by more than 50 percent.
   d. The answer depends on whether supply or demand is more elastic.
9. Peanut butter has an upward-sloping supply curve and a downward-sloping demand curve. If a 10 cent per pound tax is increased to 15 cents, the government's tax revenue
   a. increases by less than 50 percent and may even decline.
   b. increases by exactly 50 percent.
   c. increases by more than 50 percent.
   d. The answer depends on whether supply or demand is more elastic.

Answers are at the end of the chapter.

8-4 Conclusion

This chapter applied the tools of welfare economics to better understand taxes. One of the **Ten Principles of Economics** in Chapter 1 is that markets are usually a good way to organize economic activity. Chapter 7 made this principle more precise using the concepts of consumer surplus, producer surplus, and market efficiency. Here, we have seen that when the government taxes a good, it makes the allocation of resources less efficient. Taxes are costly not only because they transfer resources from market participants to the government but also because they distort incentives and create deadweight losses.

The analysis here and in Chapter 6 sheds much light on the effects of taxes, but it is not the end of the story. Microeconomists study how best to design a tax system, including how to strike the right balance between efficiency and equity. And they consider how, when a market failure impedes efficiency, well-targeted taxes can sometimes fix the problem. Macroeconomists study how taxes influence the overall economy and how policymakers can use the tax system to stabilize economic activity and promote rapid growth. So as you study economics, don't be surprised when the subject of taxation comes up yet again.

Chapter in a Nutshell

- A tax on a good reduces the welfare of buyers and sellers of that good, and the reduction in consumer and producer surplus usually exceeds the revenue raised by the government. The fall in total surplus—the sum of consumer surplus, producer surplus, and tax revenue—is called the deadweight loss of the tax.
- Taxes have deadweight losses because they cause buyers to consume less and sellers to produce less, and these changes in behavior shrink the market below the level that maximizes total surplus. Because the elasticities of supply and demand measure how much buyers and sellers respond to market conditions, larger elasticities imply larger deadweight losses.
- As a tax grows larger, it distorts incentives more, and its deadweight loss grows larger. Because a tax reduces the size of the market, however, tax revenue does not continually increase. It first rises with the size of a tax, but if the tax gets large enough, tax revenue starts to fall.

Key Concept

deadweight loss, p. 157

Questions for Review

1. What happens to consumer surplus and producer surplus when the sale of a good is taxed? How does the change in consumer and producer surplus compare with the tax revenue? Explain.
2. Draw a supply-and-demand diagram with a tax on the sale of a good. Show the deadweight loss. Show the tax revenue.
3. How do the elasticities of supply and demand affect the deadweight loss of a tax? Why do they have this effect?
4. Why do experts disagree about whether labor taxes have small or large deadweight losses?
5. What happens to the deadweight loss and tax revenue when a tax is increased?

Problems and Applications

1. The market for pizza is characterized by a downward-sloping demand curve and an upward-sloping supply curve.
   a. Draw the competitive market equilibrium. Label the price, quantity, consumer surplus, and producer surplus. Is there any deadweight loss? Explain.
   b. Suppose that the government requires each pizzeria to pay a $1 tax on each pizza sold. Illustrate the effect of this tax on the pizza market, being sure to label consumer surplus, producer surplus, government revenue, and deadweight loss. How does each area compare to the pre-tax case?
   c. If the tax were removed, pizza eaters and sellers would be better off, but the government would lose tax revenue. Suppose that consumers and producers voluntarily transferred some of their gains to the government. Could all parties (including the government) be better off than they were with a tax? Explain using the labeled areas in your graph.
2. Evaluate the following two statements. Do you agree? Why or why not?
   a. "A tax that has no deadweight loss cannot raise any revenue for the government."
   b. "A tax that raises no revenue for the government cannot have any deadweight loss."
3. Consider the market for rubber bands.
   a. If this market has very elastic supply and very inelastic demand, how would the burden of a tax on rubber bands be shared between consumers and producers? Use the tools of consumer surplus and producer surplus in your answer.
   b. If this market has very inelastic supply and very elastic demand, how would the burden of a tax on rubber bands be shared between consumers and producers? Contrast your answer with your answer to part (a).
4. Suppose that the government imposes a tax on heating oil.
   a. Would the deadweight loss from this tax likely be greater in the first year after it is imposed or in the fifth year? Explain.
   b. Would the revenue collected from this tax likely be greater in the first year after it is imposed or in the fifth year? Explain.
5. After economics class, your friend suggests that taxing food would be a good way to raise revenue because the demand for food is quite inelastic. In what sense is taxing food a "good" way to raise revenue? In what sense is it not a "good" way to raise revenue?
6. Daniel Patrick Moynihan, the senator from New York from 1977 to 2001, once introduced a bill that would levy a 10,000 percent tax on certain hollow-tipped bullets.
   a. Do you expect that this tax would raise much revenue? Why or why not?
   b. Even if the tax would raise no revenue, why might Senator Moynihan have proposed it?
7. The government places a tax on the purchase of socks.
   a. Illustrate the effect of this tax on equilibrium price and quantity in the sock market. Identify the following areas both before and after the imposition of the tax: total spending by

consumers, total revenue for producers, and government tax revenue.

b. Does the price received by producers rise or fall? Can you tell whether total receipts for producers rise or fall? Explain.

c. Does the price paid by consumers rise or fall? Can you tell whether total spending by consumers rises or falls? Explain carefully. (Hint: Think about elasticity.) If total consumer spending falls, does consumer surplus rise? Explain.

8. This chapter analyzed the welfare effects of a tax on a good. Now consider the opposite policy. Suppose that the government **subsidizes** a good: For each unit of the good sold, the government pays $2 to the buyer. How does the subsidy affect consumer surplus, producer surplus, tax revenue, and total surplus? Does a subsidy lead to a deadweight loss? Explain.

9. Hotel rooms in Smalltown go for $100, and 1,000 rooms are rented on a typical day.

a. To raise revenue, the mayor decides to charge hotels a tax of $10 per rented room. After the tax is imposed, the going rate rises to $108, and the number of rooms rented falls to 900. Calculate the amount of revenue this tax raises for Smalltown and the deadweight loss of the tax. (Hint: The area of a triangle is ½ × base × height.)

b. The mayor now doubles the tax to $20. The price rises to $116, and the number of rooms rented falls to 800. Calculate the tax revenue and deadweight loss with this larger tax. Are they double, more than double, or less than double your answers in part (a)? Explain.

10. Suppose that a market is described by the following supply and demand equations:

$$Q^S = 2P$$
$$Q^D = 300 - P.$$

a. Solve for the equilibrium price and the equilibrium quantity.

b. Suppose that a tax of T is placed on buyers, so the new demand equation is:

$$Q^D = 300 - (P + T).$$

Solve for the new equilibrium. What happens to the price received by sellers, the price paid by buyers, and the quantity sold?

c. Tax revenue is $T \times Q$. Use your answer from part (b) to solve for tax revenue as a function of T. Graph this relationship for T between 0 and 300.

d. The deadweight loss of a tax is the area of the triangle between the supply and demand curves. Recalling that the area of a triangle is ½ × base × height, solve for deadweight loss as a function of T. Graph this relationship for T between 0 and 300. (Hint: If you look sideways, the base of the deadweight loss triangle is T, and the height is the difference between the quantity sold with the tax and the quantity sold without the tax.)

e. The government now levies a tax of $200 per unit on this good. Is this a good policy? Why or why not? Can you propose a better policy?

QuickQuiz Answers

1. **a** 2. **a** 3. **b** 4. **a** 5. **d** 6. **b** 7. **d** 8. **c** 9. **a**

Chapter

9

Application: International Trade

If you check the labels on the clothes you are wearing, you will likely find that many were made in another country. For most Americans a century ago, that would not have been the case: The textile and clothing industries were then a major part of the domestic economy. But as foreign factories started to produce quality goods at a lower cost, U.S. firms shut down domestic production and laid off workers. Today, most of the textiles and clothing that Americans consume are imported.

This shift in the textile and clothing industries raises important questions: How does international trade affect economic well-being? Who gains and who loses from trade among countries, and how do the gains compare with the losses?

Chapter 3 introduced the study of international trade by applying the principle of comparative advantage. This principle says that all countries can benefit from trade because it allows each country to specialize in what it does best. But the analysis in Chapter 3 was incomplete. It did not explain how the international marketplace achieves these gains from trade or how the gains are distributed among the various market participants.

SHIFT DRIVE/SHUTTERSTOCK.COM; ERDARYE+/GETTY IMAGES

We now tackle these questions using the tools developed over the past several chapters: supply, demand, equilibrium, consumer surplus, producer surplus, and so on. These tools can help explain how international trade affects economic well-being.

9-1 The Determinants of Trade

Consider the textile market. It is well suited to studying the gains and losses from international trade: Textiles are made and traded around the world, and policymakers often consider (and sometimes implement) trade restrictions to protect domestic textile producers from foreign competitors. Here, we examine the textile market in the imaginary country of Isoland.

9-1a The Equilibrium without Trade

As our story begins, the Isolandian textile market is cut off from the rest of the world. By government decree, no one in Isoland is allowed to import or export textiles, and the penalty for violating this prohibition is so large that no one dares try.

Because there is no international trade, the market for textiles in Isoland consists solely of domestic buyers and sellers. As Figure 1 shows, the domestic price adjusts to balance the quantity supplied by domestic sellers and the quantity demanded by domestic buyers. The figure shows the consumer and producer surplus in the equilibrium without trade. The sum of consumer and producer surplus measures the total benefits that buyers and sellers receive from the textile market.

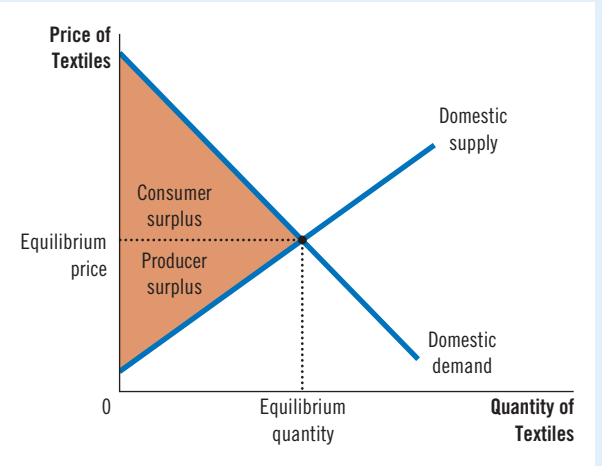
Now suppose that, in a political upset, Isoland elects Olivia Openminded as its new president. After campaigning on a platform of "change" and promising bold new ideas, President Openminded's first act is to assemble a team of economists to evaluate Isolandian trade policy. She asks them to report on three questions:

- If the government allows Isolandians to import and export textiles, what will happen to the price and quantity of textiles sold in the domestic market?
- Who will gain from free trade in textiles and who will lose, and will the gains exceed the losses?
- Should a tariff (a tax on textile imports) be part of the new trade policy?

Figure 1

The Equilibrium without International Trade

When an economy cannot trade in world markets, the price adjusts to balance domestic supply and demand. This figure shows consumer and producer surplus in an equilibrium without international trade for the textile market in Isoland.



After reviewing supply and demand in their favorite textbook (this one, of course), the Isolandian economics team begins its analysis.

9-1b The World Price and Comparative Advantage

The first issue the economists take up is whether Isoland is likely to become a textile importer or exporter. In other words, if free trade is allowed, will Isolandians end up buying or selling textiles in world markets?

To answer this question, the economists compare the current Isolandian price of textiles with the price prevailing in other countries, called the **world price**. If the world price of textiles exceeds the domestic price, then Isoland will export textiles once trade is permitted. Isolandian textile producers will be eager to receive the higher prices available abroad and will start selling their textiles to consumers in other countries. Conversely, if the world price is lower than the domestic price, Isoland will import textiles. Because foreign sellers offer a better price, Isolandian textile consumers will quickly start buying textiles from other countries.

world price
the price of a good that prevails in the world market for that good

In essence, comparing the world price and the domestic price without trade reveals whether Isoland has a comparative advantage in producing textiles. The domestic price reflects the opportunity cost: It tells us how much an Isolandian must give up to obtain one unit of textiles. If the domestic price is low, the cost of producing textiles in Isoland is low, suggesting that Isoland has a comparative advantage in producing textiles relative to the rest of the world. If the domestic price is high, then the cost of producing textiles in Isoland is high, suggesting that foreign countries have a comparative advantage in producing textiles.

As Chapter 3 discussed, trade among nations is based on comparative advantage. That is, trade is beneficial because it allows each nation to specialize in what it does best. By comparing the world price with the domestic price without trade, we can determine whether Isoland is better or worse than the rest of the world at producing textiles.

QuickQuiz

1. The country of Autarka does not allow international trade. In Autarka, you can buy a wool suit for 3 ounces of gold, while in neighboring countries, the same suit costs 2 ounces of gold. This suggests that
   a. Autarka has a comparative advantage in producing suits and would become a suit exporter if it opened up trade.
   b. Autarka has a comparative advantage in producing suits and would become a suit importer if it opened up trade.
   c. Autarka does not have a comparative advantage in producing suits and would become a suit exporter if it opened up trade.
   d. Autarka does not have a comparative advantage in producing suits and would become a suit importer if it opened up trade.
2. The nation of Openia allows free trade and exports steel. If steel exports were prohibited, the price of steel in Openia would be _____, benefiting steel _____.
   a. higher; consumers
   b. lower; consumers
   c. higher; producers
   d. lower; producers

Answers are at the end of the chapter.

9-2 The Winners and Losers from Trade

To analyze the welfare effects of trade, the Isolandian economists begin by assuming that Isoland is small compared with the rest of the world. This small-economy assumption means that Isoland's actions have a negligible effect on world markets.

Specifically, changes in Isoland's trade policy will not affect the world price of textiles. The Isolandians are said to be **price takers** in the world economy. That is, they take the price of textiles as given by the forces of supply and demand in the world market. Isoland can be an exporting country by selling textiles at the world price or an importing country by buying textiles at this price.

The small-economy assumption is not necessary to analyze the gains and losses from international trade, but the Isolandian economists know from experience (and from reading Chapter 2 of this book) that making simplifying assumptions is a key part of building a useful economic model. The assumption that Isoland is a small economy simplifies the analysis, and the basic lessons do not change in the more complicated case of a large economy.

9-2a The Gains and Losses of an Exporting Country

Figure 2 shows the Isolandian textile market when the domestic equilibrium price without trade is below the world price. Once trade is allowed, the domestic price rises to equal the world price. If market participants are all looking for the best price they can get, no seller of textiles would accept less than the world price, and no buyer would pay more than the world price.

After the domestic price has risen to equal the world price, the domestic quantity supplied differs from the domestic quantity demanded. The supply curve shows the quantity of textiles supplied by Isolandian sellers. The demand curve shows the

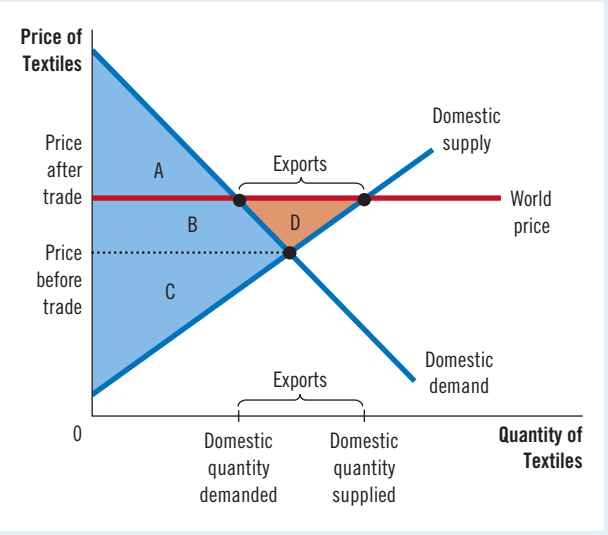
Figure 2

International Trade in an Exporting Country

Once trade is allowed, the domestic price rises to equal the world price. The supply curve shows the quantity of textiles produced domestically, and the demand curve shows the quantity consumed domestically. Exports from Isoland equal the difference between the domestic quantity supplied and the domestic quantity demanded at the world price. Sellers are better off (producer surplus rises from C to B + C + D), and buyers are worse off (consumer surplus falls from A + B to A). Total surplus rises by an amount equal to area D, indicating that trade raises the economic well-being of the country as a whole.

	Before Trade	After Trade	Change
Consumer Surplus	A + B	A	–B
Producer Surplus	C	B + C + D	+ (B + D)
Total Surplus	A + B + C	A + B + C + D	+ D

The area D shows the increase in total surplus and represents the gains from trade.



quantity of textiles demanded by Isolandian buyers. Because the domestic quantity supplied exceeds the domestic quantity demanded, Isoland sells textiles to other countries. This means that Isoland becomes a textile exporter.

Although the domestic quantity supplied and the domestic quantity demanded differ, the textile market is still in equilibrium because there is now another participant in the market: the rest of the world. One can view the horizontal line at the world price as representing the rest of the world's demand for textiles. This demand curve is perfectly elastic because Isoland, as a small economy, can sell as many textiles as it wants at the world price.

Consider the gains and losses from opening trade. Clearly, not everyone benefits. Trade forces the domestic price to rise to the world price. Domestic producers of textiles are better off because they can now sell textiles at a higher price, but domestic consumers of textiles are worse off because they now have to buy textiles at a higher price.

To measure these gains and losses, look at the changes in the consumer and producer surplus. Before trade is allowed, the price of textiles adjusts to balance supply and demand in the domestic market. Consumer surplus, the area between the demand curve and the before-trade price, is area A + B. Producer surplus, the area between the supply curve and the before-trade price, is area C. Total surplus before trade, which is the sum of consumer and producer surplus, is area A + B + C.

After trade is allowed, the domestic price rises to the world price. Consumer surplus shrinks to area A (the area between the demand curve and the world price). Producer surplus increases to area B + C + D (the area between the supply curve and the world price). Total surplus with trade is area A + B + C + D.

These welfare calculations show who wins and who loses from trade in an exporting country. Sellers benefit because producer surplus increases by area B + D. Buyers are worse off because consumer surplus decreases by area B. Because the gains of sellers exceed the losses of buyers by area D, total surplus in Isoland increases.

This analysis of an exporting country yields two conclusions:

- When a country allows trade and becomes an exporter of a good, domestic producers of the good are better off, and domestic consumers of the good are worse off.
- Trade raises the economic well-being of a nation in the sense that the gains of the winners exceed the losses of the losers.

9-2b The Gains and Losses of an Importing Country

Now, suppose that the domestic price before trade is above the world price. Once again, after trade is allowed, the domestic price becomes the world price. As Figure 3 shows, the domestic quantity supplied is less than the domestic quantity demanded. The difference between the domestic quantity demanded and the domestic quantity supplied is bought from other countries, and Isoland becomes a textile importer.

In this case, the horizontal line at the world price represents the supply of textiles from the rest of the world. This supply curve is perfectly elastic because Isoland is a small economy and can buy as much as it wants at the world price.

Consider the gains and losses from trade. Once again, not everyone benefits, but here, the winners and losers are reversed. When trade reduces the domestic price, domestic consumers are better off (they can buy textiles at a lower price), and domestic producers are worse off (they must sell at a lower price). Changes in consumer and producer surplus measure the size of the gains and losses. Before

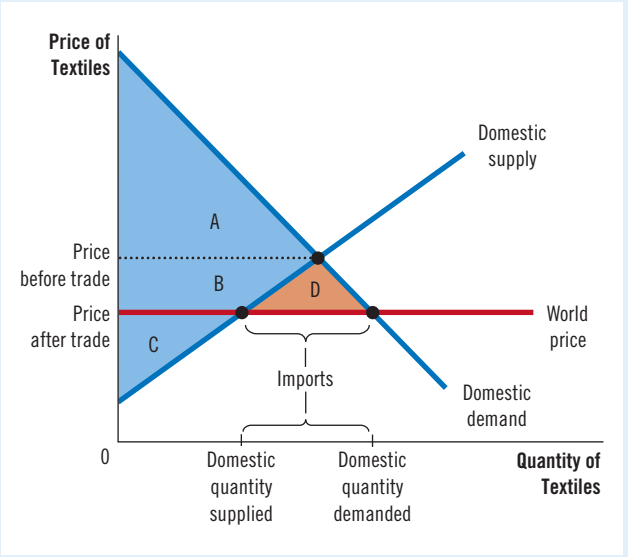
Figure 3

International Trade in an Importing Country

Once trade is allowed, the domestic price falls to equal the world price. The supply curve shows the amount produced domestically, and the demand curve shows the amount consumed domestically. Imports equal the difference between the domestic quantity demanded and the domestic quantity supplied at the world price. Buyers are better off (consumer surplus rises from A to A + B + D), and sellers are worse off (producer surplus falls from B + C to C). Total surplus rises by an amount equal to area D, indicating that trade raises the economic well-being of the country as a whole.

	Before Trade	After Trade	Change
Consumer Surplus	A	A + B + D	+ (B + D)
Producer Surplus	B + C	C	–B
Total Surplus	A + B + C	A + B + C + D	+D

The area D shows the increase in total surplus and represents the gains from trade.



trade, consumer surplus is area A, producer surplus is area B + C, and total surplus is area A + B + C. With trade, consumer surplus is area A + B + D, producer surplus is area C, and total surplus is area A + B + C + D.

These welfare calculations show who wins and who loses from trade in an importing country. Buyers benefit because consumer surplus increases by area B + D. Sellers are worse off because producer surplus falls by area B. The gains of buyers exceed the losses of sellers, and total surplus increases by area D.

This analysis of an importing country yields two conclusions parallel to those for an exporting country:

- When a country allows trade and becomes an importer of a good, domestic consumers of the good are better off, and domestic producers are worse off.
- Trade raises the economic well-being of a nation in the sense that the gains of the winners exceed the losses of the losers.

This analysis of trade helps explain, and qualify, one of the **Ten Principles of Economics** in Chapter 1: Trade can make everyone better off. What does this statement actually mean?

If Isoland opens its textile market to international trade, the change creates winners and losers, regardless of whether Isoland ends up exporting or importing

textiles. In either case, however, the gains of the winners exceed the losses of the losers, so the winners could compensate the losers and still be better off. In this sense, trade **can** make everyone better off. But **will** trade make everyone better off? Probably not. In practice, compensating the losers from international trade is rare. Without such compensation, opening an economy to international trade expands the size of the economic pie but can leave some people with a smaller slice.

This is why the debate over trade policy is often contentious. Whenever a policy creates winners and losers, the stage is set for a political battle. Nations sometimes limit trade because the losers from trade are better organized than the winners. The losers may turn their cohesiveness into political clout and lobby for trade restrictions such as tariffs or import quotas.

9-2c The Effects of a Tariff

tariff
a tax on goods produced abroad and sold domestically

The Isolandian economists next consider the effects of a **tariff**—a tax on imported goods. The economists quickly realize that a tariff on textiles will have no effect if Isoland becomes a textile exporter. That's because if no one in Isoland is interested in importing textiles, a tax on textile imports is irrelevant. The tariff matters only if Isoland becomes a textile importer. Concentrating their attention on this case, the economists compare welfare with and without the tariff.

Figure 4 shows the Isolandian market for textiles. With free trade, the domestic price equals the world price. A tariff raises the price of imported textiles above the world price by the amount of the tariff. Domestic suppliers, who compete with foreign suppliers of imported textiles, can now sell their goods for the world price plus the tariff. As a result, the price of textiles—both imported and domestic—rises by the amount of the tariff and is closer to the price without trade.

The change in price affects the behavior of domestic buyers and sellers. Because the tariff raises the price of textiles, it reduces the domestic quantity demanded from Q_1^D to Q_2^D and raises the domestic quantity supplied from Q_1^S to Q_2^S. **The tariff reduces the quantity of imports and moves the domestic market closer to its equilibrium without trade.**

Consider the gains and losses from the tariff. Because it raises the domestic price, domestic sellers are better off, and domestic buyers are worse off. In addition, the government gets the tariff revenue, which it can use for public purposes. These gains and losses are measured by the changes in consumer surplus, producer surplus, and government revenue, as in Chapter 8. These changes are summarized in the table in Figure 4.

Before the tariff, the domestic price equals the world price. Consumer surplus, the area between the demand curve and the world price, is area A + B + C + D + E + F. Producer surplus, the area between the supply curve and the world price, is area G. Government revenue equals zero. Total surplus, the sum of consumer surplus, producer surplus, and government revenue, is area A + B + C + D + E + F + G.

With a tariff, the domestic price exceeds the world price by the tariff's amount. Consumer surplus is now area A + B. Producer surplus is area C + G. Government revenue, which is the size of the tariff multiplied by the quantity of after-tariff imports, is area E. Total surplus with the tariff is area A + B + C + E + G.

To determine the total welfare effects of the tariff, add the change in consumer surplus (which is negative), the change in producer surplus (positive), and the change in government revenue (positive). The result is that total surplus decreases by area D + F. This fall in total surplus is the **deadweight loss** of the tariff.

A tariff causes a deadweight loss because, like most taxes, it distorts incentives and pushes the allocation of scarce resources away from the optimum. In this

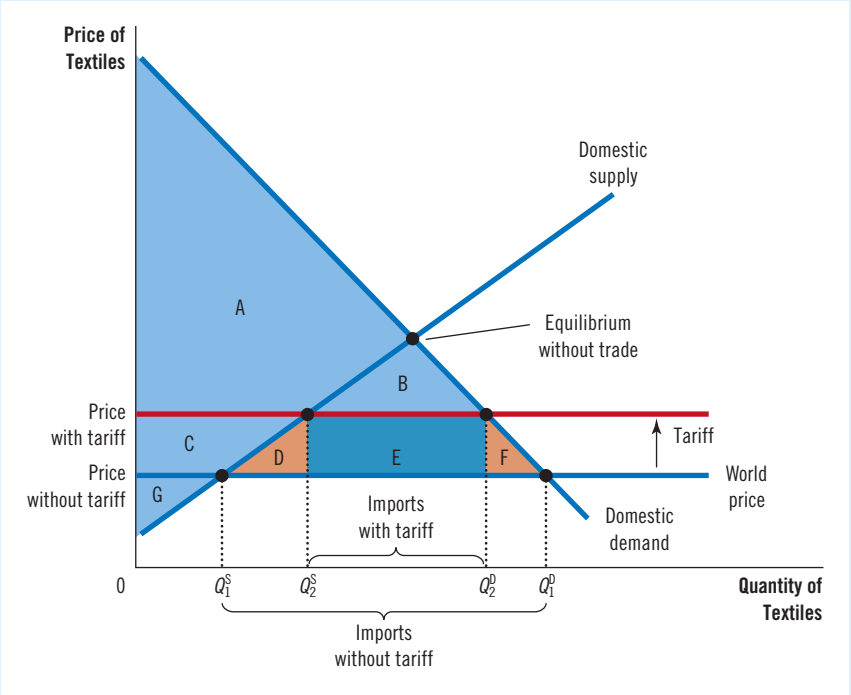
Figure 4

The Effects of a Tariff

A tariff, a tax on imports, reduces the quantity of imports and moves a market closer to the equilibrium that would exist without trade. Total surplus falls by an amount equal to area D + F. These two triangles represent the deadweight loss from the tariff.

	Before Tariff	After Tariff	Change
Consumer Surplus	A + B + C + D + E + F	A + B	−(C + D + E + F)
Producer Surplus	G	C + G	+C
Government Revenue	None	E	+E
Total Surplus	A + B + C + D + E + F + G	A + B + C + E + G	−(D + F)

The area D + F shows the fall in total surplus and represents the deadweight loss of the tariff.



case, note two effects. First, when the tariff raises the domestic price of textiles above the world price, it encourages domestic producers to increase production from Q_1^S to Q_2^S. Even though it costs more to make these units than to buy them at the world price, the tariff makes it profitable for domestic producers to manufacture them anyway. Second, when the tariff raises the price for domestic consumers, it leads them to reduce consumption from Q_1^D to Q_2^D. Even though domestic consumers are willing to pay more for these units than the world price, the tariff induces them to cut back on purchases. Area D represents the deadweight loss from the overproduction of textiles, and area F represents the deadweight loss from the underconsumption. The total deadweight loss is the sum of these two triangles.

FYI

Import Quotas: Another Way to Restrict Trade

Beyond tariffs, another way that nations sometimes restrict international trade is by putting limits on how much of a good can be imported. This book won't analyze such a policy other than to point out the conclusion: Import quotas are much like tariffs. Both tariffs and quotas reduce the quantity of imports, raise the domestic price of a good, decrease the welfare of domestic consumers, increase the welfare of domestic producers, and cause deadweight losses.

There is only one difference between these two types of trade restriction: A tariff raises revenue for the government, while an import quota generates surplus for those who obtain the permits to import. The profit for holders of import permits is the difference between the domestic price (at which they sell the imported good) and the world price (at which they buy it).

Tariffs and import quotas are even more similar if the government charges a fee for these permits. Suppose the government sets the permit fee equal to the difference between the domestic and world price. In this case, the entire profit of permit holders is paid to the government in permit fees, and the import quota works exactly like a tariff. Consumer surplus, producer surplus, and government revenue are precisely the same under the two policies.

In practice, however, countries that restrict trade with import quotas rarely do so by selling import permits. For example, the U.S. government has at times pressured Japan to "voluntarily" limit the sale of Japanese cars in the United States. In this case, the Japanese government allocates the import permits to Japanese firms, and these firms get the surplus from these permits. From the standpoint of U.S. welfare, this kind of import quota is worse than a U.S. tariff on imported cars. Both a tariff and an import quota raise prices, restrict trade, and cause deadweight losses, but at least the tariff generates revenue for the U.S. government rather than profit for foreign producers. ■

9-2d The Lessons for Trade Policy

The team of Isolandian economists writes to the new president:

Dear President Openminded,

You asked three questions about opening trade. After much hard work, we have the answers.

Question: If the government allows Isolandians to import and export textiles, what will happen to the price and quantity of textiles sold in the domestic market?

Answer: Once trade is allowed, the Isolandian price will move to equal the price prevailing around the world.

If the Isolandian price before trade is below the world price, our price will rise. The higher price will reduce the amount of textiles Isolandians consume and raise the amount Isolandians produce. Isoland will become a textile exporter because, in this case, it has a comparative advantage in producing textiles.

Conversely, if the Isolandian price before trade is above the world price, our price will fall. The lower price will increase the amount of textiles Isolandians consume and lower the amount we produce. In this case, Isoland will become a textile importer because other countries have a comparative advantage in producing textiles.

Question: Who will gain from free trade in textiles and who will lose, and will the gains exceed the losses?

Answer: The answer depends on whether the price rises or falls when trade is allowed. If the price rises, producers of textiles gain, and

consumers lose. If the price falls, consumers gain, and producers lose. But in both cases, the gains are larger than the losses, so free trade raises the total welfare of Isolandians.

Question: Should a tariff be part of the new trade policy?
Answer: A tariff has an impact only if Isoland becomes a textile importer. In this case, a tariff moves the economy closer to the no-trade equilibrium and, like most taxes, causes deadweight losses. A tariff improves the welfare of domestic producers and raises revenue for the government, but these gains are more than offset by the losses suffered by consumers. The best policy, from the standpoint of economic efficiency, would be to allow trade without a tariff.

Finally, you should bear in mind that moving to free trade in textiles will create losers as well as winners, even if the losses of the losers are smaller than the gains of the winners. So expect some of your constituents to oppose the policy. To blunt this opposition and spread the gains more equitably, you may want to make sure that the social safety net is generous enough to soften the blow for the losers.

We hope you find these answers helpful as you decide on your new policy.

Your faithful servants,
Isolandian economics team

9-2e Other Benefits of International Trade

The conclusions of the Isolandian economics team are based on the standard analysis of international trade. Their analysis uses the most fundamental tools in the economist's toolbox: supply, demand, and producer and consumer surplus. It shows that free trade entails winners and losers, but the gains of the winners exceed the losses of the losers.

The case for free trade can be made even stronger, however, because there are several other economic benefits of trade beyond those emphasized in the standard analysis. In a nutshell, here are some of them:

- **Increased variety of goods.** Goods produced in different countries are not exactly the same. German beer, for instance, is not the same as American beer. Free trade gives consumers in all countries a greater variety to choose from.
- **Lower costs through economies of scale.** Some goods can be produced at low cost only if they are produced in large quantities—a phenomenon called **economies of scale**. A firm cannot take full advantage of economies of scale if it can sell only in a small domestic market. Free trade gives firms access to world markets, allowing them to realize economies of scale more fully.
- **Increased competition.** A company shielded from foreign competitors is more likely to have market power, enabling it to raise prices above competitive levels. This is a type of market failure that hurts consumers and leads to inefficiencies. Opening trade fosters competition and gives the invisible hand a better chance to work its magic.
- **Increased productivity.** When a nation opens to international trade, the most productive firms expand their markets, while the least productive are forced out by increased competition. As resources move from the least to the most productive firms, overall productivity rises.

- **Enhanced flow of ideas.** The transfer of technological advances around the world is often linked to the exchange of the goods that embody those advances. The best way for a poor agricultural nation to quickly learn about the computer revolution, for instance, is to buy some computers from abroad rather than trying to make them from scratch.

In short, free trade increases variety for consumers, allows firms to take advantage of economies of scale, makes markets more competitive, makes the economy more productive, and fosters the spread of technology. If the Isolandian economists also took these benefits into account, their advice to the president would be even more forceful.

QuickQuiz

3. When the nation of Ectenia opens to world trade in coffee beans, the domestic price falls. Which of the following describes the situation?
   a. Domestic production of coffee rises, and Ectenia becomes a coffee importer.
   b. Domestic production of coffee rises, and Ectenia becomes a coffee exporter.
   c. Domestic production of coffee falls, and Ectenia becomes a coffee importer.
   d. Domestic production of coffee falls, and Ectenia becomes a coffee exporter.
4. When a nation opens to trade in a good and becomes an importer,
   a. producer surplus decreases, but consumer surplus and total surplus both increase.
   b. producer surplus decreases, but consumer surplus increases, so the impact on total surplus is ambiguous.
   c. producer surplus and total surplus increase, but consumer surplus decreases.
   d. producer surplus, consumer surplus, and total surplus all increase.
5. If a nation that imports a good imposes a tariff, it will increase
   a. the domestic quantity demanded.
   b. the domestic quantity supplied.
   c. the quantity imported from abroad.
   d. the efficiency of the equilibrium.
6. Which of the following policies would benefit producers, hurt consumers, and increase the amount of trade?
   a. the increase of a tariff in an importing country
   b. the reduction of a tariff in an importing country
   c. starting to allow trade when the world price is greater than the domestic price
   d. starting to allow trade when the world price is less than the domestic price

Answers are at the end of the chapter.

9-3 The Arguments for Restricting Trade

The letter from the Isolandian economics team starts to persuade President Openminded to consider allowing trade in textiles. She notes that the domestic price is now high compared with the world price. Free trade would, therefore, cause the price of textiles to fall and hurt domestic textile producers. Before implementing the new policy, she asks Isolandian textile companies to comment on the economists' advice.

Not surprisingly, the textile companies oppose free trade in textiles. They believe that the government should protect the domestic textile industry from foreign competition. Let's consider some of the arguments they might use to support their position and how the economics team would respond.

"You drive a Japanese car, drink French wine, eat Chinese food, own an American computer, buy Canadian lumber and vacation in Mexico. How can you be AGAINST free trade?!"

9-3a The Jobs Argument

Opponents of free trade often say it destroys jobs. For Isoland, for example, free trade in textiles could cause the price to fall, reducing the quantity produced domestically and slashing employment in the local textile industry. This job loss is an ugly picture and one that often occurs when trade expands.

Yet free trade creates new jobs even as it destroys some old ones. When Isolandians buy textiles from other countries, those countries get the resources to buy other goods from Isoland. Displaced Isolandian textile workers can move to industries in which Isoland has a comparative advantage. The transition may be difficult for some workers, especially in the short run. But the social safety net can assuage the hardship, and the country as a whole will enjoy a higher standard of living.

Opponents of trade are often skeptical that trade creates jobs. They might say that **everything** can be produced more cheaply abroad. With free trade, they might say, Isolandians could not be profitably employed in any industry. As Chapter 3 explains, however, the gains from trade are based on comparative advantage, not absolute advantage. This concept may be hard to grasp, but it is important. Even if one country is better than another country at producing everything, each country can gain from trading with the other. Workers will eventually find jobs in industries in which their country has a comparative advantage.

9-3b The National-Security Argument

When an industry is threatened with international competition, opponents of free trade often argue that the industry is vital to national security. For example, Isolandian steel companies might point out that steel is used to make guns and tanks. Free trade in steel could lead Isoland to become dependent on foreign countries to supply steel. If a war interrupted the foreign supply, Isoland might be unable to quickly produce enough steel and weapons to defend itself.

Economists acknowledge that protecting key industries may be appropriate when there are legitimate concerns over national security. Yet they know that this argument is often used too readily by producers eager to gain at consumers' expense.

Pay attention to who is making the national security argument. Companies have a financial incentive to exaggerate their role in national defense because protection from foreign competition can be lucrative. A nation's generals may see things differently. In fact, when the military buys an industry's output, it is a consumer and benefits from imports. Cheaper steel in Isoland, for example, would allow the Isolandian military to stockpile weapons at a lower cost.

9-3c The Infant-Industry Argument

New industries sometimes call for temporary trade restrictions to help them get started. After a period of protection, the argument goes, they will be strong enough to compete with foreign firms. Similarly, older industries sometimes say they need temporary protection until they adjust to new conditions.

Economists are often skeptical about such claims, largely because the infant-industry argument is hard to implement in practice. To apply the protection successfully, the government would need to determine which industries will eventually be profitable and decide whether the benefits of establishing them exceed the costs to consumers. Yet picking winners is extraordinarily difficult. It is made even more so by the political process, which often rewards industries with the most clout. And once a politically powerful industry is favored, the "temporary" policy can become permanent.

In addition, many economists question the infant-industry argument in principle. Suppose, for instance, that an industry is young and unable to compete against

foreign rivals, but there is reason to believe that it can be profitable in the long run. In this case, firm owners should be willing to incur temporary losses to obtain the eventual profits. Protection isn't needed. Start-up firms often incur temporary losses yet succeed in the long run, even without protection from competition.

9-3d The Unfair-Competition Argument

It is often said that free trade is desirable only if all countries play by the same rules. If companies in different countries are subject to different laws and regulations, then it is unfair (the argument goes) to expect them to compete globally. For instance, suppose Neighborland subsidizes its textile industry, lowering the costs of production for the country's textile companies. The Isolandian textile industry might argue that it should be protected from this foreign competition because Neighborland isn't competing fairly.

Would it, in fact, hurt Isoland to buy textiles from another country at a subsidized price? Yes, Isolandian textile producers would suffer, but Isolandian consumers would love the low price. The case for free trade remains the same: Consumer gains exceed producer losses. Neighborland's subsidy may be bad policy, but its taxpayers pay for it. Isoland is better off when it can buy textiles at a subsidized price. Rather than objecting to foreign subsidies, perhaps Isoland should send Neighborland a thank-you note.

9-3e The Protection-as-a-Bargaining-Chip Argument

Trade restrictions can be a bargaining chip, some politicians say. Even if free trade is desirable, they argue, trade restrictions can be useful in obtaining concessions from our trading partners. For example, Isoland might threaten to impose a tariff on textiles unless Neighborland removes its tariff on wheat. If Neighborland responds to this threat by removing its tariff, the result can be freer trade.

The problem with this strategy is that the threat may not work, leaving the country with two bad options. It can implement the trade restriction and shoot itself in the foot—reducing its own economic welfare is the way economists would put it. Or it can back down from its threat and lose prestige and future bargaining power. Why believe a tough-talking country that doesn't follow through with action? Faced with these choices, the country would probably wish that it had never made the threat in the first place.

Case Study

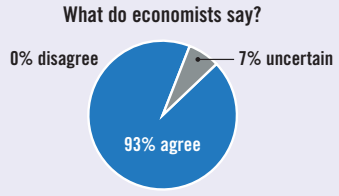
Trade Agreements and the World Trade Organization

A country can take one of two roads toward free trade. With a **unilateral** approach, it can remove its trade restrictions on its own. Great Britain did this in the 19th century, and Chile and South Korea did it more recently. Alternatively, a country can take a **multilateral** approach, reducing its trade restrictions in concert with other countries. In other words, it can bargain with its trading partners in an attempt to reduce trade restrictions around the world.

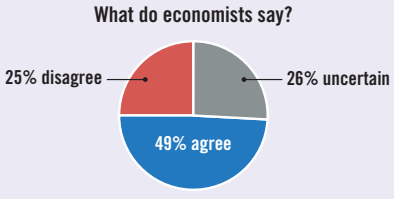
Ask the Experts

Trade Deals and Tariffs

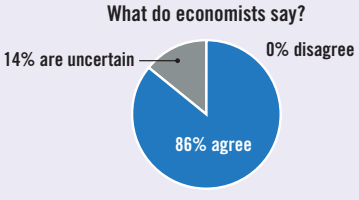
"Past major trade deals have benefited most Americans."



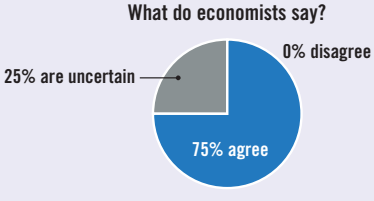
"Refusing to liberalize trade unless partner countries adopt new labor or environmental rules is a bad policy, because even if the new standards would reduce distortions on some dimensions, such a policy involves threatening to maintain large distortions in the form of restricted trade."



"The incidence of the latest round of U.S. import tariffs [in 2019] is likely to fall primarily on American households."



"The impact of the tariffs—and any Chinese countermeasures—on U.S. prices and employment is likely to be felt most heavily by lower income groups and regions."



Source: IGM Economic Experts Panel, November 11, 2014, March 27, 2013, and May 29, 2019.

One important example of the multilateral approach is the North American Free Trade Agreement (NAFTA), which in 1993 lowered trade barriers among the United States, Mexico, and Canada. An updated, new NAFTA, also known as the United States-Mexico-Canada Agreement, went into effect in 2020.

Another important multilateral pact is the General Agreement on Tariffs and Trade (GATT), a series of negotiations among many of the world's countries with the goal of promoting free trade. The United States helped to found GATT after World War II in response to the high tariffs imposed during the Great Depression of the 1930s. Many economists believe that those tariffs contributed to the worldwide economic hardship of that period. GATT successfully reduced the average tariff among member countries from more than 20 percent after World War II to less than 5 percent in 2000. And it continued at a low level for many years thereafter.

The rules established under GATT are enforced by an international institution called the World Trade Organization (WTO). The WTO was established in 1995 and has its headquarters in Geneva, Switzerland. As of 2021, 164 countries have joined the organization, accounting for 98 percent of world trade. The functions of the WTO are to administer trade agreements, provide a forum for negotiations, and handle disputes among member countries.

But the road to free trade is not a one-way street. The United States, which promoted the multilateral approach to free trade for decades, turned against it during the Trump administration and raised tariffs unilaterally in disputes with China, the European Union, and many other countries. The arguments were familiar and

In the News

Trade as a Tool for Economic Development

Free trade can help the world's poorest citizens.

Andy Warhol's Guide to Public Policy

By Arthur C. Brooks

I often ask people in my business—public policy—where they get their inspiration. Liberals often point to John F. Kennedy. Conservatives usually cite Ronald Reagan. Personally, I prefer the artist Andy Warhol, who famously declared, "I like boring things." He was referring to art, of course. But the sentiment provides solid public policy guidance as well.

Warhol's work exalted the everyday "boring" items that display the transcendental beauty of life itself. The canonical example is his famous paintings of Campbell Soup cans. Some people sneered, but those willing to look closely could see what he was doing. It is the same idea expressed in an old Zen saying, often attributed to the eighth-century Chinese Buddhist philosopher Layman Pang: "How wondrously supernatural and miraculous! I draw water and I carry wood!"

Warhol's critical insight is usually lost on most of the world. This is not because people are stupid, but because our brains are wired to filter out the mundane and focus on the novel. This turns out to be an important survival adaptation. To discern a predator, you must filter out the constant rustling of leaves and notice the strange snap of a twig.

Warhol believed that defeating this cognitive bias led to greater appreciation of beauty. It also leads to better public policy, especially in relieving poverty. For example, while our attention is naturally drawn to the latest fascinating and expensive innovations in tropical public health, many experts insist it is cheap, boring mosquito bed nets that best protect against malaria. Despite their lifesaving utility, these boring nets tend to be chronically underprovided.

We can look closer to home, too. People love to find ways to get fancy technology into poor schoolchildren's hands, but arguably the best way to help children falling behind in school is simply to devise ways to get them to show up.

But the very best example of the Warhol principle in policy is international trade. If it is progress against poverty that we're pursuing, trade beats the pants off every fancy development program ever devised. The simple mundane beauty

included many of the critiques of free trade discussed earlier. Some U.S. trading partners retaliated with higher tariffs of their own. For example, at the start of 2018, tariffs applied to very little trade between the United States and China. By the end of 2020, more than half of U.S. imports from China were subject to U.S. tariffs with an average rate of about 20 percent, and more than half of Chinese imports from the United States were subject to Chinese tariffs of about the same size. Where global tariffs will end up in the years ahead is far from certain.

What are the pros and cons of resuming the multilateral approach to free trade? One advantage is that it has the potential to result in freer trade than a unilateral approach because it can reduce trade restrictions abroad as well as at home. If international negotiations fail, however, the result could be more restricted trade than under a unilateral approach.

In addition, the multilateral approach may have a political advantage. In most markets, because producers are fewer and better organized than consumers, they have more political clout. Reducing the Isolandian tariff on textiles, for example, may be politically difficult if considered by itself. The textile companies would oppose free trade, and the buyers of textiles who would benefit are so numerous that organizing their support would be difficult. Yet suppose that Neighborland promises to reduce its tariff on wheat while Isoland reduces its tariff on textiles. In this case, the Isolandian wheat farmers, who are also politically powerful, would back the agreement. The multilateral approach can sometimes win political support when a unilateral approach cannot. ●

of making things and exchanging them freely is the best anti-poverty achievement in history.

For more than two decades, the global poverty rate has been decreasing by roughly 1 percent a year. To put this in perspective, that comes to about 70 million people—equivalent to the whole population of Turkey or Thailand—climbing out of poverty annually. Add it up, and around a billion people have escaped destitution since 1990.

Why? It isn't the United Nations or foreign aid. It is, in the words of the publication *YaleGlobal Online*, "High growth spillovers originating from large open emerging economies that utilize cross-border supply chains." For readers who don't have tenure, that means free trade in poor countries.

That mug in your hand that says "Made in China" is part of the reason that 680 million Chinese have been pulled out of absolute poverty since the 1980s. No giant collaboration among transnational technocrats or lending initiatives did that. It was because of economic reforms in China, of people making stuff, putting it on boats, and sending it to be sold in America—to you. Critics of free trade often argue that open economies lead to exploitation or environmental degradation. These are serious issues, but protectionism is never the answer. Curbing trade benefits entrenched domestic interests and works against the world's poor.

And what of claims that trade increases global income inequality? They are false. Economists at the World Bank and at LIS (formerly known as the Luxembourg Income Study Center) have shown that, for the world as a whole, income inequality has fallen for most of the past 20 years. This is chiefly because of rising incomes from globalization in the developing world….

Trade doesn't solve every problem, of course. The world needs democracy, security and many other expressions of American values and leadership as well. But in a policy world crowded with outlandish, wasteful boondoggles, free trade is just the kind of beautifully boring Warholian strategy we need. Americans dedicated to helping others ought to support it without compromise or apology. ■

Source: *New York Times*, April 12, 2015.

Questions to Discuss

1. What item that you use regularly was made in another country? What country did it come from? Who benefited from your purchase—you or the foreign producer?
2. How do you think trade between the United States and a poorer nation affects the workers in the poorer nation?

Mr. Brooks is a professor at Harvard University.

QuickQuiz

7. The nation of Lilliput imports rope from the nation of Brobdingnag, where rope producers are subsidized by the government because of their great political clout. The most efficient policy from the standpoint of Lilliput is to
   a. continue trading at the subsidized price.
   b. place a tariff on rope imports to offset the subsidy.
   c. give a similar subsidy to the rope producers of Lilliput.
   d. stop trading with Brobdingnag.

8. The goal of multilateral trade agreements is usually to
   a. equalize the level of tariffs across nations so no nation is disadvantaged relative to others.
   b. use targeted tariffs to ensure that nations produce those goods for which they have a comparative advantage.
   c. reduce tariffs in various nations simultaneously to blunt political pressure for protectionism.
   d. ensure that tariffs are used only to promote infant industries that will eventually become viable.

Answers are at the end of the chapter.

9-4 Conclusion

Economists and the public often hold different views about international trade. Public-opinion polls typically find that people are mixed on whether trade is an opportunity or a threat. Politicians often reflect this mixed verdict. By contrast, economists overwhelmingly support free trade. They view it as a way of allocating production efficiently and raising living standards both at home and abroad.

Economists view the fifty states of the United States as an ongoing experiment that confirms the virtues of free trade. Throughout its history, the United States has allowed unrestricted trade among the states, and the country has benefited from the specialization that trade allows. Florida grows oranges, Texas pumps oil, California makes wine, and so on. Americans would not enjoy the standard of living they do today if people could consume only those goods and services produced in their own states. The world could similarly benefit from free trade among countries.

To better understand economists' view of trade, return to Isoland. Suppose that President Openminded, after reading the latest poll results, ignores the advice of her economics team and decides not to allow free trade in textiles. The country remains in the equilibrium without international trade.

Then, one day, an Isolandian named Isabel Inventor discovers a way to make textiles at a very low cost. The process is secret, however. What is odd is that Inventor doesn't need traditional inputs such as cotton or wool. The only material input she needs is wheat. And even more oddly, to manufacture textiles from wheat, she hardly needs any labor input at all.

Inventor is hailed as a genius. Because everyone buys clothing, the lower cost of textiles gives all Isolandians a higher standard of living. Workers who had previously produced textiles experience some hardship when their factories close, but eventually, they find work in other industries. Some become farmers and grow the wheat that Inventor turns into textiles. Others enter new industries that emerge as a result of higher Isolandian living standards. Everyone understands that the displacement of workers in outmoded industries is an inevitable part of technological progress and economic growth.

After several years, a journalist named Roberto Reporter decides to investigate this mysterious new textile process. He sneaks into Inventor's factory and learns that she has not been making textiles at all. Instead, she has been smuggling wheat abroad in exchange for textiles from other countries. The only thing that Inventor had discovered was the gains from international trade.

When Reporter reveals the truth, the government shuts down Inventor's operation. The price of textiles rises, and workers return to jobs in textile factories. Living standards in Isoland fall back to their former levels. Inventor is jailed and held up to public ridicule. After all, she was no inventor. She was just an economist.

Chapter in a Nutshell

- The effects of free trade can be determined by comparing the domestic price before trade with the world price. A low domestic price indicates that the country has a comparative advantage in producing the good and that the country will become an exporter. A high domestic price indicates that the rest of the world has a comparative advantage in producing the good and that the country will become an importer.
- When a country allows trade and becomes an exporter of a good, producers of the good are better off, and consumers of the good are worse off. When a country allows trade and becomes an importer of a good, consumers are better off, and producers are worse off. In both cases, the gains from trade exceed the losses.
- A tariff—a tax on imports—moves a market closer to the equilibrium that would exist without trade and reduces the gains from trade. Domestic producers are better off, and the government raises revenue, but the losses to consumers exceed these gains.
- International trade yields several benefits beyond those based on comparative advantage: greater product variety for consumers, more opportunity for firms to take advantage of economies of scale, increased market competition, higher overall productivity as more productive firms expand and less productive ones contract, and improved access to state-of-the-art technology.
- There are various arguments for restricting trade: protecting jobs, defending national security, helping infant industries, preventing unfair competition, and responding to foreign trade restrictions. Although some of these arguments have merit in some cases, most economists believe that free trade is usually the better policy.

Key Concepts

world price, p. 171

tariff, p. 175

Questions for Review

1. What does the domestic price that prevails without international trade tell us about a nation's comparative advantage?
2. When does a country become an exporter of a good? An importer?
3. Draw the supply-and-demand diagram for an importing country. Identify consumer surplus and producer surplus before trade is allowed. Identify consumer surplus and producer surplus with free trade. What is the change in total surplus?
4. Describe what a tariff is and its economic effects.
5. List five arguments often given to support trade restrictions. How do economists respond to these arguments?
6. What is the difference between the unilateral and multilateral approaches to achieving free trade? Give an example of each.

Problems and Applications

1. The world price of wine is below the price that would prevail in Canada in the absence of trade.
   a. Assuming that Canadian imports of wine are a small share of total world wine production, draw a graph for the Canadian market for wine under free trade. Identify consumer surplus, producer surplus, and total surplus in an appropriate table.
   b. Now suppose that a shift of the Gulf Stream leads to an unseasonably cold summer in Europe, destroying much of the grape harvest there. What effect does this shock have on the world price of wine? Using your graph and table from part (a), show the effect on consumer surplus, producer surplus, and total surplus in Canada. Who are the winners and losers? Is Canada as a whole better or worse off?
2. Suppose that Congress imposes a tariff on imported automobiles to protect the U.S. auto industry from foreign competition. Assuming that the United States is a price taker in the world auto market, show the following on a diagram: the change in the quantity of imports, the loss to U.S. consumers, the gain to U.S. manufacturers, government revenue, and the deadweight loss associated with the tariff. The loss to consumers can be decomposed into three pieces: a gain to domestic producers, revenue for the government, and a deadweight loss. Use your diagram to identify these three pieces.
3. When China's clothing industry expands, the increase in the world supply lowers the world price of clothing.
   a. Draw an appropriate diagram to analyze how this change in price affects consumer surplus, producer surplus, and total surplus in a nation that imports clothing, such as the United States.
   b. Now, draw an appropriate diagram to show how this change in price affects consumer surplus, producer surplus, and total surplus in a nation that exports clothing, such as the Dominican Republic.
   c. Compare your answers to parts (a) and (b). What are the similarities, and what are the differences? Which country should be concerned about the expansion of the Chinese textile industry? Which country should be applauding it? Explain.
4. Consider the arguments for restricting trade.
   a. Imagine that you are a lobbyist for timber, an established industry suffering from low-priced foreign competition, and you are trying to get Congress to pass trade restrictions. Which two or three of the five arguments discussed in the chapter do you think would be most persuasive to the average member of Congress? Explain your reasoning.
   b. Now, assume you are an astute student of economics (not a hard assumption, we hope). Although all the arguments for restricting trade have their shortcomings, name the two or three arguments that seem to make the most economic sense to you. For each, describe the economic rationale for and against these arguments for trade restrictions.
5. The nation of Textilia does not allow imports of clothing. In its equilibrium without trade, a T-shirt costs $20, and the equilibrium quantity is 3 million T-shirts. After reading Adam Smith's *The Wealth of Nations* while on vacation, the president decides to open the Textilian market to international trade. The market price of a T-shirt falls to the world price of $16. The number of T-shirts consumed in Textilia rises to 4 million, while the number of T-shirts produced declines to 1 million.
   a. Illustrate the situation just described in a graph. Your graph should show all the numbers.
   b. Calculate the change in consumer surplus, producer surplus, and total surplus that results from opening trade. (Hint: Recall that the area of a triangle is ½ × base × height.)
6. China is a major producer of grains, such as wheat, corn, and rice. Some years ago, the Chinese government, concerned that grain exports were driving up food prices for domestic consumers, imposed a tax on grain exports.
   a. Draw the graph that describes the market for grain in an exporting country. Use this graph as the starting point to answer the following questions.
   b. How does an export tax affect domestic grain prices?
   c. How does it affect the welfare of domestic consumers, the welfare of domestic producers, and government revenue?
   d. What happens to total welfare in China, as measured by the sum of consumer surplus, producer surplus, and tax revenue?
7. Consider a country that imports a good from abroad. For each of the following statements, state whether it is true or false. Explain your answer.
   a. "The greater the elasticity of demand, the greater the gains from trade."
   b. "If demand is perfectly inelastic, there are no gains from trade."
   c. "If demand is perfectly inelastic, consumers do not benefit from trade."

8. Having rejected a tariff on textiles (a tax on imports), the president of Isoland is now considering the same-sized tax on textile consumption (including both imported and domestically produced textiles).
   a. Using Figure 4, identify the quantity consumed and the quantity produced in Isoland under a textile consumption tax.
   b. Construct a table similar to that in Figure 4 for the textile consumption tax.
   c. Which raises more revenue for the government—the consumption tax or the tariff? Which has a smaller deadweight loss? Explain.
9. Assume the United States is an importer of televisions and there are no trade restrictions. U.S. consumers buy 1 million televisions per year, of which 400,000 are produced domestically and 600,000 are imported.
   a. Suppose that a technological advance among Chinese television manufacturers causes the world price of televisions to fall by $100. Draw a graph to show how this change affects the welfare of U.S. consumers and U.S. producers and how it affects total surplus in the United States.
   b. After the fall in price, consumers buy 1.2 million televisions, of which 200,000 are produced domestically, and 1 million are imported. Calculate the change in consumer surplus, producer surplus, and total surplus from the price reduction.
   c. If the government responded by putting a $100 tariff on imported televisions, what would this do? Calculate the revenue that would be raised and the deadweight loss. Would it be a good policy from the standpoint of U.S. welfare? Who might support the policy?
   d. Suppose that the fall in price is attributable not to a technological advance but to a subsidy from the Chinese government to Chinese industry of $100 per television. How would this affect your analysis?
10. Consider a small country that exports steel. Suppose that a "pro-trade" government decides to subsidize the export of steel by paying a certain amount for each ton sold abroad. How does this export subsidy affect the domestic price of steel, the quantity of steel produced, the quantity of steel consumed, and the quantity of steel exported? How does it affect consumer surplus, producer surplus, government revenue, and total surplus? Is it a good policy from the standpoint of economic efficiency? (Hint: The analysis of an export subsidy is similar to the analysis of a tariff.)

QuickQuiz Answers

1. **d** 2. **b** 3. **c** 4. **a** 5. **b** 6. **c** 7. **a** 8. **c**

Chapter

10

Externalities

Firms that make and sell paper also create, as a by-product of the manufacturing process, a chemical called dioxin. Scientists say that once dioxin enters the environment, it can cause cancer, birth defects, and other health problems. This chapter considers how the release of pollutants such as dioxin fits into the analysis of markets.

One of the **Ten Principles of Economics** in Chapter 1 is that markets are usually a good way to organize economic activity. Chapters 4 through 9 explained this principle. They examined how competitive markets allocate scarce resources and showed that the equilibrium of supply and demand typically allocates resources efficiently. To use Adam Smith's metaphor, an "invisible hand" leads buyers and sellers to maximize the total benefits received by market participants.

But unfettered markets are not a cure-all. The invisible hand on its own does not prevent firms in the paper market from emitting too much dioxin. Society needs to find some other way to solve this problem. This leads us to another of the **Ten Principles of Economics**: Governments can sometimes improve market outcomes.

AZMANJAKA/E+/GETTY IMAGES; CARLOS FLORES/EYEEM/GETTY IMAGES

externality
the uncompensated impact of a person's actions on the well-being of a bystander

The market failures examined here fall under a general category called **externalities**. An **externality** arises when someone engages in an action that influences a bystander's well-being and when no compensation is paid for that effect. If the impact on the bystander is adverse, it is a **negative externality**. If it is beneficial, it is a **positive externality**.

When externalities are present, society's interest in a market outcome extends beyond the well-being of buyers and sellers in the market and includes the well-being of bystanders. Because buyers and sellers do not take into account the external effects of their actions when deciding how much to demand or supply, the market equilibrium is not efficient. In other words, societal well-being is not maximized, and government policies can potentially correct the market failure.

The release of dioxin into the environment, for instance, is a negative externality—and a grave one. Without government intervention, the producers of paper may not consider the full cost of the pollution they create, and the consumers of paper may not consider the full cost of the pollution they contribute to through their purchases. The market will allow too much dioxin to be emitted unless the government intervenes with well-designed policies. And indeed, the Environmental Protection Agency monitors dioxin levels and regulates its release.

Externalities come in many forms, as do the policy responses that try to deal with them. Here are some examples:

- The exhaust from automobiles is a negative externality because it creates smog that other people breathe and contributes to global climate change. Because drivers may ignore this externality when deciding what cars to buy and how much to use them, they tend to pollute too much. The federal government addresses this problem by setting emission standards for cars. It also taxes gasoline, subsidizes mass transit, and builds bicycle paths to reduce the amount that people drive.
- Restored historic buildings confer a positive externality because people who see them enjoy their beauty and the sense of history they convey. But building owners, who do not capture the benefits experienced by passers-by, tend to tear down older buildings too quickly. Recognizing this problem, many local governments regulate the destruction of historic buildings and provide tax breaks to owners who restore them.
- Barking dogs create a negative externality when neighbors are annoyed by their noise. Dog owners don't bear the full burden of the noise and tend to take too few precautions to prevent it. Local governments address this problem by making it illegal to disturb the peace.
- Research into new technologies provides a positive externality because it creates knowledge that others can use. If individual inventors, firms, and universities cannot capture the benefits of their inventions, they may devote too few resources to research. The federal government addresses this problem partially through the patent system, which gives inventors exclusive use of their inventions for a limited period.
- During the peak of a pandemic, busy restaurants create a negative externality by providing an opportunity for the disease to spread. Restaurant owners may ignore the health impact of their businesses on society. The government can address the problem by requiring restaurants to temporarily switch to take-out only (and perhaps compensating the restaurant owners and workers).

In all these cases, people may fail to take into account the external effects of their behavior. The government responds by trying to influence their behavior to protect the interests of bystanders.

10-1 Externalities and Market Inefficiency

In this section, we use the tools of welfare economics developed in Chapter 7 to examine how externalities affect economic well-being. The analysis shows why externalities cause markets to allocate resources inefficiently. Later sections examine the ways private individuals and public policymakers can remedy this type of market failure.

10-1a Welfare Economics: A Recap

Let's first review the lessons of welfare economics from Chapter 7, using the market for steel as an example. Figure 1 shows the supply and demand curves in the steel market.

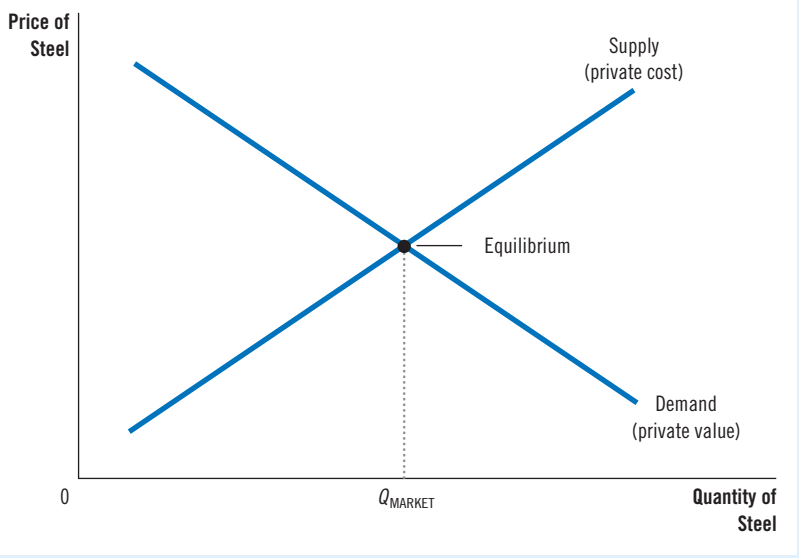
Chapter 7 showed that the supply and demand curves contain important information about costs and benefits. The demand curve reflects the value of steel to consumers, as measured by what they are willing to pay. At any quantity, the height of the demand curve shows the willingness to pay of the marginal buyer. In other words, it shows the value to the consumer of the last unit of steel bought. Similarly, the supply curve reflects the costs of producing steel. At any quantity, the height of the supply curve shows the cost to the marginal seller. In other words, it shows the cost to the producer of the last unit of steel sold.

The price adjusts to balance the supply and demand for steel. The quantity produced and consumed in the market equilibrium, shown as Q_{MARKET} in Figure 1, is efficient in the sense that it maximizes the sum of producer and consumer surplus. That is, the market allocates resources in a way that maximizes the total value to the consumers who buy and use steel minus the total costs to the producers who make and sell steel.

Figure 1

The Market for Steel

The demand curve reflects the value to buyers, and the supply curve reflects the costs of sellers. The equilibrium quantity, Q_{MARKET}, maximizes the total value to buyers minus the total costs of sellers. In the absence of externalities, the market equilibrium is efficient.



J.B. HANDELSMAN/THE NEW YORKER COLLECTION/THE CARTOON BANK

"All I can say is that if being a leading manufacturer means being a leading polluter, so be it."

10-1b Negative Externalities

Now suppose that steel factories emit pollution: For each unit of steel produced, a certain amount of smoke enters the atmosphere. Because these emissions create health and climate risks, they are a negative externality. How does this externality affect market efficiency?

Because of this externality, the cost to society of producing steel exceeds the cost to steel producers. For each unit of steel produced, the **social cost** equals the direct, private costs of the steel producers plus the costs to bystanders harmed by the emissions. Figure 2 shows the social cost of producing steel. The social-cost curve is above the supply curve because it takes into account the external costs imposed on society by steel production. The difference between these two curves reflects the cost of the pollution emitted.

What quantity of steel should be produced? Consider what a committee of benevolent social planners would do. The planners want to maximize the total surplus derived from the market—the value to consumers of steel minus the cost of producing it. They take a broad perspective, however, and understand that the cost of producing steel includes the external costs of pollution.

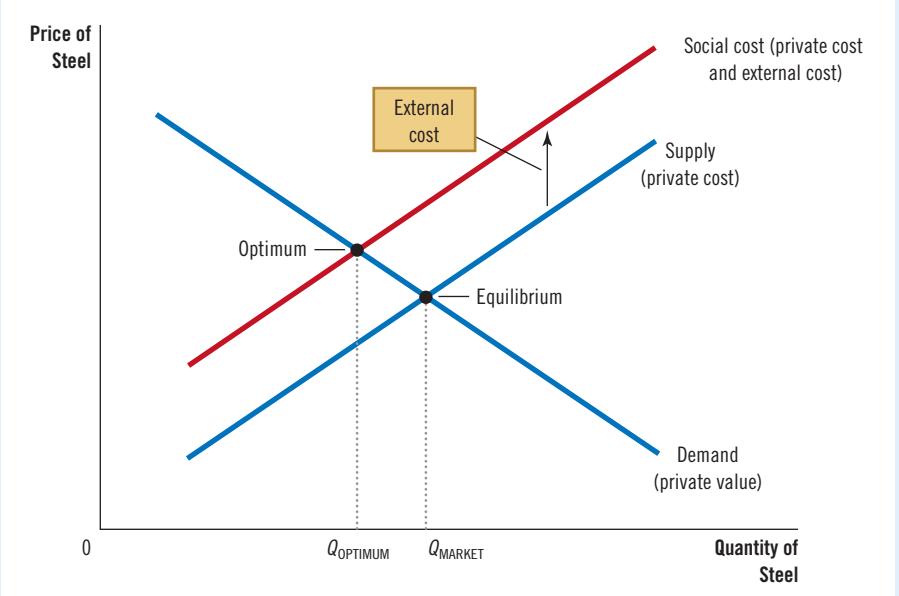
The planners would choose the production level at which the demand curve crosses the social-cost curve. This intersection determines the optimal amount of steel from the standpoint of society as a whole. Below this level of production, the value of the steel to consumers (measured by the height of the demand curve) exceeds the social cost of producing it (measured by the height of the social-cost curve). Above this level, the social cost of producing additional steel exceeds the value to consumers.

Note that the equilibrium quantity of steel, Q_{MARKET}, is larger than the socially optimal quantity, $Q_{OPTIMUM}$. This inefficiency occurs because the market equilibrium reflects only the private costs of production. In the market equilibrium, the

Figure 2

Pollution and the Social Optimum

In the presence of a negative externality, such as pollution, the social cost of the good exceeds the private cost. The optimal quantity, $Q_{OPTIMUM}$, is therefore smaller than the equilibrium quantity, Q_{MARKET}.



marginal consumer values steel at less than the social cost of producing it. That is, at Q_{MARKET}, the demand curve lies below the social-cost curve. Reducing steel production and consumption below the market equilibrium level raises total economic well-being.

How can the social planners achieve the optimal outcome? One way is to tax steel producers for each ton of steel sold. The tax would shift the supply curve upward by the size of the tax. If the tax accurately reflected the external cost of pollutants released into the atmosphere, the new supply curve would coincide with the social-cost curve. In the new market equilibrium, steel producers would produce the socially optimal quantity of steel.

Economists say that such a tax **internalizes the externality** because it gives buyers and sellers in the market an incentive to consider the external effects of their actions. Steel producers would, in essence, take the costs of pollution into account when deciding how much steel to supply because the tax would make them pay for these external costs. And, because the market price would reflect the tax on producers, consumers of steel would have an incentive to buy less. The policy is based on one of the **Ten Principles of Economics**: People respond to incentives. Later in this chapter, we discuss more fully how policymakers can deal with externalities.

internalizing the externality
altering incentives so that people take into account the external effects of their actions

10-1c Positive Externalities

Although some activities impose costs on third parties, others yield benefits. Consider education. To a large extent, the benefit of education is private. Put in strictly monetary terms, the consumers of education become more productive workers and reap much of the benefit in the form of higher wages. In addition, education helps people become more well-rounded by broadening their horizons. Beyond these private benefits, however, education also yields positive externalities. One is that a more educated population leads to more informed voters, which means better government for everyone. Another is that a more educated population tends to result in lower crime rates. A third is that a more educated population may encourage the development and dissemination of technological advances, leading to higher productivity and wages for everyone. Given these positive externalities, people may prefer to have neighbors who are well educated.

The analysis of positive externalities is similar to the analysis of negative externalities. As Figure 3 shows, the demand curve does not reflect the value to society of the good. Because the social value exceeds the private value, the social-value curve lies above the demand curve. The optimal quantity is found where the social-value curve and the supply curve intersect. The socially optimal quantity exceeds the quantity that the private market would reach on its own. In other words, without outside intervention, the market would yield too little education.

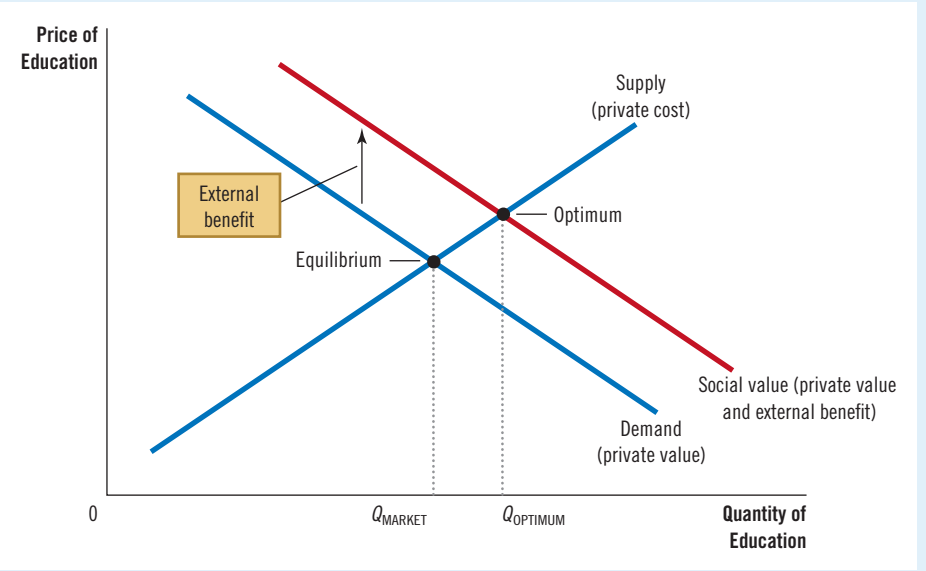
Once again, the government can correct the market failure by inducing market participants to internalize the externality. The appropriate policy to deal with positive externalities is the opposite of the policy for negative externalities. To move the market equilibrium closer to the social optimum, a positive externality requires a subsidy. In fact, that is the policy the government follows: Education is heavily subsidized through public schools, government scholarships, and tax breaks.

To summarize: **Negative externalities lead markets to produce a larger quantity than is socially desirable. Positive externalities lead markets to produce a smaller quantity than is socially desirable. To remedy the problem, the government can internalize the externality by taxing goods with negative externalities and subsidizing goods with positive externalities.**

Figure 3

Education and the Social Optimum

In the presence of a positive externality, the social value of the good exceeds the private value. The optimal quantity, $Q_{OPTIMUM}$, is therefore larger than the equilibrium quantity, Q_{MARKET}.



Case Study

Technology Spillovers, Industrial Policy, and Patent Protection

A potentially important type of positive externality is called a **technology spillover**—the impact of one firm's research and production efforts on other firms' access to technological advance.

Consider the market for industrial robots. Robots are at the frontier of a rapidly changing technology. Whenever a firm builds one, there is some chance that it will discover a new and better design. This new design may benefit not only this firm but also others in society because the design enters the collective pool of technological knowledge. That is, the new design may confer a positive externality on other producers.

In this case, the government can internalize the externality by subsidizing the production of robots. If the government paid firms a subsidy for each robot produced, the supply curve would shift down by the amount of the subsidy, and this shift would increase the equilibrium quantity of robots. To ensure that the market equilibrium equals the social optimum, the subsidy should equal the value of the technology spillover.

How large are technology spillovers, and what do they imply for public policy? This is an important question because technological progress is the key to raising living standards over time. Yet it is also a difficult question about which economists often disagree.

Some economists believe that technology spillovers are pervasive and that the government should encourage those industries that yield the largest spillovers. For instance, these economists argue that if making computer chips yields greater spillovers than making potato chips, the government should encourage the production of computer chips relative to the production of potato chips. The U.S. tax code does this in a limited way by offering special tax breaks for expenditures on research and development. Some nations go further by subsidizing specific industries that supposedly yield large technology spillovers. Government intervention that aims to promote technology-enhancing industries is sometimes called **industrial policy**.

Other economists are skeptical about industrial policy. Even if technology spillovers are common, pursuing an industrial policy requires the government to gauge the size of the spillovers from different markets. This measurement problem is difficult at best. Without accurate measurements, the political system may end up subsidizing industries with the most political clout rather than those with the largest positive externalities. And industrial policy may be pursued for reasons that have little to do with spillovers, such as protecting jobs that would otherwise disappear because of foreign competition. That may not be the most productive use of a society's resources.

Another way to deal with technology spillovers is patent protection. The patent laws protect the rights of inventors by giving them exclusive use of their inventions for a limited period. When a firm makes a breakthrough and patents the idea, it can capture much of the economic benefit for itself. The patent internalizes the externality by giving the firm a **property right** over its invention. If others want to use the technology, they must obtain permission from the inventing firm and pay it a royalty. The system is not perfect: The royalties charged by patent holders slow the dissemination of new technologies. But the patent system gives firms a greater incentive to engage in research and other activities that advance technology. ●

QuickQuiz

1. Which of the following is an example of a positive externality?
   a. Myra mows Dev's lawn and is paid $100 for the service.
   b. Myra's lawnmower emits smoke that Dev's neighbor Xavier has to breathe.
   c. Dev's newly cut lawn makes his neighborhood more attractive.
   d. Dev's neighbor Xavier offers to pay him if he keeps his lawn well groomed.

2. If the production of a good yields a negative externality, the social-cost curve lies _____ the supply curve, and the socially optimal quantity is _____ than the equilibrium quantity.
   a. above; greater
   b. above; less
   c. below; greater
   d. below; less

Answers are at the end of the chapter.

10-2 Public Policies toward Externalities

We have seen why externalities lead markets to allocate resources inefficiently but have mentioned only briefly how this inefficiency can be remedied. In practice, both public policymakers and private individuals respond to externalities in various ways. All of the remedies share the goal of moving the allocation of resources closer to the social optimum.

This section considers governmental solutions. As a general matter, the government can respond to externalities in one of two ways. **Command-and-control policies** regulate behavior directly. **Market-based policies** provide incentives so that private decision makers choose to solve the problem on their own. It will come as no surprise that economists tend to prefer market-based policies, though command-and-control policies are sometimes needed.

10-2a Command-and-Control Policies: Regulation

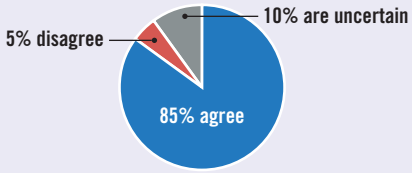
The government can remedy an externality by either requiring or forbidding certain behaviors. For example, it is a crime to dump poisonous chemicals into the water supply. In this case, the external costs to society far exceed the benefits to the polluter. The government therefore institutes a command-and-control policy that prohibits this act altogether.

Ask the Experts

Covid Vaccines

"Given the positive externalities from vaccination, an effective Covid-19 vaccine should be mandatory for every US resident (except those with health exceptions, such as infants and people with compromised immunity) with the cost covered by the federal government."

What do economists say?



Source: IGM Economic Experts Panel, June 23, 2020.

In many cases of pollution, however, the situation is not this simple. Despite the stated goals of some environmentalists, it would be impossible to prohibit all polluting activity. For example, virtually all forms of transportation—even the horse—produce some undesirable by-products. But it would not make sense for the government to ban all transportation. Instead of trying to eradicate pollution entirely, society must weigh the costs and benefits to decide the kinds and quantities of pollution it will allow. In the United States, the Environmental Protection Agency (EPA) is the government agency tasked with developing and enforcing regulations to protect the environment.

Environmental regulations can take many forms. The EPA may dictate the maximum level of pollution that a factory may emit. Sometimes, it requires firms to adopt a particular technology to reduce emissions. In all cases, to design good rules, the government regulators need to know the details about specific industries and about the alternative technologies that they could adopt. This information may be difficult for government regulators to obtain. Profit-seeking industries have little reason (other than from a general sense of social responsibility) to share what they know. Rather, they often have an incentive to conceal the adverse health effects from their production and exaggerate the cost of moving toward cleaner technologies.

10-2b Market-Based Policy 1: Corrective Taxes and Subsidies

Instead of regulating behavior in response to an externality, the government can use market-based policies to align private incentives with social efficiency. For instance, as we saw earlier, the government can internalize the externality by taxing activities with negative externalities and subsidizing activities with positive externalities. Taxes enacted to deal with the effects of negative externalities are called **corrective taxes**. They are also called **Pigovian taxes** after the economist Arthur Pigou (1877–1959), an early advocate of their use. An ideal corrective tax would equal the external cost from an activity with negative externalities, and an ideal corrective subsidy would equal the external benefit from an activity with positive externalities.

corrective tax
a tax designed to induce private decision makers to take into account the social costs that arise from a negative externality

Economists usually prefer corrective taxes to regulations as a way to deal with pollution because they can reduce pollution at a lower cost to society. To see why, consider an example.

Suppose that two factories—a paper mill and a steel mill—are each dumping 500 tons of "glop" into a river every year. Assume, further, that the glop from the two imaginary factories is identical, and that glop, while benign in small amounts, is dangerous in large quantities.

The EPA wants to reduce the amount emitted. It considers two solutions:

- Regulation: The EPA could tell each factory to reduce its pollution to 300 tons of glop per year.
- Corrective tax: The EPA could levy a tax on each factory of $50,000 for each ton of glop it emits.

The regulation would dictate a level of pollution, while the tax would give factory owners an incentive to reduce pollution. Which solution do you think is better?

In a case like this, most economists prefer the tax. To explain this preference, they would first point out that a tax is just as effective as regulation in reducing the overall level of many forms of pollution. The EPA can achieve whatever level of pollution it wants by setting the tax at the appropriate level. The higher the tax, the larger the reduction in pollution. If the tax is high enough, the factories will close altogether, reducing pollution to zero.

Although regulation and corrective taxes are both capable of reducing many types of pollution, the tax accomplishes this goal more efficiently. That's because the regulation requires each factory to reduce pollution by the same amount, which is not necessarily the least expensive way to clean up the water. It is possible that the paper mill can reduce pollution at a lower cost than the steel mill. If so, the paper mill would respond to the tax by reducing pollution substantially to avoid the tax, while the steel mill would reduce pollution less and pay the tax.

In essence, the corrective tax places a price on the right to pollute. This may seem strange if you're not accustomed to the way economists think. But just as markets allocate goods to those buyers who value them most, a corrective tax allocates pollution rights to those factories that face the highest cost of reducing it. With this flexible tool, the EPA can achieve any level of emission at the lowest total cost.

Economists also argue that corrective taxes are often better for the environment. Under the command-and-control policy of regulation, the factories have no reason to reduce emissions further once they reach the target of 300 tons of glop. By contrast, the tax gives the factories an incentive to develop cleaner technologies because a cleaner technology would reduce the amount of taxes they have to pay.

Corrective taxes differ from most other taxes. As Chapter 8 discussed, most taxes distort incentives and move the allocation of resources away from the social optimum. The reduction in economic well-being—that is, in consumer and producer surplus—exceeds the amount of revenue the government raises, resulting in a deadweight loss. But when externalities are present, society also needs to account for the well-being of bystanders. Corrective taxes give market participants the right incentives: By internalizing an externality, these taxes move the allocation of resources closer to the social optimum. Corrective taxes both raise government revenue and enhance economic efficiency.

CHRONICLE/ALAMY STOCK PHOTO

Arthur Pigou

Case Study

Why Is Gasoline Taxed So Heavily?

In many nations, gasoline is among the most heavily taxed goods. The gas tax can be viewed as a corrective tax aimed at addressing three negative externalities associated with driving:

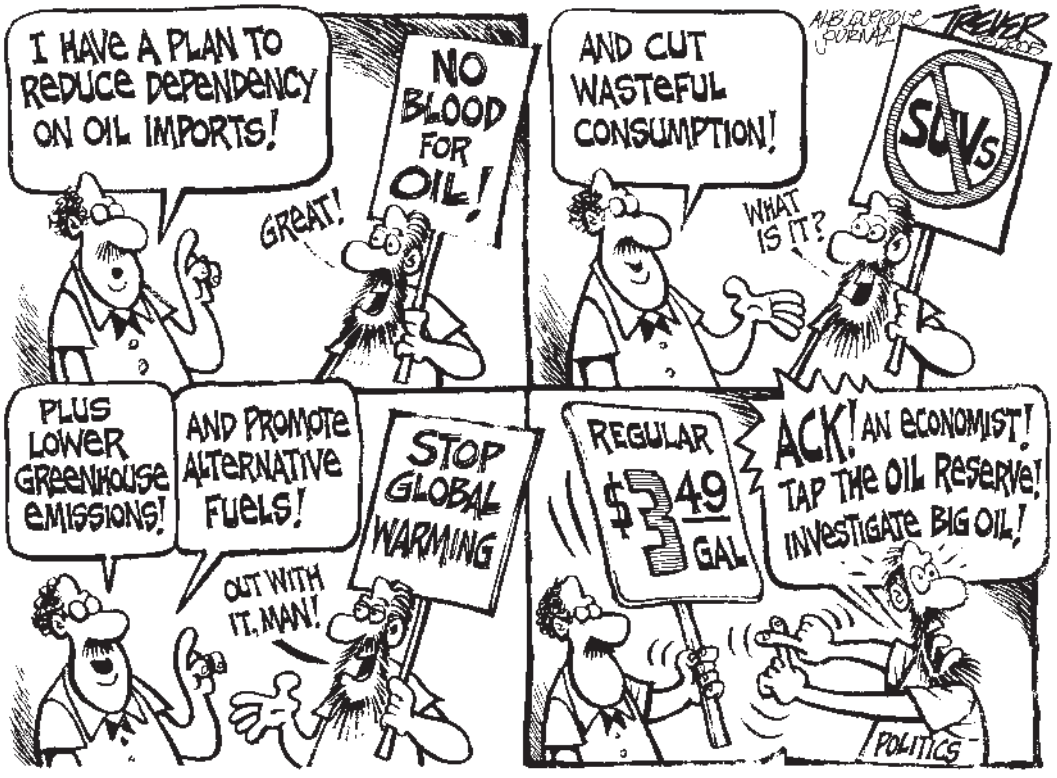
- **Pollution:** Car emissions cause smog, which increases the risk of heart and lung disease. In addition, the burning of fossil fuels such as gasoline is a primary cause of global climate change. The gas tax reduces these adverse effects by discouraging gas consumption.
- **Congestion:** Bumper-to-bumper traffic could be alleviated if there were fewer cars on the road. A gas tax keeps congestion down by encouraging people to take public transportation, ride bikes, carpool more often, and live closer to work.
- **Accidents:** When people buy large cars or sport utility vehicles, they may make themselves safer, but they also put their neighbors at risk. A person driving a smaller car is much more likely to be hurt if hit by a large car or a

sport utility vehicle than if hit by another smaller car. The gas tax is an indirect way of making people pay when their large, gas-guzzling vehicles put others at risk. It induces them to take this risk into account when choosing what vehicle to buy.

Rather than causing deadweight losses like most taxes, the gas tax makes the economy work better. It means a clearer environment, less traffic congestion, and safer roads.

How high should the tax on gasoline be? Most European countries impose gasoline taxes that are much higher than those in the United States. Many observers have suggested that the United States should also tax gasoline more heavily. A 2007 study published in the *Journal of Economic Literature* summarized the research on the size of the various externalities associated with driving. It concluded that the optimal corrective tax on gasoline was $2.28 per gallon in 2005 dollars; after adjusting for inflation, that amount is equivalent to about $3.20 per gallon in 2021 dollars. By contrast, the actual tax in the United States in 2021 was only about 55 cents per gallon.

The tax revenue from a higher gasoline tax could be used to lower taxes that distort incentives and cause deadweight losses, such as income taxes. In addition, the government regulations that require automakers to produce more fuel-efficient cars would be unnecessary, as would government subsidies to electric cars, because car consumers would face the right incentives. A higher gas tax, however, has never been politically popular. ●



10-2c Market-Based Policy 2: Tradable Pollution Permits

Returning to the paper mill and the steel mill, let's suppose that, despite the advice of its economists, the EPA adopts the regulation and requires each factory to reduce its pollution to 300 tons of glop per year. Then one day, after the regulation is in place and both mills have complied, the two firms go to the EPA with a proposal. The steel mill wants to increase its emission of glop from 300 to 400 tons. The paper mill has agreed to reduce its emission from 300 to 200 tons if the steel mill pays it $5 million. The total emission of glop would remain at 600 tons. Should the EPA allow the two factories to make this deal?

From the standpoint of economic efficiency, the answer is yes. The deal makes the owners of the two factories better off. After all, they are voluntarily agreeing to it. Moreover, the deal does not have any external effects because the total amount of pollution stays the same. Thus, social welfare is enhanced by allowing the paper mill to sell its pollution rights to the steel mill.

The same logic applies to any voluntary transfer of the right to pollute from one firm to another. When the EPA allows firms to make such deals, in essence, it creates a new scarce resource: pollution permits. A market to trade these permits develops, and that market is governed by the forces of supply and demand. The invisible hand ensures that this new market allocates the right to pollute efficiently. That is, the permits will end up in the hands of those who value them most, as judged by their willingness to pay. A firm's willingness to pay for the right to pollute, in turn, will depend on its cost of reducing pollution: The more costly it is for a firm to cut back on pollution, the more it will be willing to pay for a permit.

An advantage of allowing a market for pollution permits is that the initial allocation of the permits among firms does not matter from the standpoint of economic efficiency. Those firms that can reduce pollution at a low cost will sell whatever permits they get, while firms that can reduce pollution only at a high cost will buy whatever permits they need. If there is a market for the permits, the final allocation will be efficient regardless of the initial allocation.

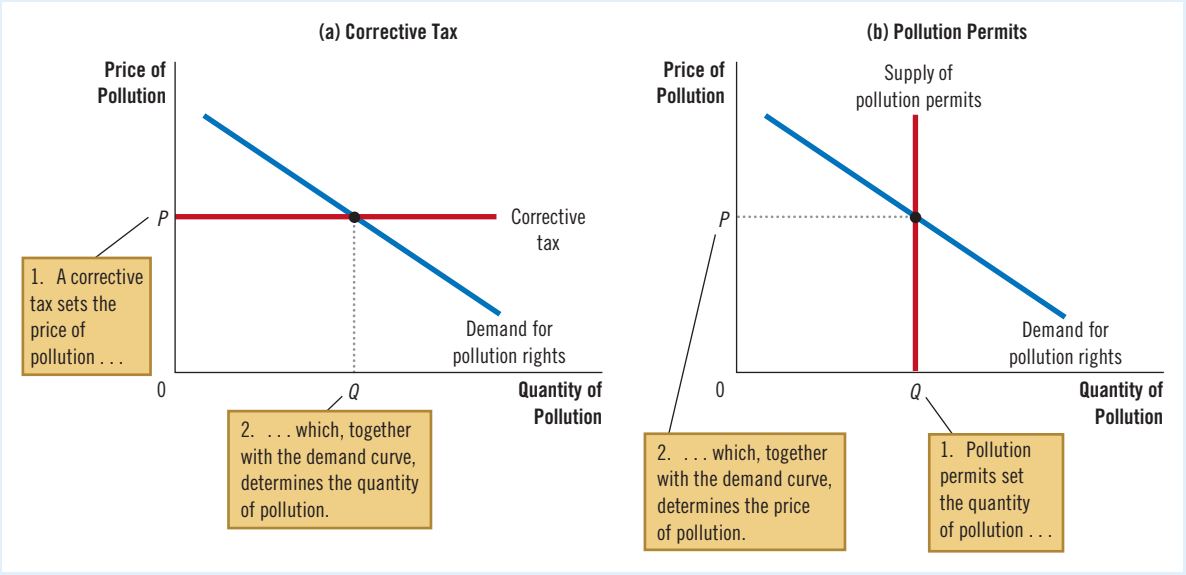
Pollution permits may seem very different from corrective taxes, but the two policies have much in common. In both cases, firms pay for their pollution. With corrective taxes, polluting firms must pay a tax to the government. With pollution permits, polluting firms must buy the permits. (Even firms that already own permits must pay to pollute, in the sense that they could have sold their permits on the open market, and, by foregoing that revenue, they incur an opportunity cost.) In the language of economics, both corrective taxes and pollution permits internalize the externality of pollution by making it costly for firms to pollute.

The similarity of the two policies can be seen by considering the market for pollution rights. Both panels in Figure 4 show the demand curve for the permission to pollute. This curve shows that the lower the price of polluting, the more firms will pollute. In panel (a), the EPA uses a corrective tax to set a price for pollution. In this case, the supply curve for pollution rights is perfectly elastic (because firms can pollute as much as they want by paying the tax), and the demand curve determines the quantity of pollution. In panel (b), the EPA sets a quantity of pollution by issuing permits. In this case, the supply curve for pollution rights is perfectly inelastic (because the quantity of pollution is fixed by the number of permits), and the demand curve determines the price. The EPA can achieve any point on the demand curve either by setting a price with a corrective tax or by setting a quantity with pollution permits.

Figure 4

The Equivalence of Corrective Taxes and Pollution Permits

In panel (a), the EPA sets a price on pollution by levying a corrective tax, and the demand curve determines the quantity of pollution. In panel (b), the EPA limits the quantity of pollution by limiting the number of pollution permits, and the demand curve determines the price of pollution. The price and quantity of pollution are the same in the two cases.



The choice between selling pollution permits and levying a corrective tax starts to matter, however, if the demand curve for pollution rights is uncertain. Suppose the EPA wants no more than 600 tons of glop dumped into the river, but because the EPA does not know the demand curve, it is not sure what size tax would hit that target. In this case, it can auction off 600 pollution permits. The auction price would, in effect, yield the corrective tax needed to achieve the EPA's goal. On the other hand, suppose the EPA knows the external cost of pollution is $50,000 per ton of glop but is uncertain how much glop factories would emit at that price. In this case, the EPA can reach the efficient outcome by setting a corrective tax of $50,000 per ton and letting the market determine the quantity of pollution.

The idea of the government auctioning off the right to pollute may at first sound like a creature of some economist's imagination. And, in fact, that is how the idea began. But increasingly, the EPA has used this system to control pollution. A notable success story has been the case of sulfur dioxide (SO_2), a leading cause of acid rain. In 1990, amendments to the Clean Air Act required power plants to reduce SO_2 emissions substantially. At the same time, the amendments set up a system that allowed plants to trade their SO_2 allowances. Initially, both industry representatives and environmentalists were skeptical of the proposal, but over time, the system reduced pollution with minimal disruption and at a low cost. Pollution permits, like corrective taxes, are now widely viewed as a cost-effective way to keep the environment clean.

10-2d Objections to the Economic Analysis of Pollution

"We cannot give anyone the option of polluting for a fee." This comment from Senator Edmund Muskie during the struggle to enact the 1972 Clean Water Act still reflects the view of some environmentalists. In 1969, the Cuyahoga River in Cleveland was so thick with industrial pollutants that it actually caught fire—a recurring problem that the legislation helped solve. Clean air and water were seen as fundamental human rights that had been despoiled. Remembering those days, many environmentalists say that economics is, at best, a secondary concern and, in some respects, a repugnant one. If clean air and water are priceless—part of the inheritance of every human—what kind of person would try to put a price on them? The environment is so important, some say, that protecting it is an absolute priority, regardless of cost.

Many modern economists are environmentalists and are committed to achieving clean air and water and ending global climate change, yet they view these public issues through a different lens. To economists, good environmental policy starts by acknowledging the first of the **Ten Principles of Economics** in Chapter 1: People face trade-offs. The value of environmental measures must be compared with their opportunity cost—that is, with what one must give up to obtain them. Eliminating all pollution is, sadly, impossible. Trying to eliminate all pollution is a lofty goal but not a feasible one—at least not in the foreseeable future—without reversing many of the technological advances that allow us to enjoy a high standard of living. Few people would be willing to accept poor nutrition, inadequate health care, or shoddy housing to make the environment as clean as possible.

The cause of environmentalism is arguably best advanced by thinking like an economist. A clean environment can be viewed as another good—one with tremendous value. Like all normal goods, it has a positive income elasticity: Rich countries can afford a cleaner environment than poor ones and usually have more rigorous environmental protection. In addition, like most other goods, clean air and water obey the law of demand: The lower the price of environmental protection, the more of it the public will want. When the economic approach of using corrective taxes and pollution permits reduces the cost of attaining clean air and water, it should increase the amount of environmental protection that the public demands.

Climate Change and Carbon Taxes

Scientists tell us that human carbon emissions are a cause of global climate change, which in turn has various harmful effects. This is a classic example of a negative externality.

Suppose a concerned citizen—let's call him Yoram—wants to reduce his carbon footprint. How might he do it?

- Yoram could buy a more fuel-efficient car, like a hybrid or an electric vehicle.
- Yoram could carpool to work.
- Yoram could use public transportation more often.
- Yoram could move closer to his job.
- Yoram could buy a smaller house that requires less energy to heat and cool.
- Yoram could adjust the thermostat to keep his home cooler in the winter and warmer in the summer.
- Yoram could put solar panels on his roof.

- Yoram could buy more energy-efficient home appliances.
- Yoram could eat more locally produced foods, which need less fuel to transport.
- Yoram could stop eating beef, as its production releases an immense amount of greenhouse gases.

By now, you get the idea. Every day, we all make lifestyle choices that affect how much carbon is emitted. These decisions are personal, but they have a global impact.

The main question for policymakers is how society can ensure that people make the right decisions, which take into account both the personal impact of their actions and the externalities. There are three approaches.

One approach is to appeal to individuals' sense of social responsibility. Some people may, in fact, reduce their carbon footprint to the optimal level out of concern for the planet. But expecting most people to act this way is unrealistic. Life is busy, people have their own priorities, and even knowing the global impact of one's own actions is a daunting task.

The second approach is to use government regulation to change the decisions that people make. An example is Corporate Average Fuel Economy, or CAFE, standards, which regulate car emissions in the United States.

But this regulatory approach is fraught with problems. One is that it creates tension between the products that consumers want to buy and the products that companies are allowed to sell. Robert A. Lutz, a former top auto executive, laments that CAFE standards are "a huge bureaucratic nightmare." He says, "CAFE is like trying to cure obesity by requiring clothing manufacturers to make smaller sizes."

A more important problem with such regulations is that they can influence only a few crucial decisions. The government can regulate the kind of car that Yoram finds at auto dealers, but it can't easily regulate how close Yoram lives to work or whether he carpools with his neighbor. Yet reducing carbon emissions at minimum cost requires a policy that encompasses all possible margins of adjustment.

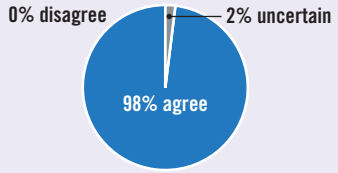
Fortunately, a policy broader in scope is possible, which brings us to the third approach to dealing with climate externalities: putting a price on carbon emissions. This can be done either by taxing them or by creating a "cap-and-trade" system of tradable pollution permits. If sellers of gasoline, electricity, beef, and so on had to pay a fee for each emission of carbon implicit in their products, that fee would be built into their prices. When making everyday decisions, Yoram and other consumers would look at the prices they face and, in effect, take into account the global impact of their choices. A price on carbon would internalize the externality. It has worked effectively in Europe, though it hasn't caught on in the United States.

But carbon taxes are a popular policy among economists, as the nearby "Ask the Experts" box shows. ●

Ask the Experts — Carbon Taxes

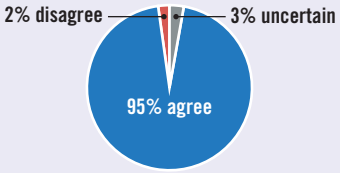
"The Brookings Institution recently described a U.S. carbon tax of $20 per ton, increasing at 4 percent per year, which would raise an estimated $150 billion per year in federal revenues over the next decade. Given the negative externalities created by carbon dioxide emissions, a federal carbon tax at this rate would involve fewer harmful net distortions to the U.S. economy than a tax increase that generated the same revenue by raising marginal tax rates on labor income across the board."

What do economists say?



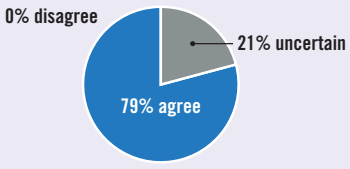
"A tax on the carbon content of fuels would be a less expensive way to reduce carbon dioxide emissions than would a collection of policies such as 'corporate average fuel economy' requirements for automobiles."

What do economists say?



"Carbon taxes are a better way to implement climate policy than cap-and-trade."

What do economists say?



Source: IGM Economic Experts Panel, December 4, 2012, December 20, 2011, and November 13, 2018.

QuickQuiz

3. When the government levies a tax equal to the external cost associated with a good's production, it _____ the price paid by consumers and makes the market outcome _____ efficient.
   a. increases; more
   b. increases; less
   c. decreases; more
   d. decreases; less
4. Which of the following statements about corrective taxes is generally NOT true?
   a. They reduce consumer surplus.
   b. They raise government revenue.
   c. They reduce the quantity sold in a market.
   d. They cause deadweight losses.
5. The government auctions off 500 units of pollution rights. The rights sell for $50 per unit, raising $25,000 of revenue for the government. This policy is equivalent to a corrective tax of _____ per unit of pollution.
   a. $10
   b. $50
   c. $450
   d. $500
6. Command-and-control regulation may be better than a corrective tax if
   a. a corrective tax would have differential effects on different industries.
   b. some polluters can reduce emissions more cheaply than others.
   c. the negative externality is so large that the optimal quantity is zero.
   d. knowledge about the cost of pollution abatement is dispersed and hard to obtain.

Answers are at the end of the chapter.

10-3 Private Solutions to Externalities

Externalities tend to make market outcomes inefficient, but government action isn't the only remedy. Sometimes, people can find private solutions.

10-3a The Types of Private Solutions

Moral codes and social sanctions can solve the problem of externalities. Consider, for instance, why most people do not litter. There are laws against littering, but they are not rigorously enforced. Most people don't litter just because it is the wrong thing to do. The Golden Rule taught to many children says, "Do unto others as you would have them do unto you." This moral injunction tells us to consider how our actions affect other people. In the less poetic language of economics, it tells us to internalize externalities.

Another private solution involves charities. For example, the Sierra Club is a nonprofit organization funded by private donors who seek to protect the environment. Colleges and universities receive gifts from alumni, corporations, and foundations in part because education has great benefits—positive externalities—for society. The government encourages these private solutions by allowing an income tax deduction for charitable donations.

The private market can often solve the problem of externalities by relying on the self-interest of the relevant parties. Sometimes, the solution takes the form of integrating different types of businesses. For example, consider an apple grower and a beekeeper who are located next to each other. Each business benefits the other, conveying a positive externality: By pollinating the flowers on the trees, the bees help the orchard produce apples, while the bees use the nectar from the apple trees to produce honey. But that doesn't mean that the two take these benefits into account. When the apple grower decides how many trees to plant and the beekeeper determines how many bees to keep, they neglect the positive externality. As a result, the apple grower plants too few trees, and the beekeeper keeps too few bees. These

externalities could be internalized if the bee and apple businesses merged: Both activities would then take place within the same firm, which could choose the optimal number of trees and bees. Internalizing externalities is one reason that some firms are involved in multiple types of businesses.

Another private approach is for the apple grower and beekeeper to negotiate a contract, setting the right number of trees and bees and perhaps a payment from one party to the other. The contract can solve the inefficiency that normally arises from these externalities and make both parties better off.

10-3b The Coase Theorem

Coase theorem
the proposition that if private parties can bargain without cost over the allocation of resources, they can solve the problem of externalities on their own

How effective is the private market in dealing with externalities? A famous result, called the **Coase theorem** after the economist Ronald Coase, suggests that it can be very effective in some circumstances. According to the Coase theorem, if private parties can bargain over the allocation of resources at no cost, then the private market will always solve the problem of externalities and allocate resources efficiently.

To see how the Coase theorem works, consider an example. Suppose that Emily owns a dog named Clifford. Clifford barks and disturbs Horace, Emily's neighbor. Emily benefits greatly from Clifford's companionship, but the dog confers a negative externality on Horace, who would prefer a sound sleep. Should Emily find Clifford a new home, or should Horace suffer restless nights because of Clifford's barking?

Consider first what outcome is socially efficient. An economist with society's interest at heart would compare the benefit that Emily gets from the dog with the cost that Horace bears from the barking. If the benefit exceeds the cost, it is efficient for Emily to keep the dog and for Horace to live with the barking. Yet if the cost exceeds the benefit, then Emily should get rid of the dog.

According to the Coase theorem, the private market will reach the efficient outcome on its own. How? Horace can offer to pay Emily to move Clifford to a new home. Emily will accept the deal if the amount of money Horace offers is greater than her benefit of keeping the dog.

By bargaining over the price, Emily and Horace can always reach the efficient outcome. For instance, suppose that Emily gets a $1,000 benefit from the dog and Horace bears a $1,500 cost from the barking. In this case, Horace can offer Emily $1,200 to get rid of the dog, and Emily will accept. Both parties are better off than they were before, and the efficient outcome is reached.

It is possible, of course, that Horace would not be willing to offer any price that Emily would accept. For instance, suppose that Emily gets a $3,000 benefit from the dog and Horace bears a $1,500 cost from the barking. In this case, Emily would turn down any offer below $3,000, while Horace would not offer any amount above $1,500. Emily ends up keeping Clifford. Given these costs and benefits, this outcome is efficient.

These examples so far assume that Emily has the legal right to keep a barking dog. In other words, Emily can keep Clifford unless Horace pays her enough to induce her to give up the dog voluntarily. But how different would the outcome be if Horace had the legal right to peace and quiet?

According to the Coase theorem, the initial distribution of rights does not matter for the market's ability to reach the efficient outcome. For instance, suppose that Horace can legally compel Emily to get rid of the dog. Having this right works to Horace's advantage, but it probably won't change the outcome. In this case, Emily can offer to pay Horace to allow her to keep the dog. If the benefit of the dog to

Emily exceeds the cost of the barking to Horace, then Emily and Horace will strike a bargain in which Emily gets to keep her dog.

Although Emily and Horace can reach the efficient outcome regardless of how rights are initially distributed, the distribution of rights is not irrelevant: It determines the distribution of economic well-being. Whether Emily has the right to a barking dog or Horace the right to peace and quiet determines who pays whom in the final bargain. But in either case, the two parties can bargain with each other and solve the externality problem. Emily will end up keeping Clifford only if her benefit exceeds Horace's cost.

To sum up: **The Coase theorem says that private economic actors can potentially solve the problem of externalities among themselves. Whatever the initial distribution of rights, the interested parties can reach a bargain in which everyone is better off and the outcome is efficient.**

10-3c Why Private Solutions Do Not Always Work

Despite the appealing logic of the Coase theorem, private individuals on their own often fail to resolve the problems caused by externalities. The Coase theorem applies only when the interested parties have no trouble reaching and enforcing an agreement. In the real world, however, bargaining does not always work, even when a mutually beneficial agreement is possible.

Sometimes, the interested parties fail to solve an externality problem because of **transaction costs**, the costs that parties incur in the process of agreeing to and following through on a bargain. In our example, imagine that Emily and Horace speak different languages so, to reach an agreement, they need to hire a translator. If the benefit of solving the barking problem is less than the cost of the translator, Emily and Horace might choose to leave the problem unsolved. In more realistic examples, the transaction costs are the expenses not of translators but of lawyers required to draft and enforce contracts.

transaction costs the costs that parties incur during the process of agreeing to and following through on a bargain

At other times, bargaining simply breaks down. The recurrence of wars and labor strikes shows that reaching an agreement can be difficult and that failing to reach one can be costly. The problem is often that each party tries to hold out for a better deal. For example, suppose that Emily gets a $1,000 benefit from having Clifford, while Horace bears a $1,500 cost from the barking. Although it is efficient for Horace to pay Emily to find another home for the dog, there are many prices that could lead to this outcome. Emily might demand $1,400, and Horace might offer only $1,100. As they haggle over the price, the inefficient outcome with the barking dog persists.

Reaching an efficient bargain is especially difficult when the number of interested parties is large because coordinating everyone is costly. For example, consider a factory that pollutes the water of a nearby lake. The pollution confers a negative externality on the local fishermen. According to the Coase theorem, if the pollution is inefficient, then the factory and the fishermen could reach a bargain in which the fishermen pay the factory not to pollute. If there are many fishermen, however, trying to coordinate them all to bargain with the factory may be almost impossible. Combating global climate change is similar: There are far too many people and firms responsible for greenhouse gas emissions for them to be able to bargain together and come up with a worldwide solution.

When private bargaining does not work, governments can sometimes play a role. A government is an institution designed for collective action. In the example of the lake, the government can act on behalf of the fishermen, even when it is impractical for the fishermen to act for themselves.

In the News

The Coase Theorem in Action

Whenever people come in close contact, externalities abound.

Don't Want Me to Recline My Airline Seat? You Can Pay Me

By Josh Barro

I fly a lot. When I fly, I recline. I don't feel guilty about it. And I'm going to keep doing it, unless you pay me to stop.

I bring this up because of a dispute you may have heard about: On Sunday, a United Airlines flight from Newark to Denver made an unscheduled stop in Chicago to discharge two passengers who had a dispute over seat reclining. According to The Associated Press, a man in a middle seat installed the Knee Defender, a $21.95 device that keeps a seat upright, on the seatback in front of him.

A flight attendant asked him to remove the device. He refused. The woman seated in front of him turned around and threw water at him. The pilot landed the plane and booted both passengers off the flight.

Obviously, it's improper to throw water at another passenger on a flight, even if he deserves it. But I've seen a distressing amount of sympathy for Mr. Knee Defender, who wasn't just instigating a fight but usurping his fellow passenger's property rights. When you buy an airline ticket, one of the things you're buying is the right to use your seat's reclining function. If this passenger so badly wanted the passenger in front of him not to recline, he should have paid her to give up that right.

I wrote an article to that effect in 2011, noting that airline seats are an excellent case study for the Coase Theorem. This is an economic theory holding that it doesn't matter very much who is initially given a property right; so long as you clearly define it and transaction costs are low, people will trade the right so that it ends up in the hands of whoever values it most. That is, I own the right to recline, and if my reclining bothers you, you can pay me to stop. We could (but don't) have an alternative system in which the passenger sitting behind

QuickQuiz

7. According to the Coase theorem,
   a. private actors can reach an agreement to solve the problem of externalities without the government.
   b. corrective subsidies are the best policy to solve the problem of positive externalities.
   c. negative externalities are a problem for society, but positive externalities are not.
   d. when two private actors amicably solve the problem of externalities, they shift the problem to a third party.

8. The Coase theorem does NOT apply if
   a. there is a significant externality between two parties.
   b. the court system vigorously enforces all contracts.
   c. transaction costs make negotiation difficult.
   d. both parties understand the externality fully.

Answers are at the end of the chapter.

10-4 Conclusion

The invisible hand is powerful but not omnipotent. A market's equilibrium maximizes the sum of producer and consumer surplus. When the buyers and sellers in the market are the only interested parties, this outcome is efficient from the standpoint of society. But when there are external effects, such as pollution, evaluating a market outcome requires accounting for the well-being of third parties. In this case, the invisible hand of the marketplace may fail to allocate resources efficiently.

In some cases, people can solve the problem of externalities on their own. The Coase theorem suggests that the interested parties can bargain among themselves and agree on an efficient solution. Sometimes, however, an efficient outcome cannot be reached, perhaps because the large number of interested parties makes bargaining difficult.

me owns the reclining rights. In that circumstance, if I really care about being allowed to recline, I could pay him to let me.

Donald Marron, a former director of the Congressional Budget Office, agrees with this analysis, but with a caveat. Recline negotiations do involve some transaction costs—passengers don't like bargaining over reclining positions with their neighbors, perhaps because that sometimes ends with water being thrown in someone's face.

Mr. Marron says we ought to allocate the initial property right to the person likely to care most about reclining, in order to reduce the number of transactions that are necessary. He further argues that it's probably the person sitting behind, as evidenced by the fact people routinely pay for extra-legroom seats.

GCHUTKA/E+/GETTY IMAGES

Mr. Marron is wrong about this last point. I understand people don't like negotiating with strangers, but in hundreds of flights I have taken, I have rarely had anyone complain to me about my seat recline, and nobody has ever offered me money, or anything else of value, in exchange for sitting upright.

If sitting behind my reclined seat was such misery, if recliners like me are "monsters," as Mark Hemingway of *The Weekly Standard* puts it, why is nobody willing to pay me to stop? People talk a big game on social media about the terribleness of reclining, but then people like to complain about all sorts of things; if they really cared that much, someone would have opened his wallet and paid me by now. ■

Questions to Discuss

1. Can you imagine offering a person sitting in front of you on an airplane some money not to recline his seat? Why or why not?
2. If a person sitting behind you on an airplane offered you some money not to recline your seat, how would you respond? Why?

Source: *New York Times*, August 27, 2014.

When people cannot solve the problem of externalities privately, the government often steps in. Yet even with government intervention, society can benefit by harnessing—rather than circumventing—market forces. The government can require market participants to bear the full costs of their actions. Corrective taxes on emissions and pollution permits, for instance, are both designed to internalize the externality of pollution. These are increasingly the policies of choice for those interested in protecting the environment. Market forces, properly redirected, are often the best remedy for market failure.

Chapter in a Nutshell

- When a transaction between a buyer and seller directly affects a third party, the effect is called an externality. If an activity yields negative externalities, such as pollution, the socially optimal quantity in a market is less than the equilibrium quantity. If an activity yields positive externalities, such as technology spillovers, the socially optimal quantity is greater than the equilibrium quantity.
- Government can remedy the inefficiencies caused by externalities. Sometimes, it regulates behavior. Other times, it internalizes an externality using corrective taxes. Another policy is to issue permits. For example, the government could protect the environment by issuing a limited number of pollution permits. The result of this policy is similar to imposing corrective taxes on polluters.
- Those affected by externalities can sometimes solve the problem privately. For instance, when one business imposes an externality on another business, the two businesses can internalize the externality by merging. Alternatively, the interested parties can solve the problem by negotiating a contract. According to the Coase theorem, if people can bargain without cost, they can reach an agreement in which resources are allocated efficiently. In many cases, however, reaching a bargain among the many interested parties is difficult, so the Coase theorem does not apply.

Key Concepts

externality, p. 190
internalizing the externality, p. 193
corrective taxes, p. 196
Coase theorem, p. 204
transaction costs, p. 205

Questions for Review

1. Give an example of a negative externality and an example of a positive one.
2. Draw a supply-and-demand diagram to explain the effect of a negative externality that occurs because of a firm's production process.
3. How does the patent system help society solve an externality problem?
4. What are corrective taxes? Why do economists prefer them to regulations as a way of protecting the environment from pollution?
5. List some of the ways that the problems caused by externalities can be solved without government intervention.
6. Imagine that you are a nonsmoker sharing a room with a smoker. According to the Coase theorem, what determines whether your roommate smokes in the room? Is this outcome efficient? How do you and your roommate reach this solution?

Problems and Applications

1. Consider two ways to protect your car from theft. The Club (a steering wheel lock) makes it difficult for a car thief to take your car. Lojack (a tracking system) makes it easier for the police to catch the car thief who has stolen it. Which of these methods confers a negative externality on other car owners? Which confers a positive externality? Do you think there are any policy implications of your analysis?
2. Consider the market for fire extinguishers.
   a. Why might fire extinguishers exhibit positive externalities?
   b. Draw a graph of the market for fire extinguishers, labeling the demand curve, the social-value curve, the supply curve, and the social-cost curve.
   c. Indicate the market equilibrium level of output and the efficient level of output. Give an intuitive explanation for why these quantities differ.
   d. If the external benefit is $10 per extinguisher, describe a government policy that would yield the efficient outcome.
3. Greater consumption of alcohol leads to more motor vehicle accidents and, thus, imposes costs on people who do not drink and drive.
   a. Illustrate the market for alcohol, labeling the demand curve, the social-value curve, the supply curve, the social-cost curve, the market equilibrium level of output, and the efficient level of output.
   b. On your graph, shade the area corresponding to the deadweight loss of the market equilibrium. (Hint: The deadweight loss occurs because some units of alcohol are consumed for which the social cost exceeds the social value.) Explain.
4. Some observers believe that the current levels of pollution in our society are too high.
   a. If society wishes to reduce overall pollution by a certain amount, why might different amounts of reduction at different firms be efficient?
   b. Command-and-control approaches often rely on uniform reductions among firms. Why are these approaches generally unable to target the firms that should undertake bigger reductions?
   c. Economists argue that appropriate corrective taxes or tradable pollution permits will result in efficient pollution reduction. How do these approaches target the firms that should undertake bigger reductions?
5. The many identical residents of Whoville love drinking Zlurp. Each resident has the following willingness to pay for the tasty refreshment:

First bottle	$5
Second bottle	4
Third bottle	3
Fourth bottle	2
Fifth bottle	1
Further bottles	0

   a. The cost of producing Zlurp is $1.50, and the competitive suppliers sell it at this price. (The supply curve is horizontal.) How many bottles

will each Whovillian consume? What is each person's consumer surplus?

b. Producing Zlurp creates pollution. Each bottle has an external cost of $1. Taking this additional cost into account, what is total surplus per person in the allocation you described in part (a)?

c. Cindy Lou Who, one of the residents of Whoville, decides on her own to reduce her consumption of Zlurp by one bottle. What happens to Cindy's welfare (her consumer surplus minus the cost of pollution she experiences)? How does Cindy's decision affect total surplus in Whoville?

d. Mayor Grinch imposes a $1 tax on Zlurp. What is consumption per person now? Calculate consumer surplus, the external cost, government revenue, and total surplus per person.

e. Based on your calculations, would you support the mayor's policy? Why or why not?

6. Bruno loves playing rock 'n' roll music at high volume. Placido loves opera and hates rock 'n' roll. Unfortunately, they are next-door neighbors in an apartment building with paper-thin walls.
   a. What is the externality here?
   b. What command-and-control policy might the landlord impose? Could such a policy lead to an inefficient outcome?
   c. Suppose the landlord lets the tenants do whatever they want. According to the Coase theorem, how might Bruno and Placido reach an efficient outcome on their own? What might prevent them from reaching an efficient outcome?

7. Figure 4 shows that for any given demand curve for the right to pollute, the government can achieve the same outcome either by setting a price with a corrective tax or by setting a quantity with pollution permits. Suppose there is a sharp improvement in the technology for controlling pollution.
   a. Using graphs similar to those in Figure 4, illustrate the effect of this development on the demand for pollution rights.
   b. What is the effect on the price and quantity of pollution under each regulatory system? Explain.

8. Suppose that the government decides to issue tradable permits for a certain form of pollution.
   a. Does it matter for economic efficiency whether the government distributes or auctions the permits? Why or why not?
   b. If the government chooses to distribute the permits, does the allocation of permits among firms matter for efficiency? Explain.

9. There are three industrial firms in Happy Valley.

Firm	Initial Pollution Level	Cost of Reducing Pollution by 1 Unit
A	30 units	$20
B	40 units	$30
C	20 units	$10

The government wants to reduce pollution to 60 units, so it gives each firm 20 tradable pollution permits.
   a. Who sells permits, and how many do they sell? Who buys permits, and how many do they buy? Briefly explain why the sellers and buyers are each willing to do so. What is the total cost of pollution reduction in this situation?
   b. How much higher would the costs of pollution reduction be if the permits could not be traded?

QuickQuiz Answers

1. **c** 2. **b** 3. **a** 4. **d** 5. **b** 6. **c** 7. **a** 8. **c**

Chapter

11

Public Goods and Common Resources

"The best things in life are free," the old song tells us, and it may well be true. Rivers, mountains, beaches, lakes, and oceans are nature's bounty, available to all. Playgrounds, parks, and parades are often provided by governments, and people usually don't have to pay anything to enjoy them.

Goods without prices don't easily fit into the kind of analysis we've practiced so far. We've focused mainly on items that are allocated through markets, in which buyers pay for what they receive and sellers are paid for what they provide. In these cases, prices guide the decisions of buyers and sellers, leading to an efficient allocation of resources. But these clear market signals are absent when goods and services are available free of charge.

Without prices, private markets on their own cannot ensure that such goods are made available and used correctly for the maximum benefit of society as a whole. This chapter examines these problems and shows that government policy can often fix the market failure and increase economic well-being. This conclusion sheds light on one of the **Ten Principles of Economics** in Chapter 1: Governments can sometimes improve market outcomes.

ROMAN SAMBORSKYI/SHUTTERSTOCK.COM, ARNO VAN ENGELAND/EYEEM/GETTY IMAGES

11-1 The Different Kinds of Goods

How well do markets work in providing people with what they want? The answer depends on the good being considered. Chapters 4 and 7 showed that for ice-cream cones, markets work efficiently: The price of ice-cream cones adjusts to balance supply and demand, and this equilibrium maximizes the sum of producer and consumer surplus. Yet Chapter 10 showed that the market cannot be counted on to produce such fine results in all cases. For example, the market on its own can't prevent steel manufacturers from polluting the air we breathe. Buyers and sellers of steel typically don't take into account the external effects of their decisions. Markets work well for ice cream but not for clean air.

When thinking about the ability of markets to efficiently allocate the economy's resources, it helps to classify goods more systematically. As we will see with some examples, two characteristics are most important:

excludability
the property of a good whereby a person can be prevented from using it

rivalry in consumption
the property of a good whereby one person's use diminishes other people's use

- **Excludability**. If people can be prevented from using a good, it is excludable. If it is impossible to prevent people from using the good, it is not excludable.
- **Rivalry in consumption**. If one person's use of a unit of a good reduces another person's ability to use it, the good is rival in consumption. If one person's use does not diminish another person's use, the good is not rival in consumption.

On a grid, these two characteristics define four categories, shown in Figure 1:

private goods
goods that are both excludable and rival in consumption

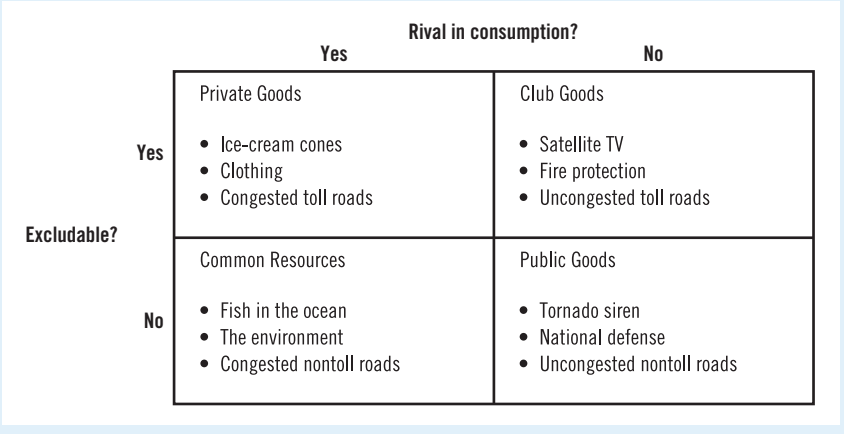
1. **Private goods** are both excludable and rival in consumption. An ice-cream cone, for instance, is excludable because it is possible to prevent someone from eating one—you just don't give it to her. It is also rival in consumption because if one person eats an ice-cream cone, another person cannot eat the same cone.

   Most goods in the economy are private goods like ice-cream cones: You don't get one unless you pay for it, and once you have it, you are the only person who benefits. When we analyzed supply and demand in Chapters 4,

Figure 1

Four Types of Goods

Goods can be grouped into four categories according to two characteristics: (1) A good is **excludable** if people can be prevented from using it. (2) A good is **rival in consumption** if one person's use of the good diminishes other people's use of it. This diagram gives examples of goods in each category.



5, and 6 and the efficiency of markets in Chapters 7, 8, and 9, we implicitly assumed that goods were both excludable and rival in consumption.

2. **Public goods** are neither excludable nor rival in consumption. They aren't excludable because people cannot be prevented from using a public good, and they aren't rival because one person's use does not reduce another person's ability to use it.

   An example is a tornado siren in a small town. Once the siren sounds, it is impossible to prevent a person from hearing it, so it is not excludable. And when one person gets the benefit of the warning, that does not reduce the benefit to others, so the siren is not rival in consumption.

3. **Common resources** are rival in consumption but not excludable. Consider fish in the ocean. They are rival in consumption because when one person catches fish, fewer fish are left for the next person. Fish are not excludable because it is hard to stop fishermen from taking fish out of a vast ocean.

4. **Club goods** are excludable but not rival in consumption. An example is satellite TV. If you don't pay your bill, the company can cut you off, making the good excludable. But your use of the satellite signal does not diminish anyone else's ability to use it, so the service is not rival in consumption. (We will later return to club goods and see that they are a type of a **natural monopoly**.)

public goods
goods that are neither excludable nor rival in consumption

common resources
goods that are rival in consumption but not excludable

club goods
goods that are excludable but not rival in consumption

Figure 1 separates goods into four distinct categories. That's helpful in thinking about these categories, but it is not completely realistic because the boundaries can be fuzzy. Whether goods are excludable or rival in consumption is often a matter of degree. Fish in an ocean may not be excludable because monitoring fishing everywhere is so hard, but a large enough coast guard could make fish partly excludable, at least in some sections of the ocean. Similarly, although fish are generally rival in consumption, this would be less true if the number of fishermen were small relative to the number of fish. (Think of North America before the arrival of European settlers.) Despite this fuzziness, however, it is useful to group goods into these four categories.

This chapter examines goods that are not excludable: public goods and common resources. Because people cannot be prevented from using them, these goods are available to everyone free of charge. But when something of value has no price attached to it, externalities arise. For example, if a private individual were to provide a public good, such as a tornado siren, other people would be better off. They would receive a benefit without paying for it—a positive externality. Similarly, when one person uses a common resource like fish in the ocean, other people are worse off because there are fewer fish to catch, yet they are not compensated for their loss. This is an externality too, but a negative one. Because of these external effects, private decisions about consumption and production can lead to an inefficient allocation of resources unless the government policy steps in to fix the problem.

QuickQuiz

1. Which categories of goods are excludable?
   a. private goods and club goods
   b. private goods and common resources
   c. public goods and club goods
   d. public goods and common resources

2. Which categories of goods are rival in consumption?
   a. private goods and club goods
   b. private goods and common resources
   c. public goods and club goods
   d. public goods and common resources

Answers are at the end of the chapter.

11-2 Public Goods

To understand how public goods differ from other goods and why they present problems for society, consider a fireworks display. It isn't excludable: Anyone can look up at the sky and see it. And it's not rival in consumption: My joy at the sight of a chrysanthemum exploding above me doesn't diminish your pleasure at seeing it.

11-2a The Free-Rider Problem

The citizens of Smalltown, U.S.A., love fireworks on the Fourth of July. Each of the town's 500 residents places a $10 value on this experience for a total benefit of $5,000. The cost of putting on the display is $1,000. Because the $5,000 benefit exceeds the $1,000 cost, it is efficient for Smalltown residents to enjoy Fourth of July fireworks.

Would the private market produce this outcome? Probably not. Imagine that Zoe, a Smalltown entrepreneur, decided to put on a fireworks display. Zoe would have trouble selling tickets because potential customers would realize that they could see the fireworks without a ticket. Because fireworks are not excludable, it is easy to be a **free rider**—the term for a person who receives the benefit of a good without paying for it. If most people would choose to be free riders rather than ticket buyers, the market would fail to provide the efficient outcome.

free rider
a person who receives the benefit of a good but avoids paying for it

This market failure can be viewed as arising from an externality. If Zoe puts on the fireworks display, she confers an external benefit on those who see it without paying for it. When deciding whether to put on the display, however, Zoe does not take the external benefits into account. Zoe isn't a philanthropist. Even though the fireworks display is socially desirable, it is not profitable, so Zoe makes the privately rational but socially inefficient decision not to put on the display.

Although the private market fails to supply the fireworks display Smalltown residents want, the solution to Smalltown's problem is obvious: The local government can sponsor a Fourth of July celebration. The town council can raise everyone's taxes by $2 and hire Zoe to produce the fireworks. Everyone in Smalltown is better off by $8—the $10 at which residents value the fireworks minus the $2 tax bill. Zoe can help Smalltown reach the efficient outcome as a public employee even though she could not do so as a private entrepreneur.

This story of Smalltown's government-sponsored fireworks is simplified but realistic. In fact, many local governments in the United States pay for fireworks on the Fourth of July. More importantly, the story shows a general lesson: Because public goods are not excludable, the free-rider problem often prevents the private market from supplying them. The government, however, can remedy the problem. If the government decides that the total benefits of a public good exceed its costs, it can provide the public good, pay for it with tax revenue, and potentially make everyone better off.

11-2b Some Important Public Goods

There are many examples of public goods. Here are three of the most critical.

National Defense The defense of a country from foreign aggressors is a classic example of a public good. Once a country has paid for a military to defend it, it is impossible to prevent a person in the country from enjoying the benefit of this

defense. And when one person enjoys the benefit of national defense, she does not reduce the benefit to anyone else. National defense is neither excludable nor rival in consumption.

National defense is also one of the most expensive public goods. In 2020, the U.S. federal government spent a total of $886 billion on national defense, or $2,682 per person. People debate whether this amount is too small or too large and whether all military spending actually makes us safer, but few doubt that some government spending for national defense is necessary. Even economists who advocate small government agree that national defense is a public good the government must provide.

DANA FRADON/THE NEW YORKER COLLECTION/THE CARTOON BANK

"I like the concept if we can do it with no new taxes."

Basic Research Research increases knowledge about many things. When evaluating the public policy toward research, it is important to distinguish general knowledge from specific technological knowledge, like the invention of a longer-lived battery, a smaller microchip, or a better digital music player. This kind of knowledge can be patented. The patent gives the inventor the exclusive right to the creation for a limited period. Anyone else who wants to use it must pay the inventor for the right to do so. In other words, the patent makes the inventor's work excludable.

By contrast, general knowledge is a public good. For example, a mathematician cannot patent a theorem. Once proven, the knowledge is not excludable: The theorem enters the collective pool of knowledge that anyone can use without charge. The theorem is also not rival in consumption: One person's use of the theorem does not prevent any other person from using it.

Profit-seeking firms spend a lot on research to develop products they can patent and sell, but most don't spend much on basic research. Their incentive, instead, is to free ride on the general knowledge created by others. As a result, in the absence of any public policy, society would devote too few resources to creating knowledge.

The government tries to provide the public good of general knowledge in various ways. Government agencies, such as the National Institutes of Health and the National Science Foundation, subsidize basic research in medicine, mathematics, physics, chemistry, biology, and even economics. Some people justify government funding of the space program on the grounds that it adds to the pool of knowledge. Determining the appropriate level of government support is difficult because the benefits are hard to measure, and the members of Congress who appropriate funds usually don't have the expertise necessary to judge what lines of research are likely to produce the largest benefits. So, while basic research is surely a public good, the public sector often won't allocate the right amount of funds for the right kinds of research.

Fighting Poverty Many government programs aim to help people who are economically disadvantaged. TANF (Temporary Assistance for Needy Families, sometimes called welfare) provides temporary income support for poor families with children. SNAP (Supplemental Nutrition Assistance Program, formerly called foods stamps) subsidizes food purchases for low-income households, and Medicaid provides them with medical care. The EITC (Earned Income Tax Credit) provides tax rebates for those who work at low-wage jobs. These and other antipoverty programs are financed by taxes on people with higher incomes.

Economists debate what role the government should play in fighting poverty. For now, note an important argument: Advocates of antipoverty programs sometimes claim that fighting poverty is a public good. Even if everyone prefers living in a society without poverty, fighting poverty is not a "good" that private actions will adequately provide.

To see why, suppose someone tried to organize a group of wealthy individuals to try to eliminate poverty. They would be providing a public good. This good would not be rival in consumption: One person's enjoyment of living in a society without poverty would not reduce anyone else's enjoyment of it. The good would not be excludable: Once poverty is eliminated, no one can be prevented from taking pleasure in this fact. As a result, there would be a tendency for people to free ride on the generosity of others, enjoying the benefits of poverty elimination without contributing to the cause.

Because of the free-rider problem, fighting poverty through private charity will probably not be sufficient. Yet government action can solve this problem. Taxing the wealthy to raise the living standards of the poor can potentially make everyone better off. The poor are better off because they now enjoy a higher standard of living, and those paying the taxes are better off because they enjoy living in a society with less poverty.

Case Study

Are Lighthouses Public Goods?

Some goods can switch between being public and private goods depending on the circumstances. For example, a fireworks display is a public good in a town with many residents. Yet if performed at Walt Disney World, a fireworks display is more like a private good because visitors to the park pay for admission.

SIEGFRIED LAYDA/THE IMAGE BANK/GETTY IMAGES

What kind of good is this?

Lighthouses have long been used as examples of public goods. They mark specific locations along the coast so passing ships can avoid treacherous waters. The benefit that a lighthouse provides to ship captains is neither excludable nor rival in consumption, so each captain has an incentive to free ride by using the lighthouse to navigate without paying for the service. Because of this free-rider problem, private markets usually fail to provide the lighthouses that captains need. As a result, most lighthouses today are operated by the government.

In some cases, however, lighthouses have been closer to private goods. On the coast of England in the 19th century, for example, some lighthouses were privately owned and operated. Instead of trying to charge ship captains for the service, however, the owner of the lighthouse charged the owner of the nearby port. If the port owner did not pay, the lighthouse owner turned off the light, and ships avoided that port.

To decide whether something is a public good, determine who benefits and whether they can be excluded from using the good. A free-rider problem arises when the beneficiaries are numerous and exclusion of any one of them is impossible. If a lighthouse benefits many ship captains, it is a public good. If it primarily benefits a single port owner, it is more like a private good. ●

11-2c The Difficult Job of Cost–Benefit Analysis

So far, we have seen that the government provides public goods because the private market on its own will not produce an efficient quantity. Yet deciding that the government must play a role is only the first step. The government must then determine what kinds of public goods to provide and in what quantities.

Suppose that the government is considering a public project, such as building a new highway. To judge whether to go ahead, the government compares the benefits for all those who would use it with the costs of building and maintaining it. A team of economists and engineers might conduct a study, called a **cost–benefit analysis**, to estimate the total costs and benefits of the project to society as a whole.

cost–benefit analysis a study that compares the costs and benefits to society of providing a public good

Cost–benefit analysts have a tough job. Because the highway will be available to everyone free of charge, there is no price with which to judge its value. Simply asking people how much they would value the highway is not reliable: Quantifying benefits is difficult using the results from a questionnaire, and respondents have little incentive to tell the truth. Those who would use the highway have an incentive to exaggerate the benefit they receive to get it built. Those who would be harmed have an incentive to exaggerate the costs so they can stop the project.

The efficient provision of public goods is, therefore, intrinsically more difficult than it is for private goods. When buyers of a private good enter a market, they reveal the value they place on the good through the prices they are willing to pay. At the same time, sellers reveal their costs with the prices they are willing to accept. The equilibrium is an efficient allocation of resources because it reflects all this information. By contrast, cost–benefit analysts do not have price signals to observe when evaluating whether the government should provide a public good and how much to provide. Their findings on the costs and benefits of public projects are rough approximations at best.

How Much Is a Life Worth?

Imagine that you have been elected to serve as a member of your local town council. The town engineer comes to you with a proposal: The town can spend $10,000 to install and operate a traffic light at an intersection that now has only a stop sign. The benefit of the traffic light is increased safety. Based on data from similar intersections, the engineer estimates that the traffic light would reduce the risk of a fatal accident over the lifetime of the traffic light from 1.6 to 1.1 percent. Should you spend the money for the new light?

To answer this question, you turn to cost–benefit analysis. But you quickly run into an obstacle: The costs and benefits must be measured in the same units if you are to compare them meaningfully. The cost is measured in dollars, but the benefit—the possibility of saving a person's life—is not directly monetary. To make your decision, you have to put a dollar value on a human life.

At first, you may say that a human life is priceless. After all, there is probably no amount of money that would induce you to give up your life or that of a loved one. This suggests that a human life has an infinite dollar value.

In cost–benefit analysis, however, this answer leads to nonsensical results. If we place an infinite value on human life, we should place traffic lights on every street corner, and we should all drive large cars loaded with the latest safety features. Yet traffic lights are not at every corner, and people sometimes choose to pay less for smaller cars without safety options such as side-impact airbags or antilock brakes. In both public and private decisions, people sometimes take risks to save money.

Once we have accepted the idea that a person's life has an implicit dollar value, how can we determine what that value is? One approach, sometimes used by courts to award damages in wrongful-death suits, is to look at the total amount that a person would have earned. Economists are often critical of this approach because it ignores other opportunity costs of losing one's life. It bizarrely implies that the life of a person who does not work has no value.

A better way to value human life is to look at the risks that people are voluntarily willing to take and how much they must be paid for taking them. For example, mortality risk varies across jobs. Construction workers in high-rise buildings face a greater risk of death on the job than office workers do. By comparing wages in risky and less risky occupations, controlling for education, experience, and other determinants of wages, economists can get some sense about what value people put on their own lives. Studies using this approach conclude that the value of a human life is about $10 million.

You can now respond to the town engineer. Because the traffic light reduces the risk of fatality by 0.5 percentage points, the expected benefit from installing the traffic light is 0.005 × $10 million, or $50,000. This estimate of the benefit exceeds the cost of $10,000, so you should approve the project. ●

QuickQuiz

3. Which is a public good?
   a. residential housing
   b. national defense
   c. restaurant meals
   d. fish in the ocean
4. Public goods are
   a. efficiently provided by market forces.
   b. underprovided in the absence of government.
   c. overused in the absence of government.
   d. a type of natural monopoly.
5. The three residents of Smallville are considering a fireworks display. Clark values this public good at $80, Lana at $50, and Pete (who dislikes fireworks) at −$30. Fireworks cost the town $120, or $40 per person. The efficient outcome is for the village
   a. to provide the public good because the median person values it more than its cost per person.
   b. to provide the public good because a majority of the residents value it more than its cost per person.
   c. to provide the public good because the total value of a majority exceeds the total cost.
   d. not to provide the public good because the total value of all residents is less than the total cost.

Answers are at the end of the chapter.

11-3 Common Resources

Tragedy of the Commons
a parable that illustrates why common resources are used more than is desirable from the standpoint of society as a whole

Common resources, like public goods, are not excludable: They are available free of charge to anyone who wants to use them. Common resources are, however, rival in consumption: One person's use of the common resource reduces other people's ability to use it. Thus, common resources give rise to a new problem: Once the good is provided, policymakers need to be concerned about how much it is used. This problem is best understood from a classic parable called the **Tragedy of the Commons**.

11-3a The Tragedy of the Commons

Of the many economic activities that take place in a small medieval town, one of the most important is raising sheep. Many families own flocks of sheep and support themselves by selling wool, which is used to make clothing.

As our story begins, the sheep spend much of their time grazing on the land surrounding the town, called the Town Commons. No family owns the land. Instead, the town residents own it collectively, and all the residents are allowed to graze their sheep on it. Collective ownership works well because land is plentiful. As long as everyone can get all the good grazing land they want, the Town Commons is not rival in consumption, and allowing residents' sheep to graze for free causes no problems. Everyone is happy.

As the years pass, the population grows, and so does the number of sheep grazing on the Town Commons. With a growing number of sheep, the land starts to lose its ability to replenish itself. Eventually, the land is grazed so heavily that it becomes barren. With no grass left on the Town Commons, raising sheep is impossible, and the once prosperous wool industry disappears. Many families lose their source of livelihood.

What causes this tragedy? Why do the shepherds allow the sheep population to grow so large that it destroys the Town Commons? The reason is that social and private incentives differ. Avoiding the destruction of the grazing land depends on the collective action of the shepherds. If they acted together, they could reduce the sheep population to a size that the Town Commons can support. Yet no single family has an incentive to reduce the size of its own flock because each flock represents only a small part of the problem.

In essence, the Tragedy of the Commons arises because of an externality. When one family's flock grazes on the common land, it reduces the quality of the land available for other families. Because people neglect this negative externality when deciding how many sheep to own, the result is an excessive number of sheep.

If the tragedy had been foreseen, the town could have solved the problem in various ways. It could have regulated the number of sheep in each family's flock, internalized the externality by taxing sheep, or auctioned off a limited number of sheep-grazing permits. That is, the medieval town could have dealt with overgrazing in the way that modern society deals with pollution.

In the case of land, however, there is a simpler solution. The town can divide the land among the town families. Each family can enclose its parcel of land with a fence and then protect it from excessive grazing. In this way, the land becomes a private good rather than a common resource. This occurred in England during the 17th century in what is known as the enclosure movement.

The Tragedy of the Commons teaches a general lesson: A person who uses a common resource diminishes other people's ability to enjoy it. Because of this negative externality, common resources tend to be used excessively. But the government can solve the problem by regulating or taxing the undesirable behavior. Alternatively, the government can sometimes turn the common resource into a private good.

This lesson has been known for thousands of years. The ancient Greek philosopher Aristotle pointed out the problem: "What is common to many is taken least care of, for all men have greater regard for what is their own than for what they possess in common with others."

11-3b Some Important Common Resources

There are many examples of common resources, leading to various government policies to mitigate the problem of overuse.

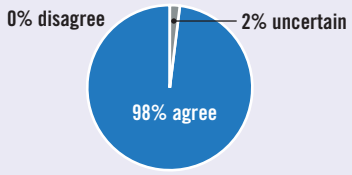
Clean Air and Water Chapter 10 showed that unfettered markets do not adequately protect the environment. Pollution is a negative externality that can be remedied with regulations or corrective taxes. This market failure can be seen

Ask the Experts

Congestion Pricing

"In general, using more congestion charges in crowded transportation networks—such as higher tolls during peak travel times in cities, and peak fees for airplane takeoff and landing slots—and using the proceeds to lower other taxes would make citizens on average better off."

What do economists say?



Source: IGM Economic Experts Panel, January 11, 2012.

as a common-resource problem. Clean air and water are common resources like open grazing land, and excessive pollution is like excessive grazing. Environmental degradation is a modern Tragedy of the Commons.

Congested Roads Roads can be either public goods or common resources. If a road is not congested, then one person's use does not affect anyone else. In this case, use is not rival in consumption, and the road is a public good. But in many urban areas, empty roads are an unattainable dream. If a road is congested, its use confers a negative externality. Whenever another vehicle drives on the road, it becomes more crowded, and others must drive more slowly. In this case, the road is a common resource.

One way for the government to address road congestion is to charge drivers a toll, which is, in essence, a corrective tax on the externality of congestion. Sometimes, as in the case of local roads, tolls are not a practical solution because the cost of collecting them is too high. But several major cities, including London and Stockholm, have found increasing tolls to be a very effective way to reduce congestion, and the idea is spreading.

Sometimes, congestion is a problem only at certain times of the day. If a bridge is heavily traveled only during rush hour, for instance, the congestion externality is largest then. The efficient solution is to charge higher tolls during rush hour. This would give drivers an incentive to alter their schedules, reducing traffic when congestion is greatest.

Another policy that responds to the problem of road congestion (discussed in the previous chapter) is the tax on gasoline. A higher gasoline tax increases the price of fuel, reduces the amount that people drive, and reduces road congestion. This tax is an imperfect solution to congestion because it affects other decisions besides the amount of driving on congested roads. In particular, the tax also discourages driving on uncongested roads, even though there is no congestion externality for these roads. And it misses congestion caused by electric cars.

Fish, Whales, and Other Wildlife Many species are common resources. Fish and whales, for instance, have commercial value, and anyone can go to the ocean and catch whatever is available. Each person has little incentive to maintain the species for the next year. Just as excessive grazing can ruin the Town Commons, excessive fishing and whaling can destroy commercially valuable marine populations.

Oceans remain one of the least regulated common resources. Two problems prevent an easy solution. First, many countries have access to the oceans, so any solution would require international cooperation among countries that hold different values. Second, because the oceans are so vast, enforcing any agreement is difficult. As a result, fishing rights have been a frequent source of tension among normally friendly countries.

Within the United States, various laws aim to manage fish and other wildlife resources. The government charges for fishing and hunting licenses, and it restricts the lengths of the fishing and hunting seasons. Fishermen are often required to throw back small fish, and hunters are permitted to kill only a limited number of animals. All these laws reduce the use of a common resource and help maintain animal populations.

Case Study

Why the Cow Is Not Extinct

Throughout history, many species of animals have been threatened with extinction. When Europeans first arrived in North America, more than 60 million buffalo roamed the continent. But buffalo hunting was so rampant that by 1900 the animal's population had fallen to about 400 before the government stepped in to protect what was left of the species. In some African countries today, elephants face a similar challenge, as poachers kill them for the ivory in their tusks.

Yet not all animals with commercial value face this threat. The beef cow is a valuable source of food, but no one worries that it will soon be extinct. The great demand for beef seems to ensure that the species will survive.

Why does the commercial value of ivory threaten the elephant, while the commercial value of beef protects the cow? The reason is that elephants are a common resource, while cows are a private good. Elephants roam freely without owners, and poachers kill as many as they can. Because poachers are numerous, each has little incentive to preserve the elephant population. By contrast, cattle live on ranches that are privately owned, and their owners have an incentive to maintain the cattle population.

Governments have tried to solve the elephant's problem in two ways. Some countries, such as Kenya, Tanzania, and Uganda, have made it illegal to kill elephants and sell their ivory. Yet these laws have been hard to enforce, and the battle between the authorities and poachers has become increasingly violent. Elephant populations have continued to dwindle. By contrast, other countries, such as Botswana, Malawi, Namibia, and Zimbabwe, have made elephants a private good by allowing people to kill elephants, but only those on their own property. Landowners have an incentive to preserve the species on their own land, and, as a result, some of those elephant populations have started to rise. With private ownership and the profit motive on its side, the African elephant might someday be as safe from extinction as the cow. ●

DOLORES M. HARVEY/SHUTTERSTOCK.COM

"Will the market protect me?"

QuickQuiz

6. Which of the following is an example of a common resource?
   a. residential housing
   b. national defense
   c. restaurant meals
   d. fish in the ocean

7. Common resources are
   a. efficiently provided by market forces.
   b. underprovided in the absence of government.
   c. overused in the absence of government.
   d. a type of natural monopoly.

8. The Mass Pike is a toll road that is congested only during rush hour. During other times of day, the use of the highway is not _____, so the efficient toll is _____.
   a. excludable; higher
   b. excludable; zero
   c. rival in consumption; higher
   d. rival in consumption; zero

Answers are at the end of the chapter.

In the News — Road Pricing

In 2021, President Biden proposed spending $2.25 trillion on infrastructure, including roads and bridges. Here, an economist argues that the nation might not get as much from this spending as it could.

How Federal Infrastructure Dollars Get Nickeled and Dimed

By Clifford Winston

Politicians and economists of all stripes have agreed on increased spending to improve the condition of U.S. infrastructure, but policymakers have failed for decades to act on this rare agreement.

President Joe Biden is planning a multitrillion-dollar infrastructure and jobs package to spur transformative change to the economy. Unfortunately, the infrastructure component of his plan will fail to significantly improve the nation's roads, bridges, and the like because it ignores the vast inefficiencies in current transportation policy that greatly reduce benefits from infrastructure spending.

Let me take you on the journey of a dollar of government spending intended to improve, for example, travel conditions on a highway. This dollar will have a long, perilous trip and encounter many dangers enroute that will divert it from its correct destination and take large, wasteful chunks out of it. By the time it reaches the wrong destination, it will fund much less than a dollar's worth of highway improvements. The dangers it encounters include inefficient road pricing and investment policy, inflated input and project costs, misallocation of highway revenues, and the slow adoption of technological innovations.

The trouble for our metaphorical dollar begins as soon as it starts trying to follow the signposts to its destination. Efficient pricing and investment serve that purpose by pointing to the amount and location of additional spending that would best benefit travelers. Efficient road pricing consists of charging motorists and truckers congestion tolls and pavement-wear fees to account for the costs they impose on other travelers by delaying them and damaging their vehicles, and for the costs of repairing and expanding the highway. Investments in a road, such as an additional lane, produce benefits, such as less delay, that are worth the cost. Efficient pricing and investment enable the road system to provide optimal travel conditions for a given level of expenditures. Moreover, they prevent waste by providing directions for how policymakers can improve road travel through additional expenditures that yield the greatest benefits. Current inefficient pricing and investment policies do not provide those directions, causing policymakers to waste money on projects that yield only small benefits.

Efficient prices for road users also serve two other important purposes. First, they fund efficient infrastructure investments. Biden wants to fund additional infrastructure spending by raising taxes on businesses. This approach is inefficient, and it may put his plan in political jeopardy. Second, efficient prices would reduce, and possibly eliminate, the highway budget deficit and the burden on general taxpayers.

Roads should be built, maintained and, when necessary, expanded at minimum cost and in a timely manner. In other words, projects should take the shortest route to completion without sacrificing quality. However, various regulations lead our

11-4 Conclusion: Property Rights and Government Action

This chapter and the previous one showed that markets do not adequately provide everything that a society needs. They don't ensure that the air we breathe is clean or that our country is defended from foreign aggressors. Instead, societies rely on the government to protect the environment and provide for the national defense.

The problems considered in these chapters arise in many different contexts, but they have a common theme. In each case, the market fails to allocate resources efficiently when something of value has no price attached to it. This happens when **property rights** are not well established, sometimes because they can't be. For example, although clean air and national defense are undeniably valuable, no private individual has the legal right to price them and profit from their use. A factory

dollar on a longer, more costly route, without increasing quality. State and federal (Davis-Bacon) regulations inflate wages and bloat the labor force hired to manage and complete highway projects. "Buy American" requirements for construction materials used in federal-aid highway projects, such as bridge repairs, raise costs even though less-expensive foreign materials of comparable quality could have been used. Moreover, the permitting process, environmental regulations on highway design, and other factors greatly extend the time to complete highway projects.

Earmarks and demonstration projects have become a growing political cost to ensure that multiyear federal transportation bills are passed. Those pet projects, as well as highway funds that are allocated throughout the country by formula without regard for efficiency, take our dollar on side trips that often go nowhere.

Finally, our dollar could take fruitful shortcuts if policymakers adopt the latest technologies to improve highways' design characteristics and maintenance at lower costs and to enhance traffic safety. But state Departments of Transportation are slow to keep up with and implement new technologies and tend to award contracts based on the minimum bid, not on the technological sophistication of the contractor.

Thus, a dollar bill intended to improve the road system is repeatedly nickeled and dimed and arrives as small change at the wrong destination because it has taken the wrong direction, gone on a longer route, made pointless side trips, and missed helpful technological shortcuts. The cumulative lost and damaged dollar bills intended for infrastructure improvements amount to billions of dollars of wasted infrastructure spending every year. Biden's new infrastructure plan, if passed, would produce the same outcome because its dollars would take the same perilous journey that previous dollars have taken.

The perilous journey persists because it has yet to create political risks that give policymakers the incentive to make infrastructure policy more efficient. But that may change. As Quentin Karpilow and I point out in our Brookings book, *Autonomous Vehicles: The Road to Economic Growth?*, autonomous vehicles are a watershed moment in the development of transportation. AVs promise not only to vastly improve road travel and generate huge benefits to travelers, shippers, and delivery companies, but also to benefit major sectors of the U.S. economy by reducing congestion and virtually eliminating vehicle accidents. However, autonomous vehicles will not live up to their promise if the inefficient infrastructure policies that have created the perilous journey and compromised road travel are not reformed.

The whole world will be watching as countries, cities, and states compete intensely to successfully develop and adopt autonomous vehicles. Policymakers who weaken their jurisdiction's AV operations by allowing the infrastructure dollar's perilous journey to persist should be prepared to incur significant political costs. Biden should take the first step to avoid those costs by making infrastructure spending more efficient instead of simply increasing it. ■

Questions to Discuss

1. Do you favor the increased use of tolls and fees for the users of roads and bridges? Why or why not?
2. Do you think the general public would embrace the increased use of tolls and fees for the users of roads and bridges? Why or why not?

Mr. Winston is an economist at the Brookings Institution.

Source: *Barron's*, March 24, 2021.

pollutes too much because no one charges the factory for the pollution it emits. The market does not provide for national defense because no one can charge those who are defended for the benefit they receive.

When the absence of property rights leads to a market failure, the government may be able to solve the problem. Sometimes, as in the sale of pollution permits, the government can help define property rights and unleash market forces. Other times, as in restricted hunting seasons, the government can regulate private behavior. Still other times, as in the provision of national defense, the government can use tax revenue to supply a good that the market fails to supply. In each of these cases, if the policy is well-planned and well-run, it can make the allocation of resources more efficient and raise economic well-being.

Chapter in a Nutshell

- Goods differ in whether they are excludable and whether they are rival in consumption. A good is excludable if it is possible to prevent someone from using it. It is rival in consumption if one person's use of the good reduces others' ability to use the same unit of the good. Markets work best for private goods, which are both excludable and rival in consumption. Markets do not work as well for other types of goods.
- Public goods are neither excludable nor rival in consumption. Examples of public goods include fireworks displays, national defense, and the discovery of fundamental knowledge. Because people are not charged for their use of the public good, they have an incentive to free ride, making private provision of the good infeasible. Governments can improve the allocation of resources by providing public goods and determining the quantity of each good with cost–benefit analysis.
- Common resources are not excludable but are rival in consumption. Examples include common grazing land, clean air, and congested roads. Because people are not charged for their use of common resources, they tend to use them excessively. Governments can remedy this problem using various methods, such as regulations and corrective taxes, to limit the use of common resources.

Key Concepts

excludability, p. 212
rivalry in consumption, p. 212
private goods, p. 212
public goods, p. 213
common resources, p. 213
club goods, p. 213
free rider, p. 214
cost–benefit analysis, p. 217
Tragedy of the Commons, p. 218

Questions for Review

1. Explain what is meant by a good being "excludable." Explain what is meant by a good being "rival in consumption." Is a slice of pizza excludable? Is it rival in consumption?
2. Define and give an example of a public good. Can the private market provide this good on its own? Explain.
3. What is cost–benefit analysis of public goods? Why is it important? Why is it hard?
4. Define and give an example of a common resource. Without government intervention, will people use this good too much or too little? Why?

Problems and Applications

1. Think about the goods and services provided by your local government.
   a. Using the categories in Figure 1, classify each of the following goods, explaining your choice:
      - police protection
      - snow plowing
      - education
      - rural roads
      - city streets
   b. Why do you think the government provides items that are not public goods?
2. Both public goods and common resources involve externalities.
   a. Are the externalities associated with public goods generally positive or negative? Is the free-market quantity of public goods generally greater or less than the socially efficient quantity? Cite examples in your answer.
   b. Are the externalities associated with common resources generally positive or negative? Is the free-market use of common resources generally greater or less than the socially efficient use? Cite examples in your answer.
3. Fredo loves watching *Downton Abbey* on his local public TV station, but he never sends any money to support the station during its fund-raising drives.
   a. What name do economists have for people like Fredo?
   b. How can the government solve the problem caused by people like Fredo?

c. Can you think of ways the private market can solve this problem? How does the option of cable TV alter the situation?

4. Wireless high-speed Internet is provided for free in the airport of the city of Communityville.
   a. At first, only a few people use the service. What type of good is this and why?
   b. Eventually, as more people find out about the service and start using it, the speed of the connection begins to fall. Now what type of good is the wireless Internet service?
   c. What problem might result and why? What is one possible way to correct this problem?

5. Four roommates are planning to spend the weekend in their dorm room watching old movies, and they are debating how many to watch. Here is their willingness to pay for each film:

	Dwayne	Javier	Salman	Chris
First film	$7	$5	$3	$2
Second film	6	4	2	1
Third film	5	3	1	0
Fourth film	4	2	0	0
Fifth film	3	1	0	0

   a. Within the dorm room, is the showing of a movie a public good? Why or why not?
   b. If it costs $8 to stream a movie, how many movies should the roommates stream to maximize total surplus?
   c. If they choose the optimal number from part (b) and then split the cost of streaming the movies equally, how much surplus does each person obtain from watching the movies?
   d. Is there any way to split the cost to ensure that everyone benefits? What practical problems does this solution raise?
   e. Suppose they agree in advance to choose the efficient number and to split the cost of the movies equally. When Dwayne is asked his willingness to pay, will he have an incentive to tell the truth? If so, why? If not, what will he be tempted to say?
   f. What does this example teach you about the optimal provision of public goods?

6. Some economists argue that private firms will not undertake the efficient amount of basic scientific research.
   a. Explain why this might be so. In your answer, classify basic research in one of the categories shown in Figure 1.
   b. What sort of policy has the United States adopted in response to this problem?
   c. It is often argued that this policy increases the technological capability of American producers relative to that of foreign firms. Is this argument consistent with your classification of basic research in part (a)? (Hint: Can excludability apply to some potential beneficiaries of a public good and not others?)

7. Two towns, each with three residents, are deciding whether to put on a fireworks display to celebrate the New Year. Fireworks cost $360. In each town, some people enjoy fireworks more than others.
   a. In the town of Bayport, each of the residents values the public good as follows:

Frank	$50
Joe	$100
Callie	$300

   Would fireworks pass a cost–benefit analysis? Explain.
   b. The mayor of Bayport proposes to decide by majority rule and, if the fireworks referendum passes, to split the cost equally among all residents. Who would vote in favor, and who would vote against? Would the vote yield the same answer as the cost–benefit analysis?
   c. In the town of River Heights, each of the residents values the public good as follows:

Nancy	$20
Bess	$140
Ned	$160

   Would fireworks pass a cost–benefit analysis? Explain.
   d. The mayor of River Heights also proposes to decide by majority rule and, if the fireworks referendum passes, to split the cost equally among all residents. Who would vote in favor, and who would vote against? Would the vote yield the same answer as the cost–benefit analysis?
   e. What do you think these examples say about the optimal provision of public goods?

8. There is often litter along highways but rarely in people's yards. Provide an economic explanation for this fact.

9. Many transportation systems, such as the Washington, D.C., Metro (subway), charge higher fares during rush hours than during the rest of the day. Why might they do this?

10. High-income people are willing to pay more than lower-income people to avoid the risk of death. For example, they are more likely to pay for cutting-edge safety features on cars. Do you think cost–benefit analysts should take this fact into account when evaluating public projects? Consider, for instance, a rich town and a poor town, both of which are considering the installation of a traffic light. Should the rich town use a higher dollar value for a human life in making this decision? Why or why not?

QuickQuiz Answers

1. **a** 2. **b** 3. **b** 4. **b** 5. **d** 6. **d** 7. **c** 8. **d**

Chapter

12

The Economics of Healthcare

All of us would like to lead long, healthy lives. And given the choice, we would prefer to do so without having to endure the surgeon's scalpel, the nurse's needle, or the dentist's drill. Yet good health rarely comes so easily. Achieving a long, healthy life often requires the use of scarce resources, and that makes it, at least in part, an economic problem. More than one out of every six dollars spent in the U.S. economy goes to some form of healthcare, including spending on physicians, nurses, dentists, hospitals, pharmaceutical drugs, and medical research scientists. Understanding the modern economy requires an appreciation of the special economics of healthcare.

This chapter begins by examining the economic forces that shape the healthcare system. The standard model of supply and demand, in which prices guide buyers and sellers to an efficient allocation of resources, explains how large parts of the world work, but the market for healthcare deviates from this benchmark in many ways. These deviations often call for government policies to ensure that healthcare resources are allocated efficiently and equitably. And in most nations, governments are deeply involved in healthcare

GAUDILAB/SHUTTERSTOCK.COM, LECHATNOIR/E+/GETTY IMAGES

markets. That involvement increased greatly in 2020 and 2021 when governments around the world needed to take emergency action to curb the fatal effects of the novel coronavirus pandemic.

The second half of this chapter looks at some of the key facts that describe the U.S. healthcare system. The system today is very different from what it was fifty years ago, and it is also different from the systems of other nations. Recognizing these differences is important for understanding the healthcare systems that exist now as well as for imagining those that could exist in the future.

The proper scope of government intervention in healthcare, and even the core principles for organizing the healthcare system, remain topics of heated discussion. This introduction to the economics of healthcare will give you a better grounding in this great debate.

12-1 The Special Characteristics of the Market for Healthcare

The standard theory of how markets work is the model of supply and demand, which we studied in Chapters 4 through 7. That model has several notable features:

1. The main interested parties are the buyers and sellers in the market.
2. Buyers are good judges of what they get from sellers.
3. Buyers pay sellers directly for the goods and services being exchanged.
4. Market prices are the primary mechanism for coordinating the decisions of market participants.
5. The invisible hand on its own leads to an efficient allocation of resources.

For many goods and services, this model offers a reasonably good description.

Yet none of the five features of the standard model reflects what occurs in the market for healthcare. The healthcare market has consumers (patients) and producers (doctors, nurses, etc.). But various factors complicate the analysis of their interactions. In particular:

1. Third parties—insurers, governments, and unwitting bystanders—often have an interest in healthcare outcomes.
2. Patients often don't know what they need and cannot evaluate the treatment they are getting.
3. Healthcare providers are often paid not by the patients but by private or government health insurance.
4. The rules established by these insurers, more than market prices, determine the allocation of resources.
5. In light of these issues, the invisible hand can't work its magic, so the allocation of resources can be highly inefficient.

Healthcare is not the only good or service in the economy that departs from the standard model of supply, demand, and the invisible hand. (Recall our discussions of externalities in Chapter 10 and public goods in Chapter 11.) But healthcare may be the most important one that departs radically from this benchmark. Examining the special features of this market is a good starting point for understanding why the government plays a large role in the provision of healthcare and why health policy is often complex and vexing.

12-1a Externalities Galore

Chapter 10 showed that market outcomes may be inefficient when there are externalities. To recap: An **externality** arises when a person engages in an activity that influences the well-being of a bystander but neither pays nor receives compensation for that effect. If the impact on the bystander is adverse, it is a **negative externality**. If it is beneficial, it is a **positive externality**. In the presence of externalities, society's interest extends beyond the well-being of buyers and sellers in the market to include the well-being of bystanders who are affected indirectly. Because buyers and sellers neglect the external effects of their actions when deciding how much to demand or supply, the externality can render the unregulated market outcome inefficient.

This general conclusion is crucial for understanding healthcare, where externalities are rampant. These externalities can call for government action to remedy the market failure.

Take vaccines, for example. If Vicki vaccinates herself against a disease, she is less likely to catch it and become a carrier who infects other people. In the language of economics, Vicki's action conveys a positive externality. If getting a jab has some cost, either in money, time, discomfort, or risk of adverse side effects, too few people will do it because they may not fully take into account the positive externalities when weighing the costs and benefits. The government can remedy this problem by requiring vaccination or encouraging it through media campaigns and incentives. The government can also increase the supply of vaccines by subsidizing their development, manufacture, and distribution.

Another example of a healthcare externality concerns medical research. When a physician or scientist figures out a new way to treat an ailment, that information enters society's pool of medical knowledge. The benefit to other physicians and patients is a positive externality. Without government intervention, there will be too little research.

Government responds to this problem in many ways. Sometimes, it grants a researcher a patent, as is the case with new pharmaceutical drugs. The patent provides an incentive for research by allowing the researcher to profit from a temporary monopoly. The patent is said to internalize the externality. Yet this approach is not perfect because the monopoly price is higher than the marginal cost of production. As we will see in Chapter 16, the high monopoly price reduces consumption of the patented treatment, leading to inefficiency as measured by the deadweight loss. Moreover, the high price may be particularly hard on patients with lower incomes.

Healthcare externalities sometimes point policymakers in conflicting directions. Again, consider vaccines. Because research into new vaccines conveys a positive externality, the government grants patent protection. But once a vaccine is devised, the high price charged by the patent holder slows its dissemination, undermining the positive externality from vaccination. One possible solution to this conflict, advocated by the economist Michael Kremer, is for the government to buy out the patent holder. This approach could put the innovation in the public domain, making it more widely available while still providing incentives for research.

Another way the government can deal with the positive externality from medical research is to subsidize research directly—and indeed it does. The annual budget of the National Institutes of Health, which funds medical research, is over $40 billion, or about $130 per person. This policy requires taxation to raise the necessary funds, and most taxation entails deadweight losses of its own. But if the externalities from the funded research exceed the cost of the research, including the deadweight losses, overall welfare can increase.

Case Study

Vaccine Hesitancy

The disease of measles illustrates the externalities central to health economics. Measles is highly contagious and dangerous, especially for children. Before a vaccine was developed in 1963, between 3 and 4 million people in the United States caught the disease every year. Tens of thousands required hospitalization, and many hundreds died.

Fortunately, immunization proved effective. By 2000, measles was largely eliminated as a health risk in the United States. During the subsequent decade, only about 60 cases were reported per year. The measles vaccine was one of the great success stories of modern medicine. In the language of economics, the research that created the vaccine had large positive externalities both in the United States and around the world.

The elimination of the disease, however, had some unintended consequences. In particular, it reduced the vigilance of some parents in ensuring that their children were vaccinated. Their decision to forgo vaccination was partly based on the belief—debunked among scientists but persistent among some segments of the public—that vaccines entail a significant risk of adverse side effects. Some claimed a religious basis for their decision not to vaccinate. Because of declining immunization, measles started to become more common again. There were 372 reported cases in 2018 and more than 1,000 in 2019.

Policymakers struggled with how to respond. In several states, lawmakers strengthened requirements that all children be vaccinated. Exemptions were allowed for valid medical reasons but not for religious or philosophical beliefs. Critics of these laws viewed them as a violation of personal liberty. Supporters viewed them as supporting the community's need to deal with the externalities inherent in infectious diseases.

A similar issue arose during the COVID-19 pandemic of 2020–2021. Soon after vaccines for this new virus were developed, they became widely and freely available. According to medical experts, eradicating the disease would require vaccinating most of the population. But because the vaccines were new, many people were hesitant. Some economists proposed paying people to get vaccinated—a Pigovian subsidy to internalize the externality and quicken the end of the pandemic. The proposal was not widely adopted at the national level, but some states and localities experimented with such incentives. In New York City, for example, those who got vaccinated were offered a voucher for a burger at Shake Shack. ●

12-1b The Difficulty of Monitoring Quality

In most markets, consumers know what they want, and after a transaction is completed, they can judge whether they are happy with what they got. Healthcare is different. When you get sick, you may not know which treatment is best. You rely on the advice of a physician who has years of specialized training. And even with hindsight, you cannot reliably judge whether the treatment offered to you was the right one. Sometimes state-of-the-art medicine fails to improve a patient's health. And given the restorative power of the human body, the wrong treatment can sometimes appear to work.

The inability of healthcare consumers to monitor the quality of the product they are buying leads to government regulations. Most importantly, the government requires physicians, dentists, nurses, and other health professionals to have licenses to practice. These licenses are granted only after an individual attends an approved school and passes rigorous tests. Those caught practicing without a license can be imprisoned. Similarly, the Food and Drug Administration (FDA) oversees the testing and release of new pharmaceutical drugs to ensure they are safe and effective.

In addition to government regulation, the medical profession monitors itself by accrediting medical schools, promoting best practices, and establishing norms of professional behavior. A physician's advice is supposed to be based entirely on the patient's best interest, not the physician's personal gain. When patients accept this advice, they rely on a degree of trust, which may be fostered by long-term relationships between doctor and patient. But conflict-of-interest problems exist in healthcare as they do elsewhere in society, and this trust is sometimes breached.

While public and private regulations of healthcare have many benefits for patients, they also have some drawbacks. For example, some economists have argued that there are too many hurdles to opening new medical schools. They suggest that the medical profession acts like a monopoly: By restricting the number of doctors, it drives up doctors' salaries and consumers' healthcare costs. Other economists have argued that the FDA is too slow in approving new drugs. Some patients who might have benefited from experimental treatments are forced to go without them. The proper balance between protecting public safety and giving people the freedom to make their own healthcare decisions is a subject of ongoing debate.

12-1c The Insurance Market and Its Imperfections

Because people don't know when they are going to get sick or what kind of medical treatments they will need, spending on healthcare is unpredictable. This uncertainty, along with people's responses to it, helps to explain why we have the health institutions that we do.

The Value of Insurance Most people exhibit **risk aversion**. That is, they dislike uncertainty. Imagine that you could choose between receiving $100,000 with certainty and flipping a coin to receive $50,000 or $150,000 with a 50–50 probability. The two options offer the same average amount, but the second is riskier. If you prefer the certain $100,000, you are risk averse.

risk aversion
a dislike of uncertainty

A similar choice arises from the randomness of health spending. Suppose that some disease affects 2 percent of the population and that everyone is equally likely to be stricken. Treatment costs $30,000 per patient. In this case, for every 100 people, 2 will get the disease, resulting in a total healthcare bill of $60,000. The population's average cost of healthcare is $60,000/100, or $600.

Here's the key point: People who are risk averse would prefer to pay $600 with certainty over a 2 percent chance of having to pay $30,000. Insurance gives people this option.

The way insurance generally works is that a person facing a risk pays a fee (called a **premium**) to an insurance company, which in return agrees to accept all or part of the risk. There are many types of insurance. Car insurance covers the risk that you get into an auto accident, fire insurance covers the risk that your house burns down, and health insurance covers the risk that you need expensive medical treatment. In our example, a health insurance company can charge a premium of $600 (or slightly more to make a profit) in exchange for promising to cover the cost of the $30,000 treatment for the 2 percent of its customers who get the disease.

Markets for insurance are useful in reducing risk, but two problems hamper their ability to do so fully and efficiently.

moral hazard
the tendency of a person who is imperfectly monitored to engage in dishonest or otherwise undesirable behavior

Moral Hazard The first problem that impedes the operation of insurance markets is **moral hazard**: When people have insurance to cover their spending on healthcare, they have less incentive to engage in behavior that will keep that spending to a reasonable level. For example, if patients don't have to pay for each visit to a

doctor, some may go whenever they experience minor symptoms (a runny nose, an achy finger). Similarly, physicians may be more likely to order tests of dubious value when they know an insurance company is picking up the tab.

Health insurance companies try to reduce moral hazard by finding ways to induce people to act more responsibly. For instance, rather than picking up the entire cost of a visit to a physician, they may charge patients **co-pays** of, say, $20 per visit to deter unnecessary visits. Insurance companies may also have strict rules about the circumstances under which they will cover the cost of certain tests ordered by physicians.

adverse selection
the tendency for the mix of unobserved attributes to become undesirable from the standpoint of an uninformed party

Adverse Selection The second problem that impedes the operation of insurance markets is **adverse selection**: If customers differ in relevant attributes (such as whether they have a chronic disease) and those differences are known to them but not to insurers, the mix of people who buy insurance may be especially expensive to insure. In particular, people with hidden conditions are more likely to buy health insurance than healthy people. As a result, for an insurance company to cover its costs, the price of health insurance must reflect the cost of a sicker-than-average person. That price may be high enough that some people with average health may forgo buying insurance. As people drop coverage, the insurance market fails to achieve its purpose of reducing the financial risk from illness.

Even worse, adverse selection can lead to a phenomenon called the **death spiral**. Suppose that insurance companies must charge everyone the same price. It might seem to make sense for a company to base the price of insurance on the health characteristics of the average person. But after it does so, the healthiest people may decide that insurance is not worth the cost and drop out of the insured pool. With a sicker-than-expected group of customers, the company has higher costs and must therefore raise the price of insurance. The price increase then induces the next healthiest group of people to drop insurance coverage, driving up the cost and price again. As this process continues, more people drop coverage, the insured pool gets less healthy, and the price keeps rising. In the end, the insurance market may disappear.

The problem of adverse selection has been central in the debate over health policy. For example, the Affordable Care Act (signed by President Obama in 2010 and often called "Obamacare") prevented health insurance companies from charging more to cover people with pre-existing medical conditions. This rule was enacted to help people with ongoing medical problems, but it was also a recipe for adverse selection: People with pre-existing conditions could view insurance as a better deal than those without them and, therefore, might be more likely to buy health insurance. Healthy people would have an incentive to wait until they got sick before buying insurance.

Lawmakers were aware of this problem. To combat it, the Affordable Care Act required **all** Americans to buy health insurance and imposed a financial penalty on those who did not. (It also gave subsidies to help households with lower incomes afford insurance.) The goal of the mandate was to increase the number of healthy people buying insurance, thereby reducing the problem of adverse selection and lowering the cost of insurance. As more healthy people entered the insurance market, those without pre-existing conditions would, in effect, subsidize those with them.

Analysts disagree on the extent to which the Affordable Care Act has improved the healthcare market. Many people (presumably healthy ones) chose to remain uninsured and pay the penalty, leading some to suggest that the penalties for not purchasing insurance were too small to prevent a significant amount of adverse selection. Yet the law has increased the number of people with health insurance, which was one of its main goals. For the population under age 65 (when Medicare

eligibility starts), the share without insurance fell from 18 percent in 2010 to 11 percent in 2020.

The mandate to buy health insurance, however, has been politically divisive. Its critics argue that penalizing people for not buying something is too great an infringement on personal liberty. At the end of 2017, President Donald Trump signed a tax bill that included an elimination of the health insurance mandate. Eliminating the mandate so far has not led to the beginning of a death spiral, perhaps because most people appreciate the value of insurance and may not need to be required to buy it.

12-1d Healthcare as a Right

Typically, when some people don't buy a good or service, the outcome is not a major problem for society. For example, suppose that admission to a water park becomes expensive and people with lower incomes choose other forms of entertainment. Some may lament that the water park is not enjoyed more widely, but few would argue this amounts to a great injustice.

Healthcare is different. When people get sick, it seems wrong to deny them treatment because they have low incomes. Healthcare, unlike admission to a water park, is widely viewed as a human right. This judgment goes beyond the traditional scope of economics, but the study of healthcare economics needs to take it into account.

In some ways, healthcare is like food: It is essential to survive. Government steps in when needed to ensure that everyone can get the essentials. For example, in the United States, resources for food are provided by the Supplemental Nutrition Assistance Program (SNAP), formerly known as the Food Stamp Program. There is, however, an important difference between food and healthcare. In recent decades, the price of food has risen more slowly than incomes, so affording an adequate diet has required a declining share of the typical household's budget. By contrast, because the cost of state-of-the-art healthcare has risen rapidly, affording it has required an increasing share of the typical household's budget.

The judgment that healthcare is a right, along with its rising cost, has led to a large role for the government. In many nations, such as Canada and England, the government runs the healthcare systems, financed mostly by taxes. These are sometimes called **single-payer** systems because one entity—the government's health service—pays all the bills.

In the United States, most people have private health insurance, often through their employers, but the government still has a large role. Medicare provides health insurance for those 65 and older, Medicaid provides health insurance for those with low incomes, the Veterans Health Administration offers healthcare to former members of the military, and the Affordable Care Act regulates the market for private health insurance and gives insurance subsidies to many households with lower incomes. Whether these programs can be improved and, if so, how, remains a topic of debate. But there is little doubt that, with healthcare often viewed as a human right, the government will continue to play a large role in the healthcare system.

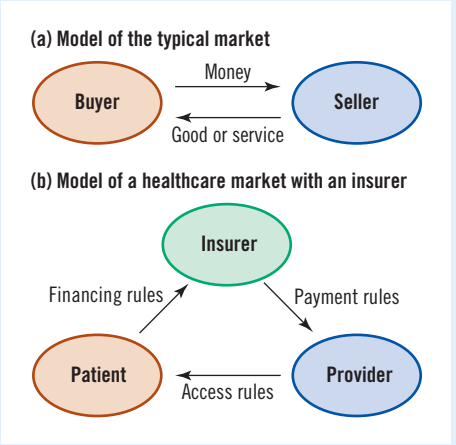
12-1e The Rules Governing the Healthcare Marketplace

The importance of health insurance, whether provided by private companies or the government, requires that the market for healthcare work differently than most other markets in the economy. Most markets, such as the market for ice cream, look like panel (a) of Figure 1. They consist of buyers and sellers. Sellers offer a good or service at the market price. Buyers who want the item simply have to offer up the right amount of money to purchase it. Exchanges are made, and soon, the sellers are counting their profits, and the buyers are enjoying their ice-cream cones.

Figure 1

How an Insurer Changes a Market

In a typical market, shown in panel (a), sellers deliver a good or service to buyers, who pay sellers a market-determined price. In the healthcare market, shown in panel (b), providers deliver healthcare to patients, but providers are paid by insurers (either the government or private companies). This arrangement requires rules for financing, access, and payment.



The market for healthcare looks more like panel (b) of Figure 1. Providers (the sellers of medical services) are not paid directly by patients (the buyers). Instead, patients pay money to insurers in the form of either a premium (if the insurer is a private company) or taxes (if the insurer is the government). The insurers use this money to compensate the providers, who in turn provide medical services to the patients.

This process requires three sets of rules to guide behavior. The first set determines the financing—that is, who pays for the insurance and how much they pay. If the insurer is the government, the payment for healthcare is part of the tax system. If the insurer is a private company, healthcare is financed by the premiums people pay for their coverage. The premium is set in the insurance market, which (like other markets) bases price on costs. In many cases, however, state and federal governments regulate the market for private insurance. For example, they may limit the extent to which companies can charge different prices based on age, gender, and pre-existing conditions. Thus, even when healthcare financing occurs between a patient and a private insurer, it is still shaped by government policy.

The second set of rules determines patients' access to healthcare. Because insured patients do not pay the full cost of each medical service they consume, there is the possibility of overuse. To mitigate the moral hazard problem, the insurer (whether the government or a private company) tries to limit access to when it makes sense. To do so, the insurer establishes rules that ration the use of medical services based on estimated costs and benefits. For example, patients may be able to get a routine check-up no more than once a year, may have access only to certain doctors with whom the insurer has contractual arrangements, or may need a referral from a general practitioner before seeing a more expensive specialist. Such access rules can be annoying, but they are necessary because, once people have insurance to pick up the cost, market prices no longer give them accurate signals about how to allocate scarce resources.

The third set of rules determines the payments from insurers to providers. These rules establish what treatment and care an insurer will pay for and how much it will pay. Insurers may deem some treatments too expensive, too experimental, or insufficiently valuable to pay for them at all. In such cases, providers will often not offer patients the services. Sometimes, providers will offer the services only

if the patient pays the full cost of the treatment (as is often the case with cosmetic procedures). The healthcare market then reverts from panel (b) in Figure 1 to the more typical market in panel (a).

The rules regarding financing, access, and payment together shape a healthcare system. For nations with government-run systems, these rules are set by public policy. For nations with more private insurance, such as the United States, these rules are often set by insurance companies as they compete for customers, subject to government regulations.

QuickQuiz

1. The market for healthcare differs from most other markets because
   a. consumers often don't know what they need.
   b. producers are paid by insurers rather than consumers.
   c. third parties have an interest in outcomes.
   d. All of the above
2. People buy health insurance to reduce
   a. moral hazard.
   b. adverse selection.
   c. uncertainty.
   d. externalities.
3. Health insurance has access rules for patients to reduce
   a. moral hazard.
   b. adverse selection.
   c. uncertainty.
   d. externalities.
4. If healthy people are less likely to buy health insurance, the market for health insurance experiences the problem of
   a. moral hazard.
   b. adverse selection.
   c. uncertainty.
   d. externalities.

Answers are at the end of the chapter.

12-2 Key Facts about the U.S. Healthcare System

Now that we understand the main economic forces at work, let's look at some data that describe the U.S. healthcare system. We first examine what the system provides, as measured by how long people live. We then see how much it costs, how much other nations pay, and how Americans pay the bill.

12-2a People Are Living Longer

Let's start with some good news: People are living much longer today than they did a century ago. Figure 2 shows life expectancy over time in the United States. Life expectancy measures how long people born today would live, on average, if they faced current mortality rates at every age. You can see that it has increased substantially over time. In 1900, life expectancy was just 47.3 years. It increased to 68.2 years by 1950 and to 78.8 years by 2019.

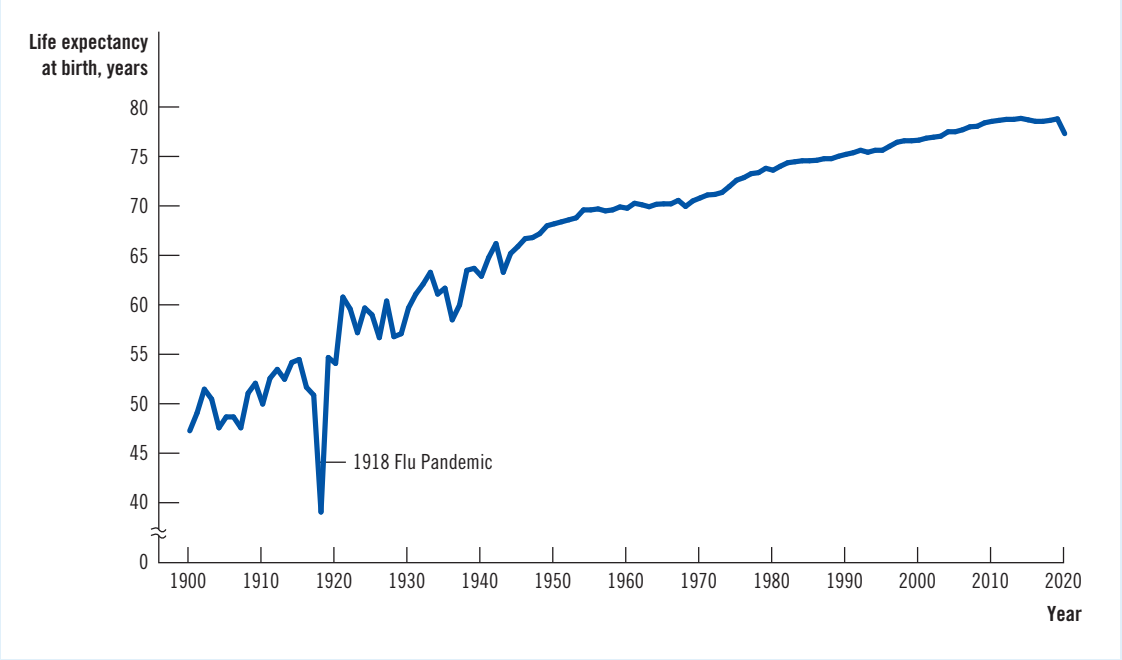
To be sure, temporary setbacks occur from time to time. The flu pandemic of 1918 is the most dramatic example, when U.S. life expectancy fell by more than 10 years. In addition, as World War II raged, life expectancy declined by 2.9 years from 1942 to 1943. More recently, the opioid epidemic curtailed many life spans in the United States, slowing the advance in life expectancy in the mid-2010s. And in 2020, the COVID-19 pandemic cost hundreds of thousands of lives, reducing U.S. life expectancy by 1.5 years—a tragic development that was expected to reverse over the next couple of years. Yet despite these setbacks, the long-term trend is highly positive.

Figure 2

U.S. Life Expectancy

The number of years a person can expect to live has increased substantially over long periods of time.

Source: Centers for Disease Control and Prevention.



A large part of the increase in life expectancy comes from a decline in infant mortality. At the beginning of the 20th century, about 10 percent of children died before the age of one. Today, the infant mortality rate is less than 0.6 percent.

Much of the credit for the historic increase in life expectancy goes to advances in medical technology. Physicians know more about how to prevent disease and how to treat medical problems when they arise. For example, highly effective vaccines, like those developed during the COVID-19 pandemic, have saved countless lives that would otherwise have been lost.

But other developments play a role as well. Improved sanitation—specifically, the availability of clean water and the adequate disposal of sewage—has reduced the spread of disease. So has the decline in smoking: Since 1960, the consumption of cigarettes per person has fallen by more than 50 percent. In addition, the rate of fatalities from car accidents is now half what it was in 1950, thanks to advances in automotive safety, such as seat belts and airbags.

12-2b Healthcare Spending Is a Growing Share of the Economy

Figure 3 shows healthcare spending in the United States as a percentage of GDP (a measure of the economy's total income). Health spending rose from 5 percent of GDP in 1960 to 18 percent in 2019, and there is no sign that the long-term trend is about to stop.

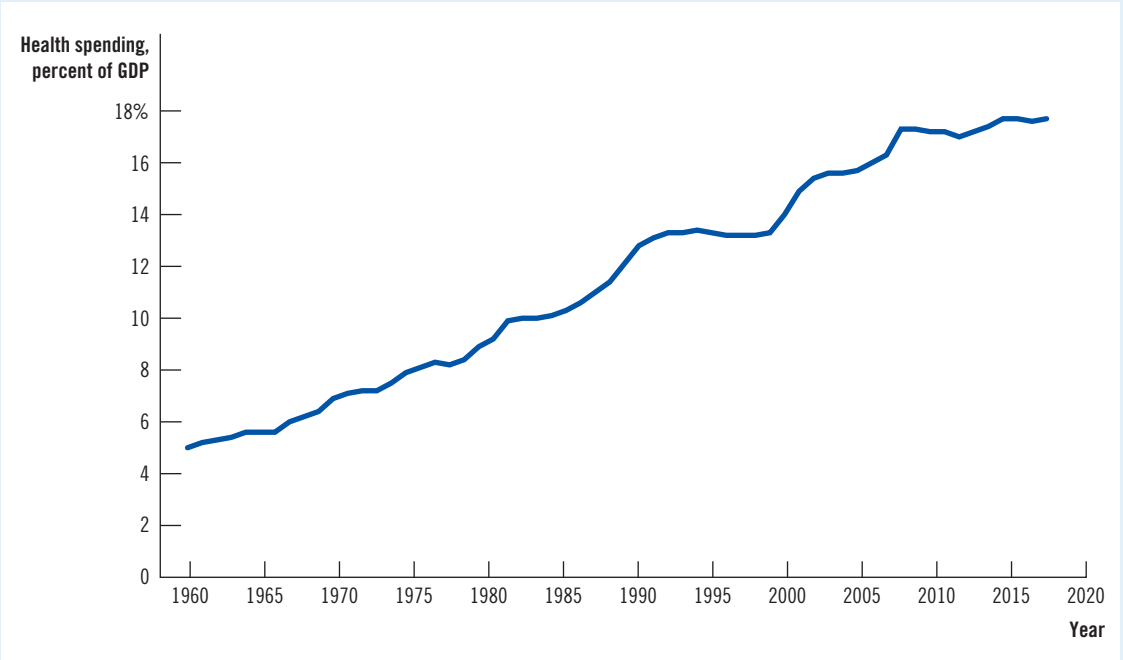
What explains this trend? Several forces are at work.

Figure 3

Health Spending as a Share of GDP: Changes over Time

The percentage of the U.S. national income devoted to healthcare has increased over time.

Source: Centers for Medicare & Medicaid Services.



First, most medical care, such as a doctor's visit, is a personal service, much like a barber's haircut or a teacher's lesson. The economist William Baumol pointed out long ago that for many providers of personal services, productivity does not change much over time. But as the rest of the economy experiences technological progress, labor productivity and overall wages increase. Those supplying personal services will come to expect rising wages along with the rest of the labor force. Yet without much productivity growth in those sectors, the only way to give these service providers higher wages is for the prices of their services (adjusted for overall inflation) to increase. In other words, when overall productivity is rising, a symptom of sectors with low productivity growth is rising costs and prices. This phenomenon is called **Baumol's cost disease**. It helps explain the rising price of many services, such as live entertainment, higher education, and healthcare. And if the demand for the services of such sectors is price inelastic, as it is for healthcare, spending on those services will increase as well.

Second, while there have been significant advances in medical technology, many of them, rather than reducing costs, have increased spending. In the past, physicians had little treatment for many diseases. Bed rest and wait-and-see (and leeches!) were sometimes the best they could offer. Today, there are more options. These new treatments extend and enhance the quality of life, but they are often expensive.

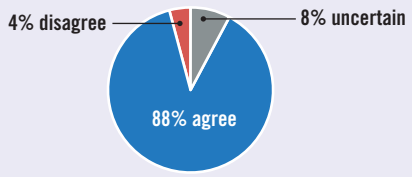
Third, changes in the population may have increased the demand for healthcare. In particular, birth rates have fallen. Fifty years ago, the average woman had about

Ask the Experts

Baumol's Cost Disease

"Because labor markets across different sectors are connected, rising productivity in manufacturing leads the cost of labor-intensive services—such as education and health care—to rise."

What do economists say?



Source: IGM Economic Experts Panel, May 16, 2017.

three children over her lifetime, compared with about two children today. This fall in birth rates, along with the increase in life expectancy, has altered the relative sizes of various age groups. The share of the U.S. population 65 years or older increased from 9 percent in 1960 to 17 percent in 2020. Because older people generally need more healthcare than the young, the aging of the population leads to greater healthcare spending.

Fourth, over time, society has become richer, and that change may have increased the share of spending on healthcare. Today, the average income per person, adjusted for inflation, is more than three times what it was in 1960. As incomes rise, people spend more on many things, but they don't increase spending on all items proportionately. How individuals choose to spend the extra income depends on preferences. For example, as incomes increase, the share of spending on food declines because the marginal value of consuming more calories declines rapidly. By contrast, the marginal value of enjoying additional years of life declines slowly, so as we get richer, we may spend a higher fraction of our budget on healthcare. In other words, healthcare may be a good with an income elasticity greater than one. Estimates based on international comparisons put this elasticity at about 1.3.

Given these four forces, the increasing share of health spending in the economy may be inevitable. In and of itself, higher healthcare spending is not necessarily a problem. But it does mean that the policy challenges in this sector will loom larger over time.

12-2c Healthcare Spending Is Especially High in the United States

Figure 4 shows healthcare spending as a percentage of GDP for seven major developed nations. One striking fact is that the United States spends an especially high fraction of its GDP on healthcare. Most developed nations spend 9 to 12 percent of GDP on healthcare, while the United States spends about 17 percent.

Critics of the U.S. healthcare system use this comparison to argue that the United States is uniquely inefficient. They point out that life expectancy is higher in some nations that pay less for healthcare, such as Canada, France, and Japan. They sometimes suggest that greater reliance on government rather than private health insurance, as is the case in most other nations, could lower costs without adversely affecting health outcomes. In particular, they say that administrative costs, plus the insurance companies' mark-up for profit, might be saved if the United States shifted to a government-financed single-payer system.

Defenders of the U.S. healthcare system accept that further reforms might reduce costs but say that reliable conclusions are hard to draw from the international comparisons. For example, the rate of obesity is higher in the United States than it is in the other six nations in Figure 4. Higher rates of obesity reduce life expectancy and increase healthcare costs. Thus, some of the international differences observed in health data may not shed light on healthcare systems but instead reflect differing approaches to diet and exercise.

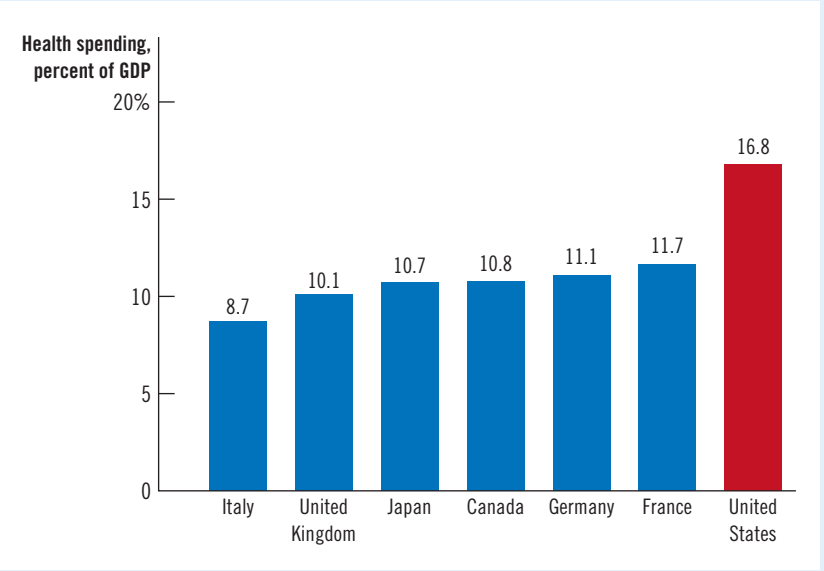
One notable and widely debated difference between the United States and other nations concerns pharmaceutical pricing. On average, Canadians spend about 30 percent less on drugs than Americans do (and residents of some European nations spend even less). Often, the same drug is much cheaper on the Canadian side of

Figure 4

Health Spending as a Share of GDP: International Comparison

The United States spends a much larger share of its income on healthcare than do other nations.

Source: The World Bank. Data are for 2019.



the border than on the American side. The reason is that Canada, with its centralized government-run health system, maintains strict controls over drug prices. Critics of the U.S. healthcare system believe that pharmaceutical companies are taking advantage of America's less centralized system by charging Americans exorbitant prices for patented drugs. They argue that the U.S. government should follow Canada's lead and undertake more aggressive regulatory policies to reduce drug prices. Defenders of the U.S. system believe that expanding price controls into the United States would reduce the incentives for pharmaceutical companies to engage in research into new drugs. Consumers would benefit from lower prices today, but they would bear the cost of a smaller range of treatments in the future. Defenders of the U.S. system also say that countries with price controls are free-riding on the research that is largely financed by higher prices in the United States.

12-2d Out-of-Pocket Spending Is a Declining Share of Health Expenditure

When you go to physicians or dentists, they are compensated in one of two ways. Either you pay them directly out of your own pocket or a third party, such as a government insurance program or private insurance company, pays them for you. Sometimes, the payment is a combination of the two.

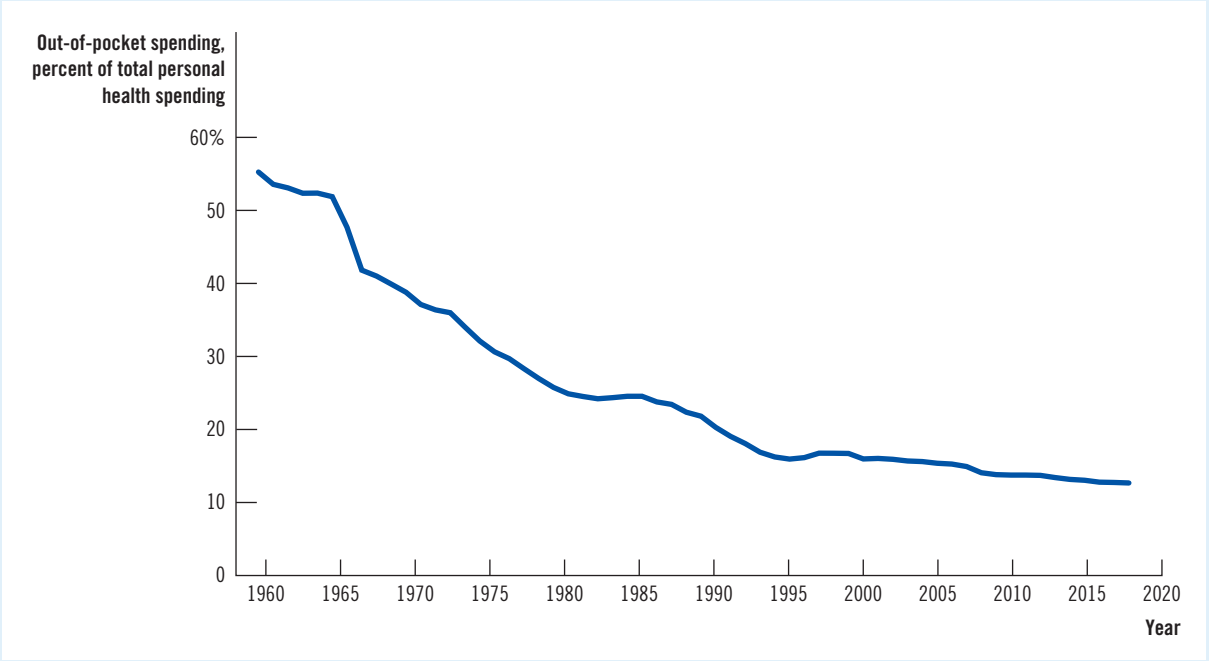
Figure 5 shows the percentage of spending on personal healthcare that is paid out of pocket in the United States. The percentage declined from 55 percent in 1960 to 13 percent in 2019. Conversely, third-party payment rose from 45 to 87 percent of health spending. Of the large amount **not** paid out of pocket, just under half is paid by private insurance companies, while just over half is paid by government insurance programs, such as Medicare (the program for those age 65 and older) and Medicaid (the program for those with low incomes).

Figure 5

Out-of-Pocket Spending as a Share of Total Personal Health Spending

The share of U.S. personal healthcare that households pay for themselves has declined over time. The share paid by third parties—government programs and private health insurers—has increased.

Source: Centers for Medicare & Medicaid Services.

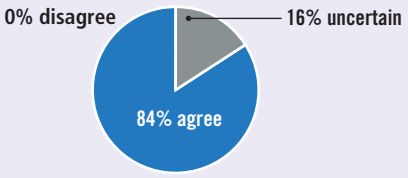


Ask the Experts

Cadillac Tax

"The 'Cadillac tax' on expensive employer-provided health insurance plans will reduce costly distortions in U.S. health care if it is allowed to take effect as scheduled in 2018."

What do economists say?



Source: IGM Economic Experts Panel, May 17, 2016

The increasing importance of health insurance is understandable. Because the need for healthcare is unpredictable, as it gets more expensive, people will seek to protect themselves from the financial risks by obtaining insurance.

Yet many economists believe that the U.S. health system has become too reliant on health insurance, especially for small or routine expenditures. They believe that excessive insurance exacerbates the moral hazard problem discussed earlier, which drives up healthcare costs. To explain excessive insurance, they note that the U.S. income tax system gives preferential treatment to employer-provided health insurance. Compensation in the form of health insurance is tax-exempt, unlike cash compensation. As a result, employees have an incentive to bargain for more generous (and thus more expensive) health insurance than they otherwise would, reducing the amount of healthcare they pay out of pocket.

The Affordable Care Act tried to remedy this problem by levying a so-called "Cadillac tax" on especially expensive

employer-provided health plans. This policy would have leveled the playing field between paying workers in the form of cash compensation and paying them in the form of generous health insurance. That is, the tax code would have no longer given an incentive for excessive insurance. The Cadillac tax was originally scheduled to go into effect in 2018 but was delayed and eventually repealed in 2019.

QuickQuiz

5. As a percentage of the economy's total income, U.S. spending on healthcare is _____ it was 50 years ago and _____ it is in other countries.
   a. much larger than; about the same as
   b. much larger than; much larger than
   c. about the same as; about the same as
   d. about the same as; much larger than
6. Baumol's cost disease arises in a service sector when that sector experiences
   a. low productivity growth.
   b. high productivity growth.
   c. low demand growth.
   d. high demand growth.
7. Compared with the past, Americans today are paying for a _____ percentage of healthcare out of pocket. This trend _____ the problem of moral hazard.
   a. larger; increases
   b. larger; decreases
   c. smaller; increases
   d. smaller; decreases

Answers are at the end of the chapter.

12-3 Conclusion: The Policy Debate over Healthcare

This chapter introduced some of the facts and economic insights useful for understanding the market for healthcare. Most of these ideas are widely accepted by the economists who study healthcare. Despite this consensus, there is an ongoing debate among U.S. policymakers about the government's role in the healthcare system.

While those on the political left disagree about many things, most agree that a large government role is essential in healthcare. They often say that private insurance companies are inefficient and tend to put profit ahead of people. Some would like the government to offer people a **public option** in the healthcare system—that is, a government-run insurance program that any person can buy into instead of purchasing private insurance. Others would like to move toward a **single-payer** system in which the government pays for healthcare for everyone out of tax revenue, as Medicare now does for those age 65 and older. They point to Canada as a worthwhile example. Centralized systems run by intelligent administrators, they say, are best able to reduce administrative inefficiency, eliminate wasteful treatment, bargain with providers for lower costs, and allocate healthcare resources most equitably to where they are most needed.

The political right is also diverse, but it is fair to say that many people there would like to reduce the government's role in the healthcare system, as in many other spheres. They acknowledge that the market for health insurance needs to be regulated but would like the regulation to be less heavy-handed than it is now.

In the News

Lessons from the Pandemic of 2020

Reducing wasteful medical spending is a good goal, but it is not easy to achieve.

Why It's So Hard to Cut Waste in Health Care

By Amy Finkelstein

Cutting waste while preserving critically important treatment is the holy grail of health care policy. The coronavirus pandemic has shown why that goal has been so stubbornly difficult to achieve.

One of the pandemic's startling effects is that despite the overwhelming need for Covid-19 treatments and vaccinations, health care spending in the United States has declined sharply. That's because people have been reluctant or unable to get medical care that isn't connected to the coronavirus. My own calculations, based on government data, show that overall total care expenditures dropped 10 percent in the first nine months of 2020, compared with the same period a year earlier.

This indiscriminate reduction in spending closely resembles the unfortunate effects of an array of policies over the past half-century intended to reduce unnecessary medical care. These too cut essential and inessential care alike.

Consider what has happened during the pandemic. Essential health care visits have plummeted: Hospital admissions for severe heart attacks and strokes and visits for routine childhood vaccinations have fallen sharply. Those reductions could have grave consequences.

At the same time, some types of care that are often overused—and can produce unnecessary follow-up treatments, anxiety and spending—also declined in frequency during the pandemic. This includes radiological scans and routine cancer screenings. Indeed, a broad spectrum of medical care has declined at roughly comparable rates, as patients avoided in-person visits and medical providers cut availability.

In short, the pandemic has pushed two basic levers used in strategies to limit wasteful medical care. These strategies aim to decrease patient demand and to restrict the availability of treatment by health care providers. And they tend to have indiscriminate effects, reducing treatment that is critical, along with care that is not essential.

The economic importance of reducing wasteful spending is hard to overstate. The health care sector accounts for almost one-fifth of the U.S. economy, and one-quarter or more of this spending is wasted, a range of studies have shown. Excessive administrative costs, high prices and inappropriate medical care are the three major culprits in wasteful spending, driving insurance premiums and taxes without helping patients.

For decades, the driving principle behind reducing needless medical care has been that if patients or their providers have financial "skin in the game," they will make prudent decisions. To encourage patients to make good decisions, insurers make them pay some share of their medical costs. High-deductible health insurance plans, in which the patient herself pays the first thousand (or more) dollars of any medical spending for the year, are one increasingly common example.

The economic theory is simple: If something becomes more expensive, people will buy less of it. And the empirical evidence is overwhelming: When patients have to pay more, they use less medical care.

The problem is they use less of all types of care. Introducing a high-deductible plan, for example, reduces scans for nonspecific low back pain and antibiotics for respiratory infections, as intended. But high deductibles also tend to reduce use of mental health services and of prescription drugs for managing chronic diabetes and high cholesterol, which can have negative consequences.

My own research shows covering uninsured individuals with Medicaid is a similarly blunt instrument for encouraging people to get critically important care. Medicaid increased preventive care, which was intended, but also increased emergency room visits, which was

They say that the best healthcare will arise if private insurers and providers openly compete for consumers. They fear that a centralized, government-run system would limit individual freedom, excessively ration care, and stifle innovation. In their view, Canada is an example of what could go wrong: Waiting times for medical procedures there can be uncomfortably long, and those who can afford it sometimes choose not to wait and instead seek treatment in the United States.

The debate over health policy is part of the larger debate over income inequality and the role of government. As Chapter 1 pointed out, society often faces a trade-off between equality and efficiency, and that fact looms large when discussing the provision of healthcare. Those on the political left generally want to achieve universal health coverage by providing government insurance or by

not. Even more targeted insurance that lowers only the price of specific and valuable care has modest effects at best, in encouraging the desired outcome.

Similarly, financial incentives that encourage physicians and hospitals to refrain from superfluous treatments may also jettison some indispensable care. Traditionally, these providers were reimbursed piece-rate, billing separately for each test and procedure, so that the more they did, the more they were paid.

To encourage only essential care, there has been a widespread shift to paying a fixed fee for a patient, regardless of what treatment is provided. A hospital might be paid $25,000 for a patient's knee replacement, for example, and then have to absorb all the associated costs, including the hospital stay, surgeon's fees and post-discharge rehabilitative care. Or a network of health care providers might be paid $10,000 a year for all of a patient's health care needs.

That way, the medical provider is on the hook for the costs of treatment. But such incentives are a double-edged sword: What encourages cost-consciousness can discourage optimal care.

Countries around the world are grappling with these issues. A recent study of health care in 10 countries—including several with single-payer systems—concluded that, to one degree or another, they all are rife with inefficient and unnecessary care.

One reason it is so hard to trim waste is that physicians and regulators have already eliminated so many demonstrably harmful or useless treatments. Patients are no longer offered patent medicines, routine bloodlettings or lobotomies.

PAT GREENHOUSE/BOSTON GLOBE/GETTY IMAGES

Amy Finkelstein

Reducing unnecessary administrative costs and curbing high prices is similarly difficult. Some red tape is purposeless, but not all of it is. Lowering prices can imperil the availability of valuable treatment. An adage about advertising applies to health care: Half of spending is wasted; we just don't know which half.

Low-value doesn't mean no-value-ever. That's presumably why Choosing Wisely, a campaign to reduce unnecessary care, identified a large number of tests and procedures that patients and physicians should "question" but not eliminate. We don't "question" whether it's wise to stick a finger in an electric socket. We warn that you shouldn't do it. But most medical procedures are different. They are a matter of judgment.

Ordering a computerized tomography, or C.T., scan for a patient the first time she complains of a nondescript headache is generally not a wise move. Neither is getting married after a first date. But occasionally that unnecessary scan catches a life-threatening problem before it's too late much as, once in a while, a whirlwind marriage leads to happily-ever-after. Unlikely isn't the same as never, which can make it hard for a well-intentioned physician or a hopeless romantic to resist.

That doesn't mean we should stop trying to cut waste. But a simple, miracle cure for excising most unnecessary medical care? Don't buy it, with or without insurance coverage. ■

Source: *New York Times*, January 24, 2021.

Questions to Discuss

1. If you faced a larger co-pay when you visited a physician, would it change how often you sought medical care? Why or why not?
2. Are you good at judging when you need medical care and when it is best to let your body heal on its own? Give some examples.
3. How much do you trust your physician in deciding what tests to order? Why?

Ms. Finkelstein is an economics professor at M.I.T.

subsidizing private insurance for households with lower incomes. But paying for these policies requires increased taxes on households with higher incomes, and those taxes are likely to distort incentives and shrink the size of the economic pie. Those on the political right emphasize the distortionary effects of taxes and income redistribution. They advocate more limited government and lower taxes and say this approach will expand the economic pie. But smaller government and less revenue mean fewer public resources to help those who struggle to get the healthcare they need.

This debate raises hard and important questions, and this chapter does not offer easy answers. But this introduction to health economics should give you a starting point for thinking through the many issues.

Chapter in a Nutshell

- The market for healthcare differs from most other markets in several ways. First, there are pervasive externalities, such as those associated with vaccination and medical research. Second, because consumers cannot easily gauge the quality of what they are buying, private and public institutions intervene to ensure that treatment is appropriate. Third, healthcare is often judged to be a right, leading to a government role to make sure that everyone has access to it.
- Spending on healthcare can be large and unpredictable, but health insurance reduces the financial risk that people face from a costly event. The problems of moral hazard and adverse selection, however, hinder the effectiveness of the market for health insurance. When people have insurance, either from a private company or a government program, the insurer establishes rules regarding financing, access, and payment.
- Since 1900, life expectancy in the United States has increased by about 30 years, largely because of advances in medical technology.
- Over the past 60 years, healthcare spending has grown substantially as a percentage of national income. Several forces are at work: Baumol's cost disease, advances in medical technology, the aging of the population, and rising incomes.
- The United States spends a higher fraction of its national income on healthcare than other developed nations. This fact has no simple explanation. Critics and defenders of the U.S. healthcare system point to different possible reasons.
- The percentage of healthcare spending paid out of pocket, rather than by insurance, has declined substantially over time. Some economists say that reliance on insurance is excessive, exacerbating moral hazard and driving up healthcare costs.

Key Concepts

risk aversion, p. 231 moral hazard, p. 231 adverse selection, p. 232

Questions for Review

1. Give two examples of externalities in the healthcare system.
2. Explain moral hazard and adverse selection in the market for health insurance.
3. Describe the three sets of rules that are necessary in a healthcare market with an insurer.
4. Give three reasons why life expectancy has been increasing over time.
5. How does spending on healthcare in the United States compare with spending 50 years ago? Give four economic forces that might help explain the trend.
6. How does healthcare spending in the United States compare with spending in other developed nations?
7. Explain the rationale for a "Cadillac tax" on expensive health insurance plans.

Problems and Applications

1. Consider how health insurance affects the quantity of healthcare services performed. Suppose that the typical medical procedure costs $100, yet a person with health insurance pays only $20 out of pocket. An insurance company pays the remaining $80. (The insurance company recoups the $80 through premiums, but the premium a person pays does not depend on how many procedures that person chooses to undertake.)
   a. Draw the demand curve in the market for medical care. (In your diagram, the horizontal axis should represent the number of medical procedures.) Show the quantity of procedures demanded if each procedure has a price of $100.
   b. On your diagram, show the quantity of procedures demanded if consumers pay only $20 per procedure. If the cost of each procedure to society is truly $100 and if individuals have health

insurance as described above, will the number of procedures performed maximize total surplus? Explain.

c. Economists often blame the health insurance system for excessive use of medical care. Given your analysis, why might the use of care be viewed as "excessive"?

d. What sort of policies might prevent this excessive use?

2. The Live-Long-and-Prosper Health Insurance Company charges $5,000 annually for a family insurance policy. The company's president suggests that the company raise the annual price to $6,000 to increase its profits. If the firm follows this suggestion, what economic problem might arise? Would the firm's pool of customers tend to become more or less healthy on average? Would the company's profits necessarily increase?

3. Complete the logic behind Baumol's cost disease of the service sector. Assume that the economy has two sectors—manufacturing and services—and that technological advances occur only in manufacturing.
   a. If a technological advance increases labor productivity in manufacturing, the demand for labor in manufacturing increases, and the equilibrium wage of manufacturing workers (increases/decreases).
   b. If some workers can move between manufacturing and services, the change identified in part (a) (increases/decreases) the supply of labor in the service sector.
   c. As a result of the shift identified in part (b), the wage of service-sector workers (increases/decreases).
   d. Due to the change identified in part (c), the costs of businesses that provide services (increase/decrease).
   e. Because of the change in costs identified in part (d), the price of services (increases/decreases).
   f. Thus, technological advances in manufacturing cause the price of services to (rise/fall).

4. Consider an example of the death spiral. An economy consists of five types of people in equal numbers with different pre-existing health problems. Here are their expected healthcare costs for the coming year:

A	$1,000
B	$2,000
C	$3,000
D	$4,000
E	$5,000

   a. If all five types of people buy health insurance, what will be the average cost of healthcare paid by insurance companies? The companies need to cover this average cost, plus other expenses (including a normal profit) of $700. Assuming the companies must charge all consumers the same price, what price will that be?
   b. While all consumers would like to be insured to reduce uncertainty, they value this feature at $1,000 and, therefore, will forgo insurance if its price exceeds their own expected healthcare costs by more than $1,000. Which consumers will buy insurance at the price calculated in part (a)?
   c. For the new insured pool identified in part (b), what is the average cost of healthcare? When insurance companies observe this new average cost, what price will they charge for insurance?
   d. When consumers see the new price in part (c), which consumers will buy insurance?
   e. For the new insured pool identified in part (d), what is the average cost of healthcare? When insurance companies observe this new average cost, what price will they charge for insurance?
   f. When consumers see the new price in part (e), which consumers will buy insurance?
   g. For the new insured pool identified in part (f), what is the average cost of healthcare? When insurance companies observe this new average cost, what price will they charge for insurance?
   h. When consumers see the new price in part (g), which consumers will buy insurance? Does this outcome now look like an equilibrium?
   i. The value of insurance to each person is the $1,000 from reduced uncertainty plus the expected cost of healthcare; the cost of providing it is the insurance premium, and the consumer surplus is the excess of value over the premium. How much consumer surplus is derived from the insurance market as described in part (h)?
   j. Now suppose the government mandates that everyone buy insurance, so we return to the situation in part (a). What surplus would each consumer get? What is total surplus in the market?
   k. If this society voted on the insurance mandate and everyone voted based on their self-interest, how would the vote turn out?
   l. To ensure that everyone obeys the mandate, how large would the penalty for failing to buy insurance need to be?

QuickQuiz Answers

1. **d** 2. **c** 3. **a** 4. **b** 5. **b** 6. **a** 7. **c**

Chapter

13

The Design of the Tax System

Al "Scarface" Capone, the 1920s gangster and crime boss, was never convicted for his violent crimes. Yet, eventually, he did go to jail—for tax evasion. Capone failed to heed Ben Franklin's observation that "in this world nothing is certain but death and taxes."

When Franklin made this claim in 1789, Americans paid less than 5 percent of their income in taxes, and that remained true for the next hundred years. Over the course of the 20th century, however, taxes became ever more important in the lives of most U.S. citizens. Today, all taxes taken together—including personal income taxes, corporate income taxes, payroll taxes, sales taxes, and property taxes—use up more than a quarter of total American income. In many European countries, the tax bite is even larger.

Taxes are inevitable because citizens expect their governments to provide them with goods and services. One of the **Ten Principles of Economics** in Chapter 1 is that markets are usually a good way to organize economic activity. But market economies rely on property rights and the rule of law, so the government provides police and courts. Another of the **Ten Principles of Economics** is that the government can sometimes improve market outcomes. When the government remedies an externality (such as pollution), provides a public good (such as national defense), or regulates the use of a

BLACKCAT/E+/GETTY IMAGES, GOWITHSTOCK/SHUTTERSTOCK.COM

common resource (such as fish in a public lake), it can raise economic well-being. But these activities can be costly. For the government to perform a range of critical functions, it must raise revenue through taxation.

Earlier chapters used the model of supply and demand to analyze taxes. Chapter 6 discussed how a tax on a good reduces the quantity sold and how the tax burden is shared by buyers and sellers depending on the elasticities of supply and demand. Chapter 8 examined how taxes affect economic well-being. In most cases, taxes cause **deadweight losses**: The reduction in consumer and producer surplus resulting from a tax exceeds the revenue raised by the government. Yet, as Chapter 10 demonstrated, taxes can enhance efficiency when they internalize externalities and correct market failures.

This chapter builds on these lessons to discuss the design of a tax system. It begins with an overview of how the U.S. government raises money. It then discusses the principles of taxation. Most people agree that taxes should impose as small a cost as possible and that the tax burden should be distributed fairly. That is, the tax system should be both **efficient** and **equitable**. As we will see, however, stating these goals is easier than achieving them.

13-1 U.S. Taxation: The Big Picture

How much of the nation's income does the government collect as taxes? Figure 1 shows government revenue, including federal, state, and local, as a percentage of total income for the U.S. economy. It shows that the role of government has grown substantially over the past century. In 1902, the government collected only 7 percent of total income; in recent years, the government has collected almost 30 percent. In other words, as the economy's income has grown, the government's tax revenue has grown even more.

Figure 1

Government Revenue as a Percentage of GDP: Changes over Time

This figure shows the revenue of the federal government and of state and local governments as a percentage of gross domestic product (GDP), which measures total income in the economy. It shows that the government plays a large role in the U.S. economy and that its role has grown over time.

Source: *Historical Statistics of the United States;* Bureau of Economic Analysis; and author's calculations.

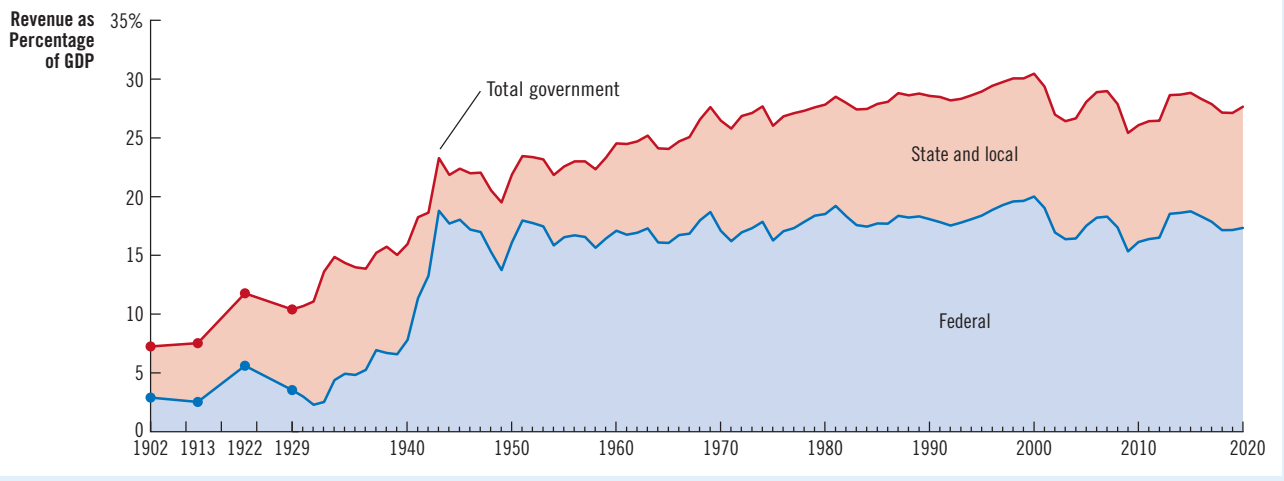


Figure 2

Government Revenue as a Percentage of GDP: International Comparisons

The percentage of income that governments take in taxes varies substantially from country to country.

Source: OECD. Data are for 2019.

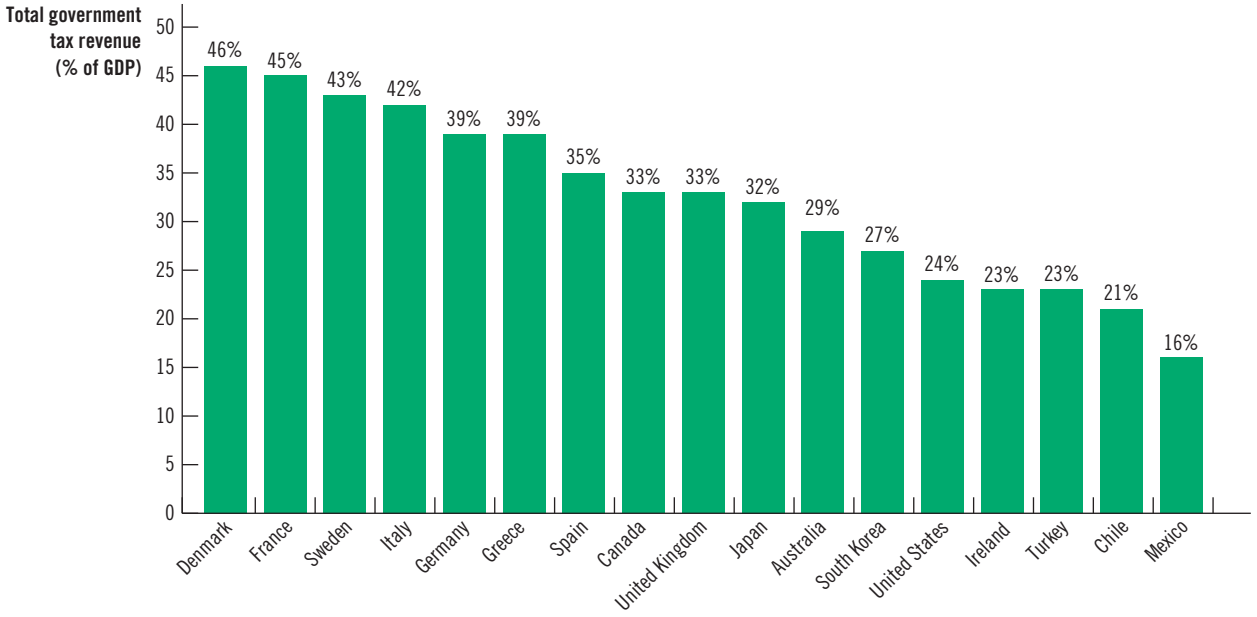


Figure 2 presents the tax burden for several major countries, as measured by tax revenue as a percentage of total income. The United States has a low tax burden compared with most other advanced economies. Many European nations have much higher taxes, which finance a more generous social safety net, including greater assistance for people experiencing poverty and unemployment.

13-1a Taxes Collected by the Federal Government

The U.S. federal government collects about two-thirds of the taxes in our economy. Table 1 shows the receipts of the federal government in 2020. Total receipts that year were about $3.7 trillion, a number so large that it is hard to comprehend. To bring this astronomical number down to earth, we can divide it by the size of the U.S. population, which was about 331 million in 2020. We then find that the average American paid $11,069 to the federal government in 2020.

Personal Income Taxes The largest source of revenue for the federal government is the personal income tax. As April 15 approaches in a typical year, almost every American family fills out a form to determine the income tax it owes the government. Each family is required to report its income from all sources: wages from working, interest on savings, dividends from corporations in which it owns shares, profits from any small businesses it operates, and so on. The family's **tax liability** (how much it owes) is based on its total income.

Table 1

Receipts of the Federal Government: 2020

Source: Bureau of Economic Analysis. Columns may not sum to total due to rounding.

Tax	Amount (billions)	Amount per Person	Percent of Receipts
Personal income taxes	$1,690	$5,106	46%
Social insurance taxes	1,421	4,293	39
Corporate income taxes	199	601	5
Other	354	1,069	10
Total	$3,664	$11,069	100%

But a family's income tax liability is not simply proportional to its income. Instead, the law requires a more complicated calculation. Taxable income is total income minus an amount based on the number of dependents (primarily children) and minus certain expenses that policymakers have deemed "deductible" (such as mortgage interest payments and charitable giving). The tax liability is then calculated from taxable income using a schedule like the one shown in Table 2.

This table presents the **marginal tax rate**—the tax rate applied to each additional dollar of income. Because the marginal tax rate rises as income rises, higher-income families pay a larger percentage of their income in taxes. Note that each tax rate in the table applies only to income within the associated range, not to a person's entire income. For example, a person with an income of $1 million still pays only 10 percent of the first $9,875. (A more complete discussion of marginal tax rates will come later in this chapter.)

Payroll Taxes Almost as important to the federal government as the personal income tax are **payroll taxes**, which are levied on the wages that a firm pays its workers. Table 1 calls this revenue **social insurance taxes** because it mainly pays for Social Security and Medicare. Social Security is an income-support program designed primarily to maintain the living standards of older people. Medicare is the government health program for them. In 2020, the total payroll tax was 15.3 percent for annual earnings up to $137,700 and 2.9 percent of earnings above $137,700, together with an additional 0.9 percent for taxpayers with high incomes (above $200,000 if single, $250,000 if married). For many middle-income households, the payroll tax is the largest tax they pay.

Table 2

The Federal Income Tax Rates: 2020

This table shows the marginal tax rates for single taxpayers. The taxes owed depend on all the marginal tax rates up to the taxpayer's income level. For example, a taxpayer with an income of $40,000 pays 10 percent of the first $9,875 of income and then 15 percent of the rest.

On Taxable Income . . .	The Tax Rate Is . . .
From $0 to $9,875	10%
From $9,876 to $40,125	15%
From $40,126 to $85,525	22%
From $85,526 to $163,300	24%
From $163,301 to $207,350	32%
From $207,351 to $518,400	35%
From $518,401 and above	37%

Corporate Income Taxes Next in magnitude, but much smaller than either personal income taxes or social insurance taxes, is the corporate income tax. A **corporation** is a business set up to have its own legal existence, distinct and separate from its owners. The government taxes each corporation based on its **profit**—the amount the corporation receives for the goods or services it sells minus the costs of producing them. Notice that corporate profits are often taxed twice. They are taxed once by the corporate income tax when the corporation earns the profits, and they are taxed again by the personal income tax when the corporation uses its profits to pay dividends to its shareholders. (An exception to this second tax occurs when the shares are held by nontaxable entities, such as university endowments and some retirement accounts.) In part to compensate for double taxation, policymakers have decided to tax dividend income at lower rates than other types of income: In 2020, the top marginal tax rate on dividend income was only 20 percent (plus a 3.8 percent Medicare tax), compared with the top marginal tax rate on ordinary income of 37 percent (plus the same 3.8 percent).

Other Taxes The last category, labeled "other" in Table 1, makes up 10 percent of receipts. This category includes **excise taxes**, which are taxes on specific goods such as gasoline, cigarettes, and alcoholic beverages. It also includes various small items, such as estate taxes and customs duties.

13-1b Taxes Collected by State and Local Governments

State and local governments collect about a third of all taxes paid. Table 3 shows the receipts of U.S. state and local governments. Total receipts for 2020 were $3 trillion, or $9,157 per person. The table also shows how this total is broken down into different kinds of taxes.

The most important taxes for state and local governments are property taxes, which make up 19 percent of receipts and are levied on property owners as a percentage of the estimated value of the land and structures.

State and local governments are also able to impose personal and corporate income taxes. In many cases, these taxes are like federal income taxes. In others, the rules are different. For example, some states tax income from wages less heavily than income earned as interest and dividends. Some states do not tax personal income at all.

Table 3

Receipts of State and Local Governments: 2020

Source: Bureau of Economic Analysis. Columns may not sum to total due to rounding.

Tax	Amount (billions)	Amount per Person	Percent of Receipts
Property taxes	$586	$1,770	19%
Personal income taxes	466	1,408	15
Sales taxes	426	1,287	14
Excise taxes	209	631	7
Corporate income taxes	76	230	3
Federal government	873	2,637	29
Other	395	1,193	13
Total	$3,031	$9,157	100%

At 14 percent of receipts, sales taxes are also big revenue raisers. They are levied as a percentage of the total amount spent at retail stores. Every time customers buy something, they pay an extra amount that retailers remit to the government. (Some states exclude certain items that are considered necessities, such as food and clothing.) Similar to sales taxes are excise taxes, which are levied on specific goods, such as gasoline, cigarettes, or alcoholic beverages. Excise taxes make up 7 percent of state and local receipts.

State and local governments also receive substantial funds from the federal government. To some extent, the federal government's policy of sharing its revenue with state governments redistributes money from high-income states (which pay more taxes) to low-income states (which receive more benefits). Often, these funds are tied to programs that the federal government wants to subsidize. For example, Medicaid, which provides healthcare for low-income individuals, is managed by the states but funded largely by the federal government.

Finally, state and local governments receive receipts from sources included in the "other" category in Table 3. These include fees for fishing and hunting licenses, tolls from roads and bridges, and fares for public buses and subways.

QuickQuiz

1. As a percent of national income, taxes in the United States are
   a. higher than in France, Germany, and the United Kingdom.
   b. lower than in France, Germany, and the United Kingdom.
   c. higher than in France and Germany but lower than in the United Kingdom.
   d. lower than in France and Germany but higher than in the United Kingdom.

2. The two largest sources of tax revenue for the U.S. federal government are
   a. personal and corporate income taxes.
   b. personal income taxes and payroll taxes for social insurance.
   c. corporate income taxes and payroll taxes for social insurance.
   d. payroll taxes for social insurance and property taxes.

Answers are at the end of the chapter.

13-2 Taxes and Efficiency

The U.S. tax system keeps changing as lawmakers embrace different approaches to raising revenue. If we were starting from scratch, how might we design a good tax system in principle? The primary aim is to fund the government, but there are many ways to do that. When choosing among alternative taxes, it is useful to consider two objectives: efficiency and equity.

One tax system is more efficient than another if it raises the same amount of revenue at a smaller cost to taxpayers. What are the costs of taxes to taxpayers? The most obvious is the tax payment itself, which is an inevitable feature of any tax system. But there are two other costs, which well-designed tax policy avoids or, at least, minimizes:

- The deadweight losses that result when taxes distort the decisions people make
- The administrative burdens imposed on taxpayers as they comply with the tax laws

An efficient tax system is one that has small deadweight losses and small administrative burdens.

13-2a Deadweight Losses

One of the **Ten Principles of Economics** is that people respond to incentives, and this includes incentives provided by the tax system. If the government taxes ice cream, people tend to eat less ice cream and more frozen yogurt. If the government taxes housing, people are likely to live in smaller houses and spend more on other things. If the government taxes labor earnings, people will be inclined to work less and enjoy more leisure.


BILL ABBOTT/CARTOONSTOCK LTD

"Before I answer, what are the tax implications?"

Because taxes distort incentives, they often entail deadweight losses, a phenomenon explained in Chapter 8. The deadweight loss of a tax is defined as the reduction in market participants' well-being in excess of the revenue raised for the government. In other words, the deadweight loss is the inefficiency that arises as people allocate resources according to tax incentives rather than the costs and benefits of the goods and services being bought and sold.

Suppose that Khalil is willing to pay $16 for a pizza and Carmen is willing to pay $12 for one. The supply of pizza is perfectly elastic at a price of $10 (which ensures that producer surplus is zero). If there is no tax on pizza, both Khalil and Carmen buy one, and each gets some surplus of value over the amount paid. Khalil's consumer surplus is $6, and Carmen's is $2, for a total of $8.

Now suppose that the government levies a $4 tax on pizza, which raises the price to $14. Khalil still buys a pizza but now has consumer surplus of only $2. Carmen doesn't buy one because now it costs more than it's worth to her, so her consumer surplus is zero. The surplus of the two consumers together falls by $6 (from $8 to $2), while the government collects tax revenue of $4 on Khalil's pizza. The tax has a deadweight loss because the decline in consumer surplus exceeds the gain in tax revenue. In this case, the deadweight loss is $2.

Notice that the deadweight loss comes not from Khalil, who pays the tax, but from Carmen, who doesn't. The $4 reduction in Khalil's surplus exactly offsets the amount of revenue the government collects. The deadweight loss arises because the tax induces Carmen to change her behavior. When the tax raises the price of pizza, Carmen is worse off, but there is no offsetting government revenue. This reduction in Carmen's welfare is the deadweight loss of the tax.

Finally, recall that not all taxes that alter incentives lead to deadweight losses. As Chapter 10 showed, when there are externalities, a market on its own can lead to inefficient outcomes, and the right tax can correct the problem. For example, if the wafting smell of pizza baking in a brick oven makes passersby hungry and unhappy, perhaps because it reminds them that pizza is delicious but not very healthy, a tax on pizza could conceivably enhance efficiency. Corrective taxes also raise tax revenue, which can be used to reduce taxes that create deadweight losses.

Case Study

Should Income or Consumption Be Taxed?

When taxes cause people to change their behavior—such as inducing Carmen to buy less pizza—the taxes can make the allocation of resources less efficient. Much government revenue comes from the personal income tax, and a case study in Chapter 8 discussed how this tax discourages some people from working as hard as they otherwise might. That's not the end of the story, though. This tax causes another inefficiency: It discourages people from saving.

Consider Sammy Saver, a 25-year-old deciding whether to save $1,000. If Sammy puts this money in a savings account that earns 6 percent and leaves it there, he will have $10,286 when he retires at age 65. Yet if the government taxes one-fourth of his interest income each year, the effective interest rate is only 4.5 percent.

After 40 years of earning 4.5 percent, the $1,000 grows to only $5,816, which is 43 percent less than what it would have been without taxation. That's why taxes on interest income make saving less attractive.

Some economists advocate eliminating the tax system's disincentive toward saving. Rather than taxing the amount of income that people earn, the government could tax the amount that people spend. Under this proposal, all income that is saved is free from taxation until the saving is later spent. This alternative system, called a **consumption tax**, would not distort people's saving decisions.

Some provisions of current law already make the system a bit like a consumption tax. People can put a limited amount of their income into tax-sheltered accounts, like Individual Retirement Accounts and 401(k) plans. This income, along with the accumulated interest it earns, avoids taxation until the money is withdrawn at retirement. For people who do most of their saving through these retirement accounts, their tax bill is, in effect, based on their consumption rather than their income.

European countries often favor consumption taxes. Most of them raise a significant amount of government revenue through a value-added tax, or a VAT. A VAT is like the retail sales tax that many U.S. states use. But rather than collecting all of the tax at the retail level when the consumer buys the final good, the government collects the tax in stages as the item is being produced (that is, as value is added along the chain of production).

Some U.S. policymakers have proposed that the U.S. tax code move further in the direction of taxing consumption rather than income. In 2005, the economist Alan Greenspan, then chair of the Federal Reserve, offered this advice to a presidential commission on tax reform: "As you know, many economists believe that a consumption tax would be best from the perspective of promoting economic growth—particularly if one were designing a tax system from scratch—because a consumption tax is likely to encourage saving and capital formation. However, getting from the current tax system to a consumption tax raises a challenging set of transition issues." ●

13-2b Administrative Burden

Many people would place the task of filling out tax forms toward the bottom of their personal list of important but unpleasant chores, somewhere between flossing their teeth and cleaning their toilet. The tax system's administrative burden is a cause of inefficiency. It includes not only the time spent filling out forms on deadline but also the hours needed throughout the year to keep records for tax purposes and the resources the government uses to enforce the tax laws.

Many people—especially those in higher tax brackets—hire lawyers and accountants to help with their taxes. These experts in the complex tax laws fill out tax forms for their clients and help them arrange their affairs in a way that reduces the amount of taxes owed. This behavior is legal tax avoidance, as opposed to illegal tax evasion. As the famed jurist Learned Hand put it, "Anyone may arrange his affairs so that his taxes shall be as low as possible; he is not bound to choose that pattern which best pays the treasury. There is not even a patriotic duty to increase one's taxes."

Critics of the tax system say that advisers help their clients avoid taxes by taking advantage of the detailed provisions of the tax code, often dubbed "loopholes." In some cases, loopholes are congressional mistakes: They arise from ambiguities or omissions in the tax laws. More often, they arise because Congress has chosen to

favor specific types of behavior. For example, the federal tax code gives preferential treatment to investors in municipal bonds because Congress wanted to make it easier for state and local governments to borrow money. To some extent, this provision benefits states and localities, but it also benefits taxpayers with higher incomes. Most loopholes are well known by those in Congress who make tax policy, but what looks like a loophole to one taxpayer may look like a justifiable tax provision to another.

The resources devoted to complying with the tax laws are themselves a type of deadweight loss. The government gets only the amount of taxes paid. By contrast, the taxpayer loses not only this amount but also the time and money spent documenting, computing, and avoiding taxes.

The administrative burden of the tax system could be reduced by simplifying the tax laws. Yet simplification is often politically difficult. Most people are ready to eliminate the loopholes that benefit others, but few are eager to give up the loopholes that they benefit from themselves. In the end, the complexity of the tax law results from the political process as taxpayers with their own special interests lobby for their causes.

13-2c Marginal Tax Rates versus Average Tax Rates

When discussing income taxes, economists distinguish between two notions of the tax rate: the average and the marginal. The **average tax rate** is total taxes paid divided by total income. The **marginal tax rate** is the increase in taxes from an additional dollar of income.

average tax rate
total taxes divided by total income

marginal tax rate
the increase in taxes from an additional dollar of income

lump-sum tax
a tax that is the same amount for every person

For example, suppose that the government taxes 20 percent of the first $50,000 of income and 50 percent of all income above $50,000. Under this tax, a person who makes $60,000 pays a tax of $15,000: 20 percent of the first $50,000 ($0.20 \times \$50{,}000 = \$10{,}000$) plus 50 percent of the remaining $10,000 ($0.50 \times \$10{,}000 = \$5{,}000$). In this case, the average tax rate is $15,000/$60,000, or 25 percent. But if this person earned an additional dollar of income, that dollar would be subject to the 50 percent tax rate, so the amount owed to the government would rise by $0.50. Thus, the marginal tax rate is 50 percent.

The marginal and average tax rates each contain a useful piece of information. If we are trying to gauge the sacrifice made by a taxpayer, the average tax rate is more appropriate because it measures the fraction of income paid in taxes. By contrast, if we are trying to gauge how the tax system distorts incentives, the marginal tax rate is more meaningful. One of the **Ten Principles of Economics** in Chapter 1 is that rational people think at the margin. A corollary to this principle is that the marginal tax rate measures how much the tax system discourages people from working. If you are thinking about working a few more hours, the marginal tax rate determines how much the government takes from your additional earnings. It is the marginal tax rate, therefore, that determines the deadweight loss of an income tax.

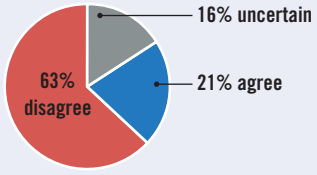
13-2d Lump-Sum Taxes

Suppose the government imposes a simple tax of $8,000 on everyone. That is, everyone owes the same amount, regardless of earnings, circumstances, or any actions that a person might take. Such a tax is called a **lump-sum tax**.

ASK THE EXPERTS Top Marginal Tax Rates

"Raising the top federal marginal tax on earned personal income to 70 percent (and holding the rest of the current tax code, including the top bracket definition, fixed) would raise substantially more revenue (federal and state, combined) without lowering economic activity."

What do economists say?



Source: IGM Economic Experts Panel, January 16, 2019.

A lump-sum tax clearly shows the difference between average and marginal tax rates. For someone with an income of $40,000, the average tax rate of an $8,000 lump-sum tax is 20 percent; for a person with an income of $80,000, the average tax rate is 10 percent. For both people, the marginal tax rate is zero because nothing is owed on an additional dollar of income.

A lump-sum tax is the most efficient tax possible. Because a person's decisions do not alter the amount owed, the tax does not distort incentives and, therefore, does not cause deadweight losses. Because everyone can easily calculate the amount owed and because there is no benefit to hiring tax lawyers and accountants, the lump-sum tax imposes a minimal administrative burden.

If lump-sum taxes are so efficient, why are they rare? Because efficiency isn't everything. A lump-sum tax would take the same amount from everyone regardless of income, and most people would consider that unfair. To explain real-world tax systems, we must look at the other major goal of tax policy: equity.

QuickQuiz

3. Betty gives piano lessons. She has an opportunity cost of $50 per lesson and charges $60. She has two students: Archie, who has a willingness to pay of $70, and Veronica, who has a willingness to pay of $90. When the government puts a $20 tax on piano lessons and Betty raises her price to $80, the deadweight loss is _____, and the tax revenue is _____.
   a. $10; $20
   b. $10; $40
   c. $20; $20
   d. $20; $40
4. If the tax code exempts the first $20,000 of income from taxation and then taxes 25 percent of all income above that level, a person who earns $50,000 has an average tax rate of _____ percent and a marginal tax rate of _____ percent.
   a. 15; 25
   b. 25; 15
   c. 25; 30
   d. 30; 25
5. Lump-sum taxes
   a. have a zero marginal tax rate.
   b. have a zero average tax rate.
   c. are costly to administer.
   d. impose large deadweight losses.

Answers are at the end of the chapter.

13-3 Taxes and Equity

Ever since American colonists dumped imported British tea into Boston Harbor, tax policy has generated some of the most heated debates in American politics. The heat is rarely fueled by questions of efficiency. Instead, it arises from disagreements over how the tax burden should be distributed. Russell Long, the Senator from Louisiana who had immense influence over federal tax policy from 1948 to 1987, once mimicked the public debate with this ditty:

> Don't tax you.
> Don't tax me.
> Tax that fella behind the tree.

Of course, if we rely on the government to provide some of the goods and services we want, someone must pay taxes to fund those goods and services. How should the burden of taxes be divided among the population? How do we evaluate whether

a tax system is fair? Everyone agrees that the tax system should be equitable, but there is much disagreement about how to make it so.

13-3a The Benefits Principle

The **benefits principle** of taxation states that people should pay taxes based on what they receive from government services. This approach tries to make public goods like private goods. It seems fair that a person who often eats ice cream pays more in total for ice cream than a person who rarely eats it. Similarly, according to this principle, a person who gets great benefit from a public good should pay more for it than a person who gets little benefit.

benefits principle
the idea that people should pay taxes based on the benefits they receive from government services

The gasoline tax, for instance, is sometimes justified using the benefits principle. In some states, revenues from a gas tax are used to build and maintain roads, and interstate highways are maintained, in part, from revenues from the federal gas tax. Because those who buy gas are the same people who use the roads, the gas tax might be viewed as a fair way to pay for this government service. (The recent spread of electric cars, however, has made the gas tax less well targeted, leading some policymakers to suggest a miles-driven tax instead.)

The benefits principle can also be used to argue that wealthy citizens should pay higher taxes than poorer ones. Why? Simply because the wealthy benefit more from public services. Consider, for example, the benefits of police protection from theft. Property owners with much to protect benefit more from the police than do people with less to protect. Therefore, according to the benefits principle, the wealthy should contribute more to the cost of maintaining the police force. The same argument can be used for many other public services, such as fire protection, national defense, and the court system.

It is even possible to use the benefits principle to argue for antipoverty programs funded by taxes on the wealthy. As Chapter 11 discussed, people may prefer living in a society without poverty, suggesting that antipoverty programs are a public good. If the wealthy place a greater dollar value on this public good than members of the middle class do, perhaps just because the wealthy have more to spend, then according to the benefits principle, they should be taxed more heavily to pay for these programs.

13-3b The Ability-to-Pay Principle

Another way to evaluate the equity of a tax system is called the **ability-to-pay principle**, which states that taxes should be levied according to how well a person can shoulder the burden. It is sometimes said that everyone should make an "equal sacrifice" to support the government. The magnitude of a person's sacrifice, however, depends not only on the size of his tax payment but also on his income and other circumstances: A $1,000 tax paid by a person with a low income may require a larger sacrifice than a $10,000 tax paid by a person with a much higher income.

ability-to-pay principle
the idea that taxes should be levied on a person according to how well that person can shoulder the burden

The ability-to-pay principle leads to two corollary notions: vertical equity and horizontal equity. **Vertical equity** states that taxpayers with a greater ability to pay should contribute a larger amount. **Horizontal equity** states that taxpayers with similar abilities to pay should contribute the same amount. These notions of equity are widely accepted, but applying them to evaluate a tax system is rarely straightforward.

vertical equity
the idea that taxpayers with a greater ability to pay taxes should pay larger amounts

horizontal equity
the idea that taxpayers with similar abilities to pay taxes should pay the same amount

Vertical Equity If taxes are based on the ability to pay, then the rich should pay more than the poor. But how much more should the rich pay? The debate over tax policy often focuses on this question.

Table 4

Three Tax Systems

	Proportional Tax		Regressive Tax		Progressive Tax	
Income	Amount of Tax	Percent of Income	Amount of Tax	Percent of Income	Amount of Tax	Percent of Income
$50,000	$12,500	25%	$15,000	30%	$10,000	20%
100,000	25,000	25	25,000	25	25,000	25
200,000	50,000	25	40,000	20	60,000	30

proportional tax
a tax for which taxpayers at all income levels pay the same fraction of income

regressive tax
a tax for which taxpayers with high incomes pay a smaller fraction of their income than do taxpayers with low incomes

progressive tax
a tax for which taxpayers with high incomes pay a larger fraction of their income than do taxpayers with low incomes

Consider the three tax systems in Table 4. In each case, taxpayers with higher incomes pay more. Yet the systems differ in how quickly taxes rise with income. The first system is **proportional** because all taxpayers pay the same fraction of income. The second is **regressive** because high-income taxpayers pay a smaller fraction of their income, even though they pay a larger amount. The third system is **progressive** because high-income taxpayers pay a larger fraction of their income.

Which of these three tax systems is most fair? Economic theory alone does not offer an answer. Equity, like beauty, is in the eye of the beholder. As the next case study shows, the U.S. tax system is generally progressive, reflecting the consensus view that those with higher incomes should pay a larger percentage of their income in taxes.

Case Study

How the Tax Burden Is Distributed

The debate over tax policy often concerns whether those with high incomes pay their fair share. There is no objective way to make this judgment. In evaluating the issue for yourself, however, it is useful to know how much families with different incomes pay under the current tax system.

Table 5 presents some data on how federal taxes are distributed among income classes. These figures are for 2018, the most recent year available as this book was

Table 5

The Burden of Federal Taxes

Source: Congressional Budget Office and author's calculations. Figures are for 2018.

Quintile	Average Market Income	Taxes as a Percentage of Market Income	Taxes Less Transfers as a Percentage of Market Income
Lowest	$16,600	0.0%	–127.1%
Second	35,900	11.1	–44.0
Middle	63,900	15.5	–11.1
Fourth	104,000	18.8	4.6
Highest	310,000	25.4	21.3
Top 1%	1,987,500	30.3	29.7

going to press, and were tabulated by the Congressional Budget Office (CBO). They include all federal taxes—personal income taxes, payroll taxes, corporate income taxes, and excise taxes—but not state and local taxes.

To construct the table, households are ranked according to their income and placed into five groups of equal size, called **quintiles**. The table also presents data on the richest 1 percent of Americans (who represent the top sliver of the highest quintile). The second column of the table shows the average market income of each group. Market income measures what a household earns from its economic activity, including wages and salaries, business income, interest, capital gains, dividends, and pension benefits. The lowest quintile had average market income of $16,600, and the highest quintile had average market income of $310,000. The top 1 percent had average market income of almost $2 million.

The third column of the table shows total taxes as a percentage of income (the average tax rate). As you can see, the U.S. federal tax system is progressive. The lowest quintile of households paid about zero percent of their incomes in taxes: That's because the taxes they paid, such as the payroll tax, were offset by tax rebates they received, such as the earned income tax credit. The middle quintile paid 15.5 percent of their income in taxes. The highest quintile paid 25.4 percent, and the top 1 percent paid 30.3 percent.

These numbers on taxes are a good starting point for understanding how the burden of government is distributed, but they give an incomplete picture. Money flows not only from households to the government in the form of taxes but also from the government back to households in the form of transfer payments, including Social Security, unemployment insurance benefits, Medicare (a health program for older Americans), Medicaid (a health program for those with low incomes), SNAP benefits (a program formerly known as food stamps), and housing assistance. In some ways, transfer payments are the opposite of taxes.

Treating transfers as negative taxes substantially alters the distribution of the tax burden, as shown in the last column of the table. The change is small for high-income households: The highest quintile paid 21.3 percent of its income to the government, even after transfers are subtracted, and the top 1 percent paid 29.7 percent. But the average tax rates for the lowest three quintiles become negative numbers. That is, typical households in the bottom three-fifths of the income distribution received more in transfers than they paid in taxes. This is particularly true for those with the lowest incomes. While the lowest quintile had average market income of only $16,600, its average income after taxes and transfers was $37,700. The lesson is clear: To fully understand the progressivity of government policies, one must consider both what people pay and what they receive.

Finally, note that the numbers in Table 5 may be out of date. In 2021, President Biden proposed increasing taxes for those with the highest incomes and expanding tax credits for lower-income households with children. ●

Horizontal Equity If taxes are based on ability to pay, then similar taxpayers should pay similar amounts of taxes. But families differ in many ways. To judge whether a tax code is horizontally equitable, one must first decide which differences are relevant for a family's ability to pay and which are not.

Suppose the Garcia and Jackson families each have annual income of $100,000. The Garcias have no children, but Mr. Garcia has an illness that results in medical expenses of $30,000. The Jacksons are in good health, but they have three children,

two of whom are in college, generating tuition bills of $60,000. Would it be fair for these two families to pay the same tax because they have the same income? Would it be fair to give the Garcias a tax break to help them offset their high medical expenses? Would it be fair to give the Jacksons a tax break to help them with their tuition expenses?

These questions do not have easy answers. In practice, the U.S. tax code is filled with special provisions that alter a family's tax obligations based on its specific circumstances.

13-3c Tax Incidence and Tax Equity

Tax incidence—the study of who bears the burden of taxes—is central to evaluating tax equity. As Chapter 6 pointed out, the person who bears the burden of a tax is not always the person who gets the tax bill from the government. Because taxes alter supply and demand, they alter equilibrium prices. As a result, they affect people beyond those who, according to statute, pay the tax. When evaluating the vertical and horizontal equity of any tax, it is important to take these indirect effects into account.

Many discussions of tax equity ignore the indirect effects of taxes and are based on what economists mockingly call the **flypaper theory** of tax incidence. According to this theory, the burden of a tax, like a fly on flypaper, sticks wherever it first lands. This assumption, however, is rarely valid.

For example, a person not trained in economics might argue that a tax on expensive fur coats is vertically equitable because most buyers of furs are wealthy. Yet if these buyers can easily substitute other luxuries for furs, then a tax on furs might only reduce the sale of furs. In the end, the burden of the tax will fall more on those who make and sell furs than on those who buy them. Because most workers who make furs are not wealthy, the equity of a fur tax could be quite different from what the flypaper theory indicates.

Case Study

Who Pays the Corporate Income Tax?

The corporate income tax provides a good example of the importance of tax incidence for tax policy. The corporate tax is popular among some voters. After all, corporations are not people. Voters are always eager to get a tax cut and let some impersonal corporation pick up the tab.

BILL PUGLIANO/GETTY IMAGES NEWS/GETTY IMAGES

This worker pays part of the corporate income tax.

But before deciding that the corporate income tax is a good way for the government to raise revenue, consider who bears the burden of the corporate tax. This is a difficult question on which economists disagree, but one thing is certain: **People pay all taxes**. When the government levies a tax on a corporation, the corporation is more like a tax collector than a taxpayer. The burden of the tax ultimately falls on people—the owners, customers, or workers of the corporation.

Why might workers and customers pay part of the corporate income tax? Consider an example. Suppose that the U.S. government decides to raise the tax on the income earned by car companies. At first, this tax hurts

the companies' owners, who receive less after-tax profit. But over time, these owners will respond to the tax. Because producing cars is less profitable, they invest less in building new car factories. Instead, they invest their wealth in other ways—for example, by buying larger houses or by building factories in other industries or countries. With fewer domestic car factories, the supply of cars declines, as does the demand for autoworkers. A tax on corporations making cars causes the price of cars to rise and the wages of autoworkers to fall.

This issue rose to prominence in the early days of the Trump administration. The tax bill signed into law by President Trump in 2017 cut the corporate tax rate from 35 to 21 percent. The president's economic advisers argued that the long-term effect of the policy would be increased capital accumulation, productivity, and wages. Critics of the bill agreed that these growth effects would occur but believed they would be small. In their view, the main benefits of the corporate tax cut would accrue to the corporations' owners, who tend to be wealthy. Yet advocates and critics agreed on this: Evaluating the fairness of any tax change requires paying careful attention to tax incidence. ●

QuickQuiz

6. A toll is a tax on citizens who use toll roads. This policy can be viewed as an application of
   a. the benefits principle.
   b. horizontal equity.
   c. vertical equity.
   d. tax progressivity.
7. In the United States, taxpayers in the top 1 percent of the income distribution pay about _____ percent of their income in federal taxes.
   a. 5
   b. 10
   c. 20
   d. 30
8. If the corporate income tax induces businesses to reduce their capital investment, then
   a. the tax does not have any deadweight loss.
   b. corporate shareholders benefit from the tax.
   c. workers bear some of the burden of the tax.
   d. the tax achieves the goal of vertical equity.

Answers are at the end of the chapter.

13-4 Conclusion: The Trade-Off between Equity and Efficiency

Equity and efficiency are the two most important goals of a tax system. But these two goals can conflict, especially when equity is judged by progressivity. People often disagree about tax policy because they attach different weights to these goals.

The history of tax policy shows how political leaders differ in their views on equity and efficiency. When Ronald Reagan was elected president in 1980, the marginal tax rate on the earnings of the richest Americans was 50 percent. On interest income, the marginal tax rate was 70 percent. Reagan argued that such high tax rates greatly

In the News

The Value-Added Tax

The United States does not have a value-added tax, but the tax is common throughout the rest of the world.

How about This? A Tax That Discourages Tax Evasion

By Seema Jayachandran

You may not like paying taxes, but your burden will be even higher if others don't pay their fair share.

That, in a nutshell, goes a long way toward explaining why a value-added tax, or VAT, is used by just about every major country in the world except the United States.

Although the concept has never caught on in this country, the VAT has been a powerful, well-mannered weapon for progress. That is especially true in less-affluent nations, because it reduces tax evasion in a relatively effective and gentle way, as an emerging body of research shows.

Rich countries collect 34 percent of gross domestic product in taxes on average (the United States collects about 27 percent). But most low- and middle-income countries—a range of places, from Mali to Malaysia—collect much less, typically just 10 to 20 percent of G.D.P. The main reason is that it is harder to collect taxes in less-advanced nations, where a larger share of economic activity is informal, making tax evasion easier. Yet tax revenue, which enables governments to provide public goods like roads and schools, is at least as important in poorer countries.

Enter the VAT, which taxes personal income indirectly by collecting it from businesses. A value-added tax also has a built-in, self-enforcing feature.

Here is how it works.

At its most basic, a VAT taxes the value a business adds to a good or service as it is being produced. The added value may be thought of as the price at which the business sells its product minus the cost of producing it.

For example, a wholesale bakery earns revenue by selling bread to grocery stores. Subtract from that the bakery's spending on flour, yeast and other ingredients. The difference is the added value on which the bakery is taxed. In addition, the flour company would pay a VAT on the revenue from its flour minus what it pays for wheat and the like.

When you tally the value added at every stop on the supply chain, from wheat farmer to bread eater, you get the retail price of the bread. Thus a VAT is a tax on consumption and it can be easier to administer than a personal income tax.

For one thing, the government collects the VAT from businesses (of which there are relatively few) rather than households (of which there are many). For another, especially in less-advanced economies, personal income tax systems are rife with problems because so many people are self-employed, with latitude to fudge their reported income.

In the United States, the Internal Revenue Service estimates, self-employment and farm income are underreported by over 60 percent. Even so, the personal income tax system in the United States works pretty well because most people have an employer. That's not the case in many other countries.

A VAT resembles a sales tax, with an important difference: It is paid at every stage of production, not just the point of sale. That makes a VAT marvelously self-enforcing, because one firm's tax deductions are another firm's tax liability.

When the baker buys flour, it is in her financial interest to inform the tax authority

distorted incentives to work and save. In other words, he claimed that these high tax rates cost too much in terms of efficiency. Reagan signed into law large cuts in tax rates in 1981 and then again in 1986. When Reagan left office in 1989, the top marginal tax rate was only 28 percent.

When Bill Clinton ran for president in 1992, he argued that the rich were not paying their fair share of taxes. In other words, the low tax rates on the rich violated his view of vertical equity. In 1993, President Clinton signed into law a bill that raised the top marginal tax rate to about 40 percent.

In the years that followed, the pendulum of political debate continued to swing. President George W. Bush reprised many of Reagan's themes and reduced the

about the purchase, so that she can deduct the cost from her tax base. That information alerts the tax authority about the flour producer's income.

These theoretical advantages are well known, but Dina Pomeranz, an economist at the University of Zurich, found a way to test how important they are in practice. In a study published in the American Economic Review in 2015, she collaborated with the tax authority in Chile, using the prospect of audits to uncover where fraud was rampant.

The Chilean tax authority sent letters to a randomly selected set of firms, all of which were required to pay a VAT, informing them that they were under special scrutiny. Professor Pomeranz found that after the letters were received, the reporting of business-to-business sales hardly budged, suggesting that income was already being reported accurately.

In contrast, businesses like grocery stores that sold to consumers started reporting more income and paying more VAT.

Why was there a difference? With a VAT, there is a last-mile problem: A store customer buying bread had no incentive to ask for a receipt, because she couldn't deduct the cost from her income as businesses could, so many business-to-consumer transactions were underreported.

Governments have tried to solve the last-mile problem by adding a reward for customers who ask for a receipt. The state of São Paulo, Brazil, started a program in 2007 that gave customers a rebate of roughly 1 percent on their retail receipts. (To make the program more enticing, some of the rebate is in the form of raffle tickets for prizes worth up to $500,000.)

Retailers had to submit the receipts to the tax authority, and consumers in São Paulo could check online whether stores had submitted their receipts and complain to the tax authority if they hadn't. In essence, shoppers were enlisted as citizen tax auditors.

The Brazilian program succeeded in increasing tax collection among retailers 21 percent, according to a research study by Joana Naritomi at the London School of Economics that analyzed the program's first four years and is scheduled to appear in the American Economic Review.

The success of VAT systems in emerging economies is welcome news for the beneficiaries of that tax revenue. It might be useful one day in the United States, too, and for another reason. Because a VAT taxes consumption, rather than income, it tends to encourage saving and investment.

It has its problems, too. A big one is that it is regressive. If you have barely enough money to get by, you can't afford to save: Your entire income is spent on essentials and, thus, taxed. Richer people are taxed on a much smaller share of their income.

But a VAT can be paired with progressive tax policies, like a tax credit for low-income people, that counterbalance this deficiency, as argued by William Gale of the Brookings Institution in his book "Fiscal Therapy."

In the end, a VAT's major advantage is that it can bring in a great deal of revenue with comparatively little administrative cost and without as much need for intrusive enforcement. It could be an appealing option for the United States one day. ■

Source: *New York Times*, May 19, 2019.

Questions to Discuss

1. Would you favor the introduction of a value-added tax in the United States? Why or why not?
2. Why do you think the United States differs from so many other nations in not having value-added tax?

Ms. Jayachandran is an economics professor at Northwestern University.

top tax rate to 35 percent in 2003. President Barack Obama again emphasized vertical equity, and in 2013, the top rate was back at about 40 percent. But then Donald Trump was elected president, and he signed into law a cut in the top rate to 37 percent starting in 2018. In 2021, President Biden proposed raising it to 45 percent for taxpayers with annual incomes above $25 million.

Economics alone cannot determine the best way to balance the goals of efficiency and equity. This issue involves political philosophy as well as economics. But economists have an important role in this debate: They can shed light on the trade-offs that society inevitably faces when designing the tax system and can help avoid policies that sacrifice efficiency without enhancing equity.

Chapter in a Nutshell

- The government raises revenue using a variety of taxes. The most important ones for the U.S. federal government are personal income taxes and payroll taxes for social insurance. The most important for state and local governments are property taxes, personal income taxes, and sales taxes.
- The efficiency of a tax system involves the costs it imposes on taxpayers. There are two costs of taxes beyond the transfer of resources from the taxpayer to the government. The first is the deadweight loss that arises as taxes alter incentives and distort the allocation of resources. The second is the administrative burden of complying with the tax laws.
- The equity of a tax system concerns whether the tax burden is distributed fairly among the population. According to the benefits principle, it is fair for people to pay taxes based on the benefits they receive from the government. According to the ability-to-pay principle, it is fair for people to pay taxes based on their capacity to handle the financial burden. When evaluating the equity of a tax system, it is important to remember a lesson from the study of tax incidence: The distribution of tax burdens is not the same as the distribution of tax bills.
- When considering changes in the tax laws, policymakers often face a trade-off between efficiency and equity. Much of the debate over tax policy arises because people give different weights to these two goals.

Key Concepts

average tax rate, p. 255
marginal tax rate, p. 255
lump-sum tax, p. 255
benefits principle, p. 257
ability-to-pay principle, p. 257
vertical equity, p. 257
horizontal equity, p. 257
proportional tax, p. 258
regressive tax, p. 258
progressive tax, p. 258

Questions for Review

1. Over the past century, has the government's tax revenue grown more or less slowly than the rest of the economy?
2. Explain how corporate profits are taxed twice.
3. Why is the burden of a tax to taxpayers greater than the revenue received by the government?
4. Why do some economists advocate taxing consumption rather than income?
5. What is the marginal tax rate on a lump-sum tax? How is this related to the efficiency of the tax?
6. Give two arguments why people with high incomes should pay more taxes than people with low incomes.
7. What is the concept of horizontal equity, and why is it hard to apply?

Problems and Applications

1. The information in many of the tables in this chapter can be found in the *Economic Report of the President*, which appears annually. Using a recent issue of the report at your library or on the Internet, answer the following questions and provide some numbers to support your answers. (Hint: The website of the Government Printing Office is www.gpo.gov.)
   a. Figure 1 shows that government revenue as a percentage of total income has increased over time. Is this increase primarily attributable to changes in federal government revenue or in state and local government revenue?
   b. Looking at the combined revenue of the federal government and state and local governments, how has the composition of total revenue changed over time? Are personal income taxes more or less important? Social insurance taxes? Corporate profits taxes?

2. Suppose you are a typical person in the U.S. economy. You pay 4 percent of your income in a state income tax and 15.3 percent of your labor earnings in federal payroll taxes (employer and employee shares combined). You also pay federal income taxes as in Table 2. How much tax of each type do you pay if you earn $40,000 a year? Taking all taxes into account, what are your average and marginal tax rates? What happens to your tax bill and to your average and marginal tax rates if your income rises to $80,000?
3. Some states exclude necessities, such as food and clothing, from their sales tax. Other states do not. Discuss the merits of this exclusion. Consider both efficiency and equity.
4. Prior to 2018, state and local taxes (SALT) could be fully deducted when calculating taxable income for the federal income tax. Starting in 2018, the SALT deduction was limited to $10,000.
   a. Who benefited most from the unlimited SALT deduction—residents of high-tax states like California and New York or residents of low-tax states like Florida and Texas?
   b. How do you think the limitation of the SALT deduction affected the migration of people among the states?
   c. How do you think the limitation of the SALT deduction affected the propensity of state and local governments to raise taxes?
5. When Serena Saver owns an asset (such as a share of stock) that rises in value, she has an "accrued" capital gain. If she sells the asset, she "realizes" the gains that have previously accrued. Under the U.S. income tax system, realized capital gains are taxed, but accrued gains are not.
   a. Explain how individuals' behavior is affected by this rule.
   b. Some economists believe that cuts in capital gains tax rates, especially temporary ones, can raise tax revenue. How might this be so?
   c. Do you think it is a good rule to tax realized but not accrued capital gains? Why or why not?
6. Suppose that your state raises its sales tax from 5 percent to 6 percent. The state revenue commissioner forecasts a 20 percent increase in sales tax revenue. Is this plausible? Explain.
7. The Tax Reform Act of 1986 eliminated the deductibility of interest payments on consumer debt (mostly credit cards and auto loans) but maintained the deductibility of interest payments on mortgages and home equity loans. What do you think happened to the relative amounts of borrowing through consumer debt and home equity debt?
8. Categorize each of the following funding schemes as examples of the benefits principle or the ability-to-pay principle.
   a. Visitors to many national parks pay an entrance fee.
   b. Local property taxes support elementary and secondary schools.
   c. An airport trust fund collects a tax on each plane ticket sold and uses the money to improve airports and the air traffic control system.

QuickQuiz Answers

1. **b** 2. **b** 3. **c** 4. **a** 5. **a** 6. **a** 7. **d** 8. **c**

Chapter

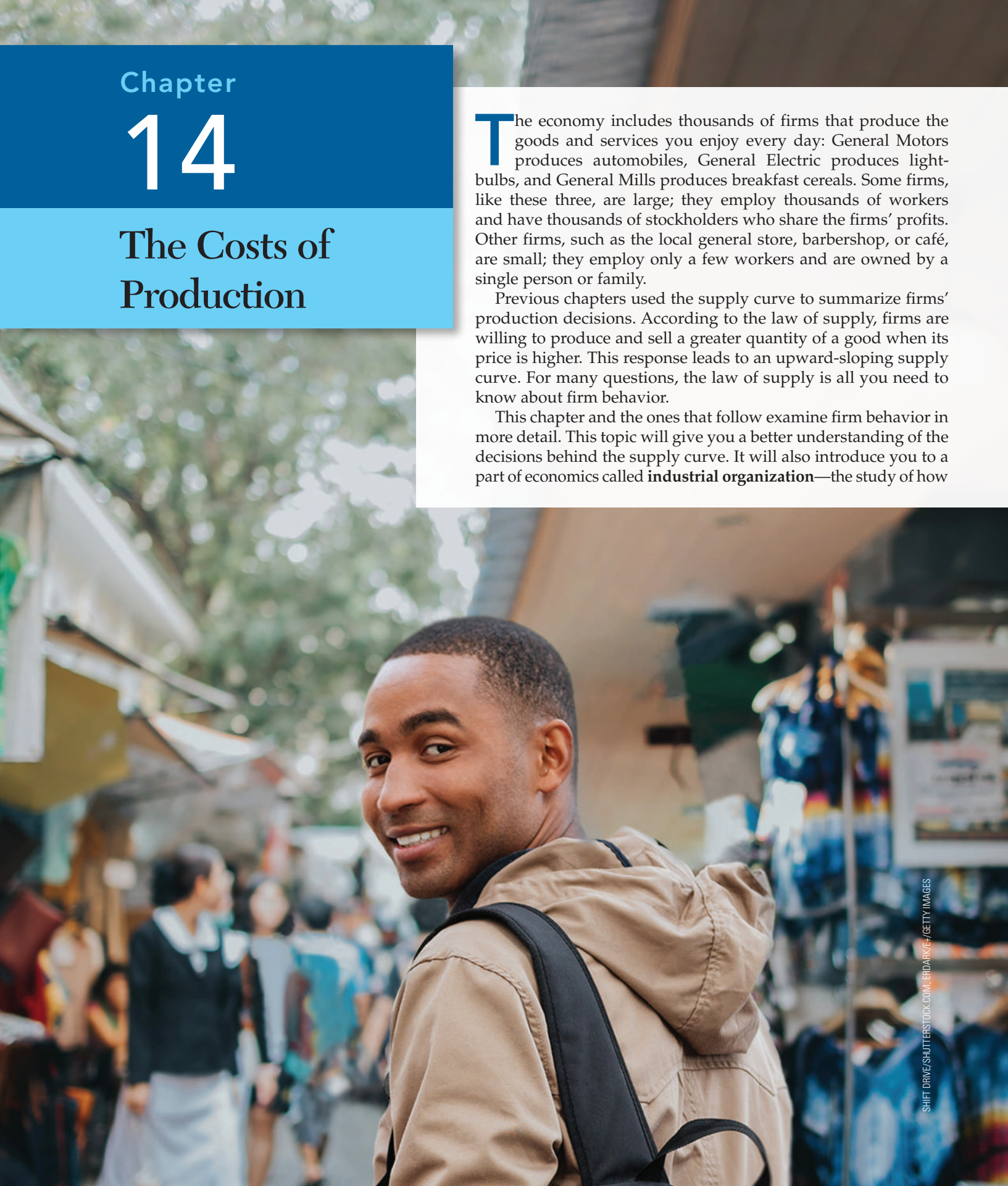
14

The Costs of Production

The economy includes thousands of firms that produce the goods and services you enjoy every day: General Motors produces automobiles, General Electric produces lightbulbs, and General Mills produces breakfast cereals. Some firms, like these three, are large; they employ thousands of workers and have thousands of stockholders who share the firms' profits. Other firms, such as the local general store, barbershop, or café, are small; they employ only a few workers and are owned by a single person or family.

Previous chapters used the supply curve to summarize firms' production decisions. According to the law of supply, firms are willing to produce and sell a greater quantity of a good when its price is higher. This response leads to an upward-sloping supply curve. For many questions, the law of supply is all you need to know about firm behavior.

This chapter and the ones that follow examine firm behavior in more detail. This topic will give you a better understanding of the decisions behind the supply curve. It will also introduce you to a part of economics called **industrial organization**—the study of how



SHIFT DRIVE/SHUTTERSTOCK.COM; ERDARYE+/GETTY IMAGES

firms' decisions about prices and quantities depend on the market conditions they face. The town in which you live, for instance, may have several pizzerias but only one cable television company. This raises a key question: How does the number of firms affect the prices in a market and the efficiency of the market outcome? The field of industrial organization addresses this question.

Before turning to these issues, it's important to understand the costs of production. All firms, from Delta Air Lines to your local deli, incur costs while making the goods and services they sell. As the coming chapters show, a firm's costs are a key determinant of its production and pricing decisions. This chapter defines the variables that economists use to measure such costs and considers the relationships among these variables.

A word of warning: This topic is a bit technical. To be frank, you might even call it boring. Hang in there. This material provides the foundation for the fascinating topics that follow.

14-1 What Are Costs?

Enter Chloe's Cookie Factory. Chloe, the owner, bakes irresistible cookies. To do so, she buys flour, sugar, chocolate chips, and other cookie ingredients. She also buys mixers and ovens and hires workers to run this equipment for her. She then sells the cookies to grateful consumers. Some of the issues that Chloe faces in her business apply to all firms.

14-1a Total Revenue, Total Cost, and Profit

To understand the decisions a firm makes, let's begin by asking what it is trying to do. Chloe may have started her firm because of an altruistic desire to provide the world with cookies or simply out of love for the cookie business, but the truth, in her case, is that she started it to make money. Economists typically assume that the goal of a firm is to maximize profit, and they find that this assumption works well in most cases.

What is a firm's profit? The amount that it receives for the sale of its output (cookies) is **total revenue**. The amount that it pays to buy inputs (flour, sugar, workers, ovens, and so forth) is **total cost**. As the business owner, Chloe gets to keep any revenue above her costs. That is, a firm's **profit** equals its total revenue minus its total cost:

total revenue
the amount a firm receives for the sale of its output

total cost
the market value of the inputs a firm uses in production

profit
total revenue minus total cost

$$\text{Profit} = \text{Total revenue} - \text{Total cost}.$$

Chloe wants to make her profit as large as possible.

To see how a firm maximizes profit, the first step is to measure its total revenue and total cost. Total revenue is simple: It is the quantity of output the firm produces multiplied by the sales price. If Chloe produces 10,000 cookies and sells them at $2 a cookie, her total revenue is $20,000. The measurement of a firm's total cost, however, is more subtle.

14-1b Why Opportunity Costs Matter

When measuring costs at Chloe's Cookie Factory or any other firm, keep in mind one of the **Ten Principles of Economics** from Chapter 1: The cost of something is what you give up to get it. Recall that the **opportunity cost** of an item refers to all the things that must be forgone to acquire it. When economists speak of a firm's cost of production, they include all the opportunity costs of making its output of goods and services.

Some of these opportunity costs are obvious. When Chloe pays $1,000 for flour, that $1,000 is an opportunity cost because Chloe can no longer use that $1,000 to buy something else. Similarly, when Chloe hires workers to make the cookies, their wages are part of the firm's costs. Because these opportunity costs require the firm to pay out some money, they are called **explicit costs**.

explicit costs
input costs that require an outlay of money by the firm

By contrast, some of a firm's opportunity costs, called **implicit costs**, do not require a cash outlay. These costs may not be immediately obvious, but they are nonetheless meaningful. Imagine that Chloe is good with computers and could earn $100 per hour working as a programmer. For every hour that she works at her cookie factory, she gives up $100 in programming income, and this forgone income is also part of her costs. The total cost of Chloe's business is the sum of her explicit and implicit costs.

implicit costs
input costs that do not require an outlay of money by the firm

Economists and accountants analyze businesses differently. Economists are interested in how firms make production and pricing decisions. Because these decisions are based on both explicit and implicit costs, economists include both when measuring a firm's costs. But accountants keep track of the money that flows into and out of firms, so they measure the explicit costs but usually ignore the implicit ones.

The difference between the methods of economists and accountants is easy to see at Chloe's Cookie Factory. When Chloe gives up the opportunity to earn money as a programmer, her accountant will not count this as a cost of her cookie business. Because no money flows out of the business to pay for this cost, it never appears on the accountant's financial statements. An economist, however, will count the forgone income as a cost because it will affect the decisions that Chloe makes in her cookie business. For example, if Chloe's wage as a programmer rises from $100 to $300 per hour, she might decide that running her cookie business is too costly. She might choose to shut down the factory so she can work as a programmer.

14-1c The Cost of Capital Is an Opportunity Cost

An implicit cost of almost every business is the opportunity cost of the money—economists call it financial capital—that has been invested in it. Suppose, for instance, that Chloe used $300,000 of her savings to buy the cookie factory. If she had instead left this money in a savings account that paid an interest rate of 5 percent, she would have earned $15,000 per year. To own her cookie factory, therefore, Chloe has given up $15,000 a year in interest income. This forgone $15,000 is one of the implicit opportunity costs of Chloe's business.

The cost of capital is a prime case in which economists and accountants view businesses differently. An economist views the $15,000 in interest income that Chloe gives up every year as an implicit cost. Chloe's accountant, however, will not show this $15,000 as a cost because no money flows out of the business to pay for it.

To further explore the difference between the methods of economists and accountants, let's change the example slightly. Suppose that Chloe did not have the entire $300,000 to buy the factory but instead, used $100,000 of her own savings and borrowed $200,000 from a bank at an interest rate of 5 percent. Chloe's accountant, who only measures explicit costs, will now count the $10,000 interest paid on the bank loan every year as a cost because this money flows out of the firm. By contrast, according to an economist, the opportunity cost of owning the business is still $15,000. The opportunity cost equals the interest on the bank loan (an explicit cost of $10,000) plus the forgone interest on savings (an implicit cost of $5,000).

14-1d Economists and Accountants Measure Profit Differently

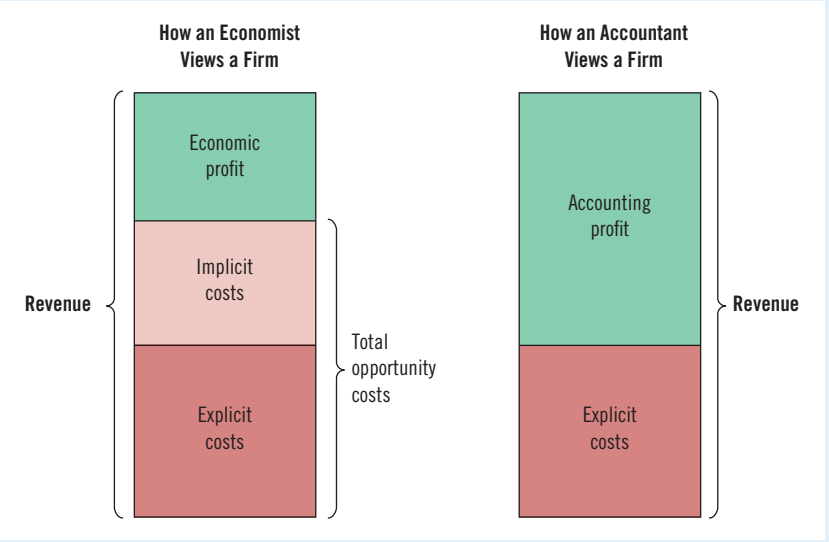
Now let's return to the firm's objective: profit. Economists and accountants measure this differently, too. An economist measures a firm's **economic profit** as its total

economic profit
total revenue minus total cost, including both explicit and implicit costs

Figure 1

Economists versus Accountants

Because economists include all opportunity costs when analyzing a firm, while accountants measure only explicit costs, economic profit is smaller than accounting profit.



accounting profit
total revenue minus total explicit cost

revenue minus all its opportunity costs (explicit and implicit) of producing the goods and services sold. An accountant measures the firm's **accounting profit** as its total revenue minus only its explicit costs.

Figure 1 summarizes this difference. Notice that because the accountant ignores the implicit costs, accounting profit is larger than economic profit. For a business to be profitable from an economist's standpoint, total revenue must exceed all the opportunity costs, both explicit and implicit.

Economic profit is an important concept because it motivates the firms that supply goods and services. As we will see, a firm that makes positive economic profit will stay in business. It is covering all its opportunity costs and has some revenue left to reward the firm's owners. When a firm makes economic losses (that is, when economic profits are negative), it is failing to bring in enough revenue to cover all the costs of production. Unless conditions change, the owners will eventually close the business and exit the industry. To understand business decisions, keep an eye on economic profit.

QuickQuiz

1. Farmer McDonald gives banjo lessons for $20 per hour. One day, he spends 10 hours planting $100 worth of seeds on his farm. What total cost has he incurred?
   a. $100
   b. $200
   c. $300
   d. $400

2. Xavier opens a lemonade stand for two hours. He spends $10 for ingredients and sells $60 worth of lemonade. In the same two hours, he could have mowed his neighbor's lawn for $40. Xavier earns an accounting profit of _____ and an economic profit of _____.
   a. $50; $10
   b. $90; $50
   c. $10; $50
   d. $50; $90

Answers are at the end of the chapter.

14-2 Production and Costs

Firms incur costs when they buy inputs to produce the goods and services they plan to sell. This section examines the link between a firm's production process and its total cost. Once again, consider Chloe's Cookie Factory.

The analysis that follows makes a simplifying assumption: The size of Chloe's factory is fixed, and Chloe can vary the quantity of cookies produced only by changing the number of workers she employs. This assumption is realistic in the short run but not in the long run. That is, Chloe cannot build a larger factory overnight, but she could do so over the next year or two. This analysis, therefore, describes the production decisions that Chloe faces in the short run. The relationship between costs and time horizon is examined more fully later in the chapter.

14-2a The Production Function

Table 1 shows how the quantity of cookies produced per hour at Chloe's factory depends on the number of workers. As you can see in columns (1) and (2), if there are no workers in the factory, Chloe produces no cookies. When there is 1 worker, she produces 50 cookies. When there are 2 workers, she produces 90 cookies and so on. Panel (a) of Figure 2 presents a graph of these two columns of numbers. The number of workers is on the horizontal axis, and the number of cookies produced is on the vertical axis. This relationship between the quantity of inputs (workers) and quantity of output (cookies) is called the **production function**.

production function the relationship between the quantity of inputs used to make a good and the quantity of output of that good

Table 1

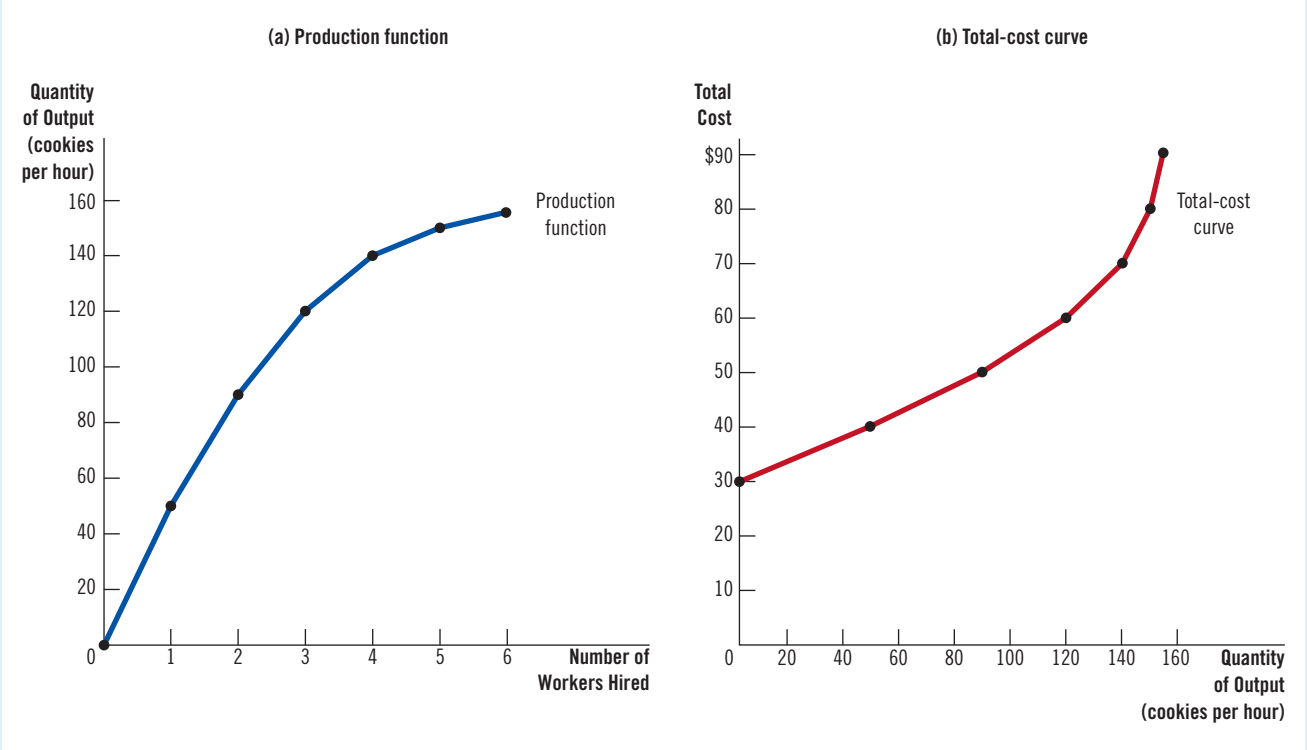
A Production Function and Total Cost: Chloe's Cookie Factory

(1)	(2)	(3)	(4)	(5)	(6)
Number of Workers	Output (quantity of cookies produced per hour)	Marginal Product of Labor	Cost of Factory	Cost of Workers	Total Cost of Inputs (cost of factory + cost of workers)
0	0		$30	$0	$30
		50			
1	50		30	10	40
		40			
2	90		30	20	50
		30			
3	120		30	30	60
		20			
4	140		30	40	70
		10			
5	150		30	50	80
		5			
6	155		30	60	90

Figure 2

Chloe's Production Function and Total-Cost Curve

The production function in panel (a) shows the relationship between the number of workers hired and the quantity of output produced. Here, the number of workers hired (on the horizontal axis) is from column (1) in Table 1, and the quantity of output (on the vertical axis) is from column (2). The production function gets flatter as the number of workers increases, reflecting diminishing marginal product. The total-cost curve in panel (b) shows the relationship between the quantity of output and total cost of production. Here, the quantity of output produced (on the horizontal axis) is from column (2) in Table 1, and the total cost (on the vertical axis) is from column (6). The total-cost curve gets steeper as the quantity of output increases because of diminishing marginal product.



marginal product
the increase in output that arises from an additional unit of input

One of the **Ten Principles of Economics** in Chapter 1 is that rational people think at the margin. As future chapters show, this idea is the key to understanding the decisions a firm makes about how many workers to hire and how much output to produce. To take a step toward understanding these decisions, column (3) in the table gives the marginal product of a worker. The **marginal product** of any input in the production process is the change in the quantity of output obtained from one additional unit of that input. When the number of workers goes from 1 to 2, cookie production increases from 50 to 90, so the marginal product of the second worker is 40 cookies. When the number of workers goes from 2 to 3, cookie production increases from 90 to 120, so the marginal product of the third worker is 30 cookies. In the table, the marginal product is shown halfway between two rows because it represents the change in output as the number of workers increases from one level to another.

Notice that as the number of workers increases, the marginal product declines. The second worker has a marginal product of 40 cookies, the third has a marginal product of 30 cookies, and the fourth has a marginal product of 20 cookies. This property is called **diminishing marginal product**. At first, when there are only a few workers at the factory, they have easy access to the kitchen equipment. As Chloe increases hiring, the workers must share equipment and deal with more crowded conditions. Eventually, the kitchen becomes so overcrowded that workers get in each other's way. Hence, as more workers are hired, each extra worker contributes fewer additional cookies to total production.

diminishing marginal product
the property whereby the marginal product of an input declines as the quantity of the input increases

Diminishing marginal product is apparent in Figure 2. The production function's slope ("rise over run") measures the change in Chloe's output of cookies ("rise") for each additional input of labor ("run"). That is, the slope of the production function measures the marginal product. As the number of workers increases, the marginal product declines, and the production function becomes flatter.

14-2b From the Production Function to the Total-Cost Curve

Columns (4), (5), and (6) in Table 1 show Chloe's cost of producing cookies. In this example, the cost of Chloe's factory is $30 per hour, and the cost of hiring a worker is $10 per hour. If she hires 1 worker, her total cost is $40 per hour. If she hires 2, her total cost is $50 per hour, and so on. This information ties things together. The table shows how the number of workers hired determines the quantity of cookies produced and the total cost of production.

The goal of the next several chapters is to study firms' production and pricing decisions. For this purpose, the most important relationship in Table 1 is between quantity produced [in column (2)] and total cost [in column (6)]. Panel (b) of Figure 2 graphs these two columns of data with quantity produced on the horizontal axis and total cost on the vertical axis. This graph is called the **total-cost curve**.

Now compare the total-cost curve in panel (b) with the production function in panel (a). These two curves are opposite sides of the same coin. The total-cost curve grows steeper as the amount produced rises, while the production function becomes flatter as production rises. These changes in slope occur for the same reason. High production of cookies means that Chloe's kitchen is crowded with workers. In this case, because of diminishing marginal product, each additional worker adds little to production. That's why the production function is relatively flat. But flip this logic around: When the kitchen is crowded, producing an additional cookie requires a lot more labor, which is costly. Therefore, when the quantity produced is large, the total-cost curve is relatively steep.

QuickQuiz

3. Farmer Greene faces diminishing marginal product. If she plants no seeds on her farm, she gets no harvest. If she plants 1 bag of seeds, she gets 3 bushels of wheat. If she plants 2 bags, she gets 5 bushels. If she plants 3 bags, she gets
   a. 6 bushels.
   b. 7 bushels.
   c. 8 bushels.
   d. 9 bushels.

4. Diminishing marginal product explains why, as output increases,
   a. the production function and total-cost curve both get steeper.
   b. the production function and total-cost curve both get flatter.
   c. the production function gets steeper, and the total-cost curve gets flatter.
   d. the production function gets flatter, and the total-cost curve gets steeper.

Answers are at the end of the chapter.

14-3 The Many Measures of Cost

The analysis of Chloe's Cookie Factory showed how a firm's total cost reflects its production function. From data on total cost, we can derive several related measures of cost that will be later useful in analyzing production and pricing decisions. Now, consider the example in Table 2, which presents cost data on Chloe's neighbor—Caleb's Coffee Shop.

Column (1) in the table shows the number of cups of coffee that Caleb might produce, ranging from 0 to 10 cups per hour. Column (2) shows Caleb's total cost of producing coffee. Figure 3 plots Caleb's total-cost curve. The quantity of coffee [from column (1)] is on the horizontal axis, and total cost [from column (2)] is on the vertical axis. Caleb's total-cost curve has a shape like Chloe's. In particular,

Table 2

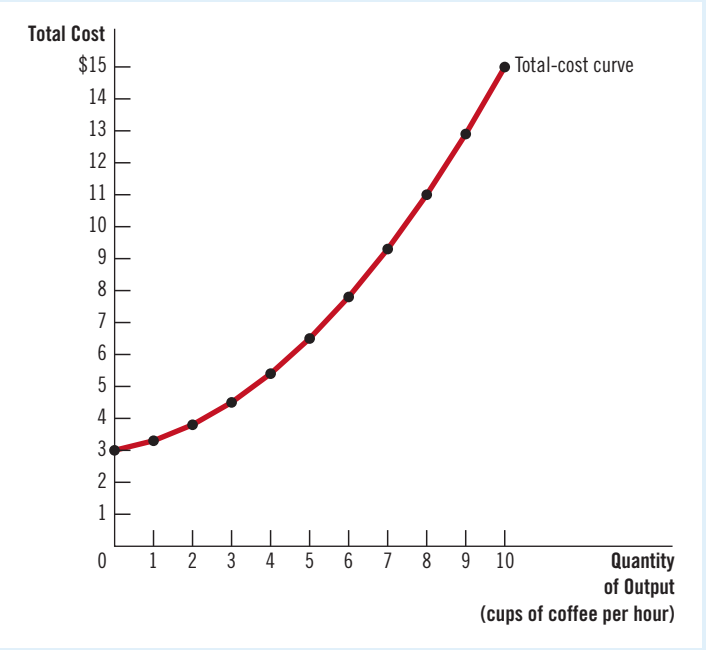
The Various Measures of Cost: Caleb's Coffee Shop

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Output (cups of coffee per hour)	Total Cost	Fixed Cost	Variable Cost	Average Fixed Cost	Average Variable Cost	Average Total Cost	Marginal Cost
0	$3.00	$3.00	$0.00	—	—	—	
							$0.30
1	3.30	3.00	0.30	$3.00	$0.30	$3.30	
							0.50
2	3.80	3.00	0.80	1.50	0.40	1.90	
							0.70
3	4.50	3.00	1.50	1.00	0.50	1.50	
							0.90
4	5.40	3.00	2.40	0.75	0.60	1.35	
							1.10
5	6.50	3.00	3.50	0.60	0.70	1.30	
							1.30
6	7.80	3.00	4.80	0.50	0.80	1.30	
							1.50
7	9.30	3.00	6.30	0.43	0.90	1.33	
							1.70
8	11.00	3.00	8.00	0.38	1.00	1.38	
							1.90
9	12.90	3.00	9.90	0.33	1.10	1.43	
							2.10
10	15.00	3.00	12.00	0.30	1.20	1.50	

Figure 3

Caleb's Total-Cost Curve

Here, the quantity of output produced (on the horizontal axis) is from column (1) in Table 2, and the total cost (on the vertical axis) is from column (2). As in Figure 2, the total-cost curve gets steeper as the quantity of output increases, reflecting diminishing marginal product.



it becomes steeper as the quantity produced rises, which (as we have discussed) reflects diminishing marginal product.

14-3a Fixed and Variable Costs

Caleb's total cost can be divided into two types. **Fixed costs** do not vary with the quantity of output produced. They are incurred even if the firm produces nothing at all. Caleb's fixed costs include rent, which is the same regardless of how much coffee he produces. Similarly, if Caleb needs to hire a full-time bookkeeper to pay bills, regardless of the quantity of coffee produced, the bookkeeper's salary is a fixed cost. The third column in Table 2 shows Caleb's fixed cost, which in this example is $3.00.

fixed costs
costs that do not vary with the quantity of output produced

On the other hand, **variable costs** change as the firm alters the quantity of output produced. Caleb's variable costs include the cost of coffee beans, milk, sugar, and paper cups: The more coffee Caleb produces, the more of these items he needs to buy. Similarly, if Caleb has to hire more workers to make more cups of coffee, the salaries of these workers are variable costs. Column (4) in the table shows Caleb's variable cost. The variable cost is 0 if he produces nothing, $0.30 if he produces 1 cup of coffee, $0.80 if he produces 2 cups, and so on.

variable costs
costs that vary with the quantity of output produced

A firm's total cost is the sum of fixed and variable costs. In Table 2, total cost in column (2) equals fixed cost in column (3) plus variable cost in column (4).

14-3b Average and Marginal Cost

As the owner, Caleb decides how much to produce. When making this decision, he will want to consider how the level of production affects his costs. Caleb might ask his production supervisor the following two questions about the cost of producing coffee:

- How much does it cost to make the typical cup of coffee?
- How much does it cost to increase production of coffee by 1 cup?

These questions might seem to have the same answer, but they do not. Both answers are important for understanding how firms make production decisions.

To find the cost of the typical unit produced, divide the firm's costs by the quantity of output it produces. For example, if the firm produces 2 cups of coffee per hour, its total cost is $3.80, and the cost of the typical cup is $3.80/2, or $1.90. Total cost divided by the quantity of output is **average total cost**. Because total cost is the sum of fixed and variable costs, average total cost can be expressed as the sum of average fixed cost and average variable cost. **Average fixed cost** equals the fixed cost divided by the quantity of output, and **average variable cost** equals the variable cost divided by the quantity of output.

average total cost
total cost divided by the quantity of output

average fixed cost
fixed cost divided by the quantity of output

average variable cost
variable cost divided by the quantity of output

marginal cost
the increase in total cost that arises from an extra unit of production

Average total cost tells us the cost of the typical unit, but it does not say how much total cost will change as the firm alters its production level. Column (8) in Table 2 shows the amount that total cost rises when the firm increases production by 1 unit of output. This number is called **marginal cost**. For example, if Caleb increases production from 2 to 3 cups, total cost rises from $3.80 to $4.50, so the marginal cost of the third cup of coffee is $4.50 minus $3.80, or $0.70. In the table, the marginal cost appears halfway between any two rows because it represents the change in total cost as the quantity of output increases from one level to another.

It is helpful to express these definitions mathematically:

$$\text{Average total cost} = \text{Total cost}/\text{Quantity}$$
$$ATC = TC/Q$$

and

$$\text{Marginal cost} = \text{Change in total cost}/\text{Change in quantity}$$
$$MC = \Delta TC/\Delta Q.$$

Here, Δ, the Greek letter delta, represents the change in a variable. These equations show how average total cost and marginal cost are derived from total cost. **Average total cost tells us the cost of a typical unit of output if total cost is divided evenly over all the units produced. Marginal cost tells us the increase in total cost that arises from producing an additional unit of output.** The next chapter explains why business managers like Caleb need to keep in mind the concepts of average total cost and marginal cost when deciding how much of their product to supply to the market.

14-3c Cost Curves and Their Shapes

Just as graphs of supply and demand were useful when analyzing the behavior of markets, graphs of average and marginal cost will help when analyzing the behavior of firms. Figure 4 graphs Caleb's costs using the data from Table 2. The horizontal axis measures the quantity the firm produces, and the vertical axis measures marginal and average costs. The graph shows four curves: average total cost (*ATC*), average fixed cost (*AFC*), average variable cost (*AVC*), and marginal cost (*MC*).

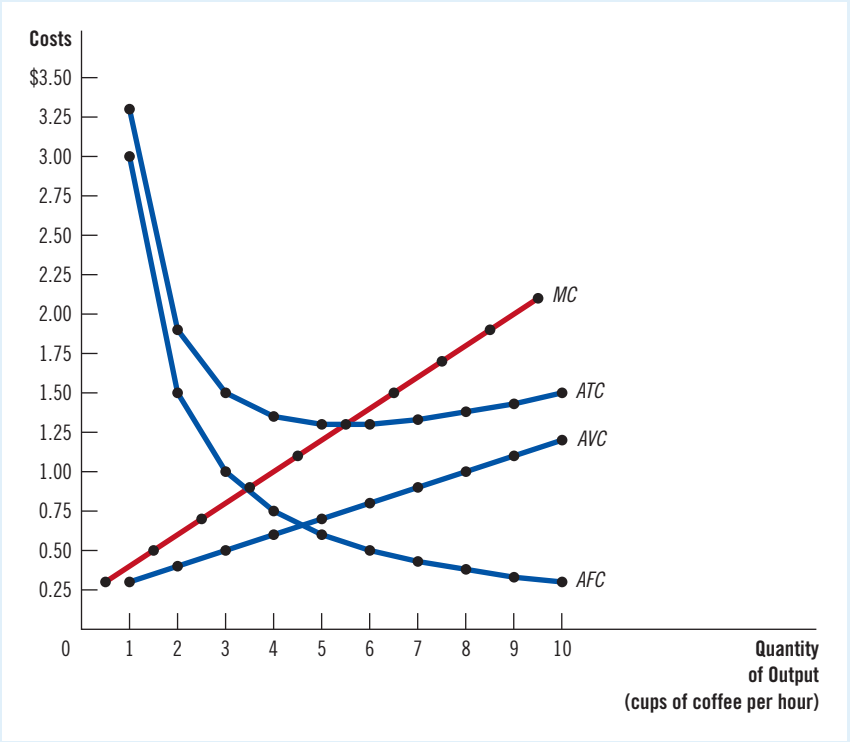
The cost curves shown here for Caleb's Coffee Shop have some features that are common to the cost curves of many firms in the economy. Note three features in particular: the shape of the marginal-cost curve, the shape of the average-total-cost curve, and the relationship between marginal cost and average total cost.

Rising Marginal Cost Caleb's marginal cost rises as the quantity of output produced increases. This upward slope reflects diminishing marginal product. When Caleb produces a small quantity of coffee, he has few workers, and much of his

Figure 4

Caleb's Average-Cost and Marginal-Cost Curves

This figure shows the average total cost (*ATC*), average fixed cost (*AFC*), average variable cost (*AVC*), and marginal cost (*MC*) for Caleb's Coffee Shop. These curves are all obtained by graphing the data in Table 2. They show three common features: (1) Marginal cost rises with the quantity of output. (2) The average-total-cost curve is U-shaped. (3) The marginal-cost curve crosses the average-total-cost curve at the minimum of average total cost.



equipment isn't being used. Because he can easily put these idle resources to use, the marginal product of an extra worker is large, and the marginal cost of producing an extra cup of coffee is small. But when Caleb produces a large quantity of coffee, his shop has a lot of workers, and his equipment is busy. If Caleb produces more coffee by adding new workers, they will operate in crowded conditions and may need to wait to use the equipment. Therefore, when the quantity of coffee produced is already high, the marginal product of an extra worker is low, and the marginal cost of producing an extra cup of coffee is large.

U-Shaped Average Total Cost Caleb's average-total-cost curve in Figure 4 is U-shaped: It first falls and then rises. To understand why it takes this shape, remember that average total cost is the sum of average fixed cost and average variable cost. Average fixed cost declines as output rises because the fixed cost is getting spread over a larger number of units. But average variable cost usually rises as output increases because of diminishing marginal product.

The U-shape occurs because average total cost reflects the shapes of both average fixed cost and average variable cost. At very low levels of output, such as 1 or 2 cups per hour, average total cost is very high. Even though average variable cost is low, average fixed cost is high because the fixed cost is spread over only a few units. As output increases, the fixed cost is spread over more units. Average fixed cost declines rapidly at first and then more slowly. As a result, average total cost also declines until the firm's output reaches 5 cups of coffee per hour, when average total cost is $1.30 per cup. When the firm produces more than 6 cups per

hour, however, the increase in average variable cost becomes the dominant force, and average total cost starts rising. The tug of war between average fixed cost and average variable cost generates the U-shape in average total cost.

efficient scale
the quantity of output that minimizes average total cost

The bottom of the U-shape occurs at the quantity that minimizes average total cost. This quantity is sometimes called the **efficient scale** of the firm. For Caleb, the efficient scale is 5 or 6 cups of coffee per hour. If he produces more or less than this amount, his average total cost rises above the minimum of $1.30. At lower levels of output, average total cost is higher than $1.30 because the fixed cost is spread over so few units. At higher levels of output, average total cost is higher than $1.30 because the marginal product of inputs has diminished significantly. At the efficient scale, these two forces are balanced to yield the lowest average total cost.

The Relationship between Marginal Cost and Average Total Cost If you look at Figure 4 (or back at Table 2), you will see something that may be surprising at first. **Whenever marginal cost is less than average total cost, average total cost is falling. Whenever marginal cost is greater than average total cost, average total cost is rising.** This feature of Caleb's cost curves is not a coincidence: It is true for all firms.

To see why, consider an analogy. Average total cost is like your cumulative grade point average. Marginal cost is like the grade you get in the next course you take. If your grade in your next course is less than your grade point average, your grade point average will fall. If your grade in your next course is higher than your grade point average, your grade point average will rise. The mathematics of average and marginal costs is exactly the same as the mathematics of average and marginal grades.

This relationship between average total cost and marginal cost has an important corollary: **The marginal-cost curve crosses the average-total-cost curve at its minimum.** Why? At low levels of output, marginal cost is below average total cost, so average total cost is falling. But after the two curves cross, marginal cost rises above average total cost. As a result, average total cost must start to rise at this level of output. That's why this point of intersection is the minimum of average total cost. As the next chapter shows, minimum average total cost plays a central role in the analysis of competitive firms.

14-3d Typical Cost Curves

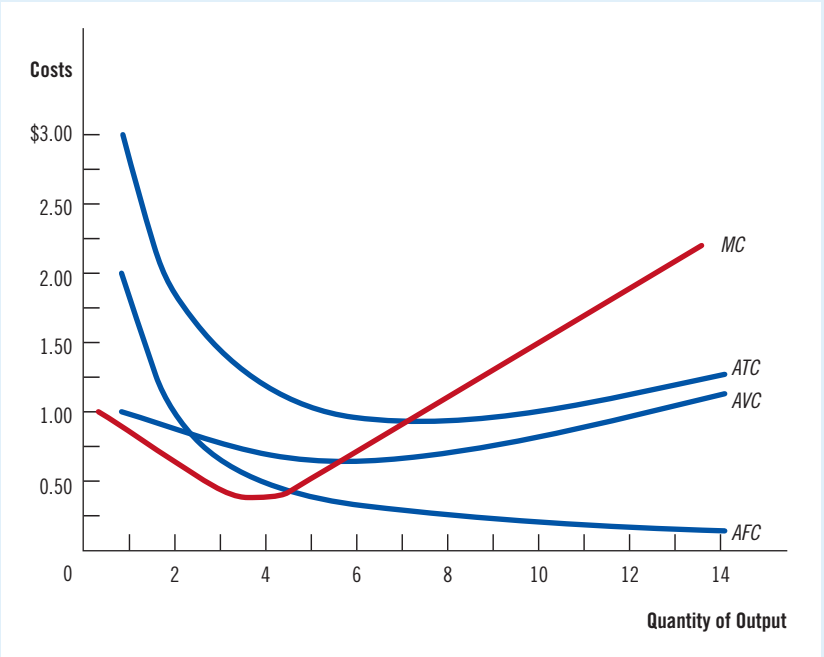
In the examples so far, the firms have exhibited diminishing marginal product and rising marginal cost at all levels of output. This simplifying assumption was useful because it allowed us to focus on the features of cost curves that are most important in analyzing firm behavior. Yet actual firms are often more complex. In many businesses, marginal product does not start to fall immediately after the first worker is hired. Depending on the production process, the second or third worker might have a higher marginal product than the first because a team of workers can divide tasks and work more productively than one person. A firm with this pattern would have increasing marginal product for a while before diminishing marginal product set in.

Figure 5 shows the cost curves for such a firm, including average total cost (*ATC*), average fixed cost (*AFC*), average variable cost (*AVC*), and marginal cost (*MC*). At low levels of output, the firm experiences increasing marginal product, and the marginal-cost curve falls. Eventually, the firm starts to experience diminishing marginal product, and the marginal-cost curve starts to rise. This combination of increasing then diminishing marginal product also makes the average-variable-cost curve U-shaped.

Figure 5

Cost Curves for a Typical Firm

Many firms experience increasing marginal product before diminishing marginal product. As a result, they have cost curves shaped like those in this figure. Notice that marginal cost and average variable cost fall for a while before starting to rise.



Despite these differences, the cost curves in Figure 5 and those in the previous example share three important properties:

- Marginal cost eventually rises with the quantity of output.
- The average-total-cost curve is U-shaped.
- The marginal-cost curve crosses the average-total-cost curve at the minimum of average total cost.

QuickQuiz

5. A firm is producing 1,000 units at a total cost of $5,000. When it increases production to 1,001 units, its total cost rises to $5,008. For this firm,
   a. marginal cost is $5, and average variable cost is $8.
   b. marginal cost is $8, and average variable cost is $5.
   c. marginal cost is $5, and average total cost is $8.
   d. marginal cost is $8, and average total cost is $5.
6. A firm is producing 20 units with an average total cost of $25 and a marginal cost of $15. If it increases production to 21 units, which of the following must occur?
   a. Marginal cost will decrease.
   b. Marginal cost will increase.
   c. Average total cost will decrease.
   d. Average total cost will increase.
7. The government imposes a $1,000 per year license fee on all pizza restaurants. As a result, which cost curves shift?
   a. average total cost and marginal cost
   b. average total cost and average fixed cost
   c. average variable cost and marginal cost
   d. average variable cost and average fixed cost

Answers are at the end of the chapter.

14-4 Costs in the Short Run and in the Long Run

Earlier in this chapter, we noted that a firm's costs might depend on the time horizon under consideration. Let's examine why this is the case.

14-4a The Relationship between Short-Run and Long-Run Average Total Cost

For many firms, the division of total costs between fixed and variable costs depends on the time horizon. Consider, for instance, a car manufacturer such as Ford Motor Company. Over only a few months, Ford cannot adjust the number or sizes of its factories. The only way it can produce additional cars is to hire more workers at the factories it already has. The cost of these factories is, therefore, a fixed cost in the short run. But over several years, Ford can expand its factories' sizes, build new ones, or close old ones. In the long run, its factories represent a variable cost.

Because many decisions are fixed in the short run but variable in the long run, a firm's long-run and short-run cost curves differ. Figure 6 shows an example. The figure presents three short-run average-total-cost curves—for a small, medium, and large factory. It also presents the long-run average-total-cost curve. As the firm moves along the long-run curve, it adjusts the size of the factory to the quantity of production.

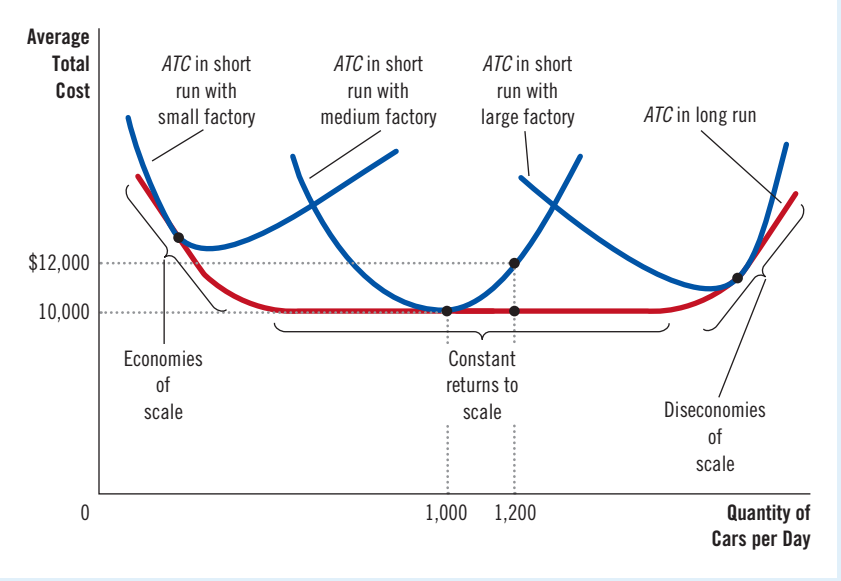
This graph shows how short-run and long-run costs are related. The long-run average-total-cost curve has a much flatter U-shape than the short-run average-total-cost curve. In addition, all the short-run curves lie on or above the long-run curve. This is because firms have greater flexibility in the long run. In essence, in the long run, the firm gets to choose which short-run curve it wants. But in the short run, it has to use whatever short-run curve it already has, determined by decisions made in the past.

The figure shows how a change in production alters costs over different time horizons. When Ford wants to increase production from 1,000 to 1,200 cars per day, it has no choice in the short run but to hire more workers at its existing medium-sized

Figure 6

Average Total Cost in the Short and Long Runs

Because fixed costs are variable in the long run, the average-total-cost curve in the short run differs from the average-total-cost curve in the long run.



factory. Because of diminishing marginal product, average total cost rises from $10,000 to $12,000 per car. In the long run, however, Ford can expand both the size of the factory and its workforce, and average total cost returns to $10,000.

How long does it take a firm to get to the long run? It depends on the firm. A major manufacturer, such as a car company, may need a year or more to build a larger factory. But a person running a coffee shop can buy another coffee maker within a few days. There is no single answer to how long it takes a firm to adjust its production facilities.

14-4b Economies and Diseconomies of Scale

The shape of the long-run average-total-cost curve conveys important information about a firm's production processes. In particular, it tells us how costs vary with the scale—that is, the size—of a firm's operations. When long-run average total cost declines as output increases, there are said to be **economies of scale**. When the opposite occurs and long-run average total cost rises as output increases, there are **diseconomies of scale**. And when long-run average total cost does not vary with the level of output, there are **constant returns to scale**. In Figure 6, Ford has economies of scale at low levels of output, constant returns to scale at intermediate levels of output, and diseconomies of scale at high levels of output.

economies of scale
the property whereby long-run average total cost falls as the quantity of output increases

diseconomies of scale
the property whereby long-run average total cost rises as the quantity of output increases

constant returns to scale
the property whereby long-run average total cost stays the same as the quantity of output changes

What might cause economies or diseconomies of scale? Economies of scale often arise because higher production levels allow **specialization** among workers, which enables them to become better at specific tasks. For instance, if Ford hires a large number of workers and produces many cars, it can reduce costs using modern assembly-line production. Diseconomies of scale can arise because of **coordination problems** that often occur in large organizations. The more cars Ford produces, the more stretched the management team becomes, and the less effective the managers become at keeping costs down.

This analysis shows why long-run average-total-cost curves are often U-shaped. At low levels of production, the firm benefits from increased size because it can take

Lessons from a Pin Factory

"Jack of all trades, master of none." This adage means that a person who tries to do everything ends up doing nothing very well, and it helps explain the shapes of cost curves. If a firm wants its workers to be as productive as possible, it should expect each worker to master a limited number of tasks. But this organization of work is possible only if a firm employs many workers and produces a large quantity of output.

In his book *The Wealth of Nations*, Adam Smith described a visit he made to a pin factory. Smith was impressed by the specialization among the workers and the resulting economies of scale. He wrote,

> One man draws out the wire, another straightens it, a third cuts it, a fourth points it, a fifth grinds it at the top for receiving the head; to make the head requires two or three distinct operations; to put it on is a peculiar business; to whiten it is another; it is even a trade by itself to put them into paper.

Smith reported that because of this specialization, the factory produced thousands of pins per worker every day. He conjectured that if the workers had chosen to work separately, rather than as a team of specialists, "they certainly could not each of them make twenty, perhaps not one pin a day." Because of specialization, a large pin factory could achieve higher output per worker and lower average cost per pin than a small factory.

The specialization that Smith observed in the pin factory is common in the modern economy. If you want to build a house, you could do all the work yourself. But most people these days turn to a builder, who in turn hires carpenters, plumbers, electricians, painters, and people in many other trades. These workers become better at particular jobs than if they were generalists. The use of specialization to achieve economies of scale is one reason modern societies are as prosperous as they are. ■

advantage of greater specialization, and coordination problems are not yet acute. At high levels of production, the benefits of specialization have already been realized, and coordination problems become more severe as the firm grows larger. As a result, long-run average total cost is falling at low levels of production because of increasing specialization and rising at high levels of production because of growing coordination problems.

QuickQuiz

8. If a higher level of production allows workers to specialize in particular tasks, a firm will likely exhibit _____ of scale and _____ average total cost.
   a. economies; falling
   b. economies; rising
   c. diseconomies; falling
   d. diseconomies; rising

9. If Boeing produces 9 jets per month, its long-run total cost is $9 million per month. If it produces 10 jets per month, its long-run total cost is $11 million per month. Boeing exhibits
   a. rising marginal cost.
   b. falling marginal cost.
   c. economies of scale.
   d. diseconomies of scale.

Answers are at the end of the chapter.

14-5 Conclusion

This chapter has developed tools to study how firms make production and pricing decisions. You should now understand what economists mean by the term **costs** and how costs vary with the quantity of output a firm produces. Table 3 summarizes some of the definitions we have encountered.

By themselves, a firm's cost curves do not tell us what decisions the firm will make. But they are a key component of that decision, as the next chapter shows.

Table 3

The Many Types of Cost: A Summary

Term	Definition	Mathematical Description
Explicit costs	Costs that require an outlay of money by the firm	
Implicit costs	Costs that do not require an outlay of money by the firm	
Fixed costs	Costs that do not vary with the quantity of output produced	FC
Variable costs	Costs that vary with the quantity of output produced	VC
Total cost	The market value of all the inputs that a firm uses in production	$TC = FC + VC$
Average fixed cost	Fixed cost divided by the quantity of output	$AFC = FC/Q$
Average variable cost	Variable cost divided by the quantity of output	$AVC = VC/Q$
Average total cost	Total cost divided by the quantity of output	$ATC = TC/Q$
Marginal cost	The increase in total cost that arises from an extra unit of production	$MC = \Delta TC/\Delta Q$

Chapter in a Nutshell

- A firm's goal is to maximize profit, which equals total revenue minus total cost.
- When analyzing a firm's behavior, it is important to include all the opportunity costs of production. Some opportunity costs, such as the wages a firm pays its workers, are explicit. Others, like the wages the firm owner forgoes by not taking another job, are implicit. While accounting profit considers only explicit costs, economic profit accounts for both explicit and implicit costs.
- A firm's costs reflect its production process. A typical firm's production function gets flatter as the quantity of an input increases, displaying the property of diminishing marginal product. As a result, the total-cost curve gets steeper as the quantity produced rises.
- A firm's total cost can be separated into fixed costs and variable costs. Fixed costs remain constant when the firm alters the quantity of output produced. Variable costs change when the firm alters the quantity of output produced.
- From a firm's total cost, two related measures of cost are derived. Average total cost is total cost divided by the quantity of output. Marginal cost is the amount by which total cost rises if output increases by 1 unit.
- When analyzing firm behavior, it is often useful to graph average total cost and marginal cost. For a typical firm, marginal cost rises with the quantity of output. Average total cost first falls as output increases and then rises as output increases further. The marginal-cost curve always crosses the average-total-cost curve at the minimum of average total cost.
- A firm's costs often depend on the time horizon considered. In particular, many costs are fixed in the short run but variable in the long run. As a result, when the firm changes its level of production, average total cost may rise more in the short run than in the long run.

Key Concepts

total revenue, p. 268
total cost, p. 268
profit, p. 268
explicit costs, p. 269
implicit costs, p. 269
economic profit, p. 269
accounting profit, p. 270
production function, p. 271
marginal product, p. 272
diminishing marginal product, p. 273
fixed costs, p. 275
variable costs, p. 275
average total cost, p. 276
average fixed cost, p. 276
average variable cost, p. 276
marginal cost, p. 276
efficient scale, p. 278
economies of scale, p. 281
diseconomies of scale, p. 281
constant returns to scale, p. 281

Questions for Review

1. What is the relationship between a firm's total revenue, total cost, and profit?
2. Give an example of an opportunity cost that an accountant would not count as a cost. Why would the accountant ignore this cost?
3. What is marginal product, and what is meant by diminishing marginal product?
4. Draw a production function that exhibits diminishing marginal product of labor. Draw the associated total-cost curve. (In both cases, be sure to label the axes.) Explain the shapes of the two curves you have drawn.
5. Define **total cost**, **average total cost**, and **marginal cost**. How are they related?
6. Draw the marginal-cost and average-total-cost curves for a typical firm. Explain why the curves have the shapes they do and why they intersect where they do.
7. How and why does a firm's average-total-cost curve in the short run differ from its average-total-cost curve in the long run?
8. Define **economies of scale** and explain why they might arise. Define **diseconomies of scale** and explain why they might arise.

Problems and Applications

1. This chapter discusses many types of costs: opportunity cost, total cost, fixed cost, variable cost, average total cost, and marginal cost. Fill in the type of cost that best completes each sentence:
   a. What you give up in taking some action is called the _____.
   b. _____ is falling when marginal cost is below it and rising when marginal cost is above it.
   c. A cost that does not depend on the quantity produced is a(n) _____.
   d. In the ice-cream industry in the short run, _____ includes the cost of cream and sugar but not the cost of the factory.
   e. Profits equal total revenue minus _____.
   f. The cost of producing an extra unit of output is the _____.

2. Buffy is thinking about opening an amulet store. She estimates that it would cost $350,000 per year to rent the location and buy the merchandise. In addition, she would have to quit her $80,000 per year job as a vampire hunter.
   a. Define **opportunity cost**.
   b. What is Buffy's opportunity cost of running the store for a year?
   c. Buffy thinks she can sell $400,000 worth of amulets in a year. What would her accountant consider the store's profit?
   d. Should Buffy open the store? Explain.
   e. How much revenue would the store need to generate for Buffy to earn positive economic profit?

3. A commercial fisherman notices the following relationship between hours spent fishing and the quantity of fish caught:

Hours	Quantity of Fish (in pounds)
0 hours	0 lb.
1	10
2	18
3	24
4	28
5	30

   a. What is the marginal product of each hour spent fishing?
   b. Use these data to graph the fisherman's production function. Explain its shape.
   c. The fisherman has a fixed cost of $10 (his pole). The opportunity cost of his time is $5 per hour. Graph the fisherman's total-cost curve. Explain its shape.

4. Nimbus, Inc. makes brooms and sells them door-to-door. Here is the relationship between the number of workers and Nimbus's output during a given day:

Workers	Output	Marginal Product	Total Cost	Average Total Cost	Marginal Cost
0	0		_____	_____	
		_____			_____
1	20		_____	_____	
		_____			_____
2	50		_____	_____	
		_____			_____
3	90		_____	_____	
		_____			_____
4	120		_____	_____	
		_____			_____
5	140		_____	_____	
		_____			_____
6	150		_____	_____	
		_____			_____
7	155		_____	_____	

   a. Fill in the column of marginal products. What pattern do you see? How might you explain it?
   b. A worker costs $100 a day, and the firm has fixed costs of $200. Use this information to fill in the column for total cost.
   c. Fill in the column for average total cost. (Recall that $ATC = TC/Q$.) What pattern do you see?
   d. Now fill in the column for marginal cost. (Recall that $MC = \Delta TC/\Delta Q$.) What pattern do you see?
   e. Compare the column for marginal product with the column for marginal cost. Explain the relationship.
   f. Compare the column for average total cost with the column for marginal cost. Explain the relationship.

5. You are the chief financial officer for a firm that sells gaming consoles. Your firm has the following average-total-cost schedule:

Quantity	Average Total Cost
600 consoles	$300
601	301

Your current level of production is 600 consoles, all of which have been sold. Someone calls, desperate to buy one of your consoles. The caller offers you $550 for it. Should you accept the offer? Why or why not?

6. Consider the following cost information for a pizzeria:

Quantity	Total Cost	Variable Cost
0 dozen pizzas	$300	$0
1	350	50
2	390	90
3	420	120
4	450	150
5	490	190
6	540	240

   a. What is the pizzeria's fixed cost?
   b. Construct a table in which you calculate the marginal cost per dozen pizzas using the information on total cost. Also, calculate the marginal cost per dozen pizzas using the information on variable cost. What is the relationship between these sets of numbers? Explain.

7. Your cousin Vinnie owns a painting company with fixed costs of $200 and the following schedule for variable costs:

Quantity of Houses Painted per Month	1	2	3	4	5	6	7
Variable Costs	$10	$20	$40	$80	$160	$320	$640

   Calculate average fixed cost, average variable cost, and average total cost for each quantity. What is the efficient scale of the painting company?

8. The city government is considering two tax proposals:
   - A lump-sum tax of $300 on each producer of hamburgers
   - A tax of $1 per burger, paid by producers of hamburgers

   a. Which of the following curves—average fixed cost, average variable cost, average total cost, and marginal cost—would shift as a result of the lump-sum tax? Why? Show this in a graph. Label the graph as precisely as possible.
   b. Which of these same four curves would shift as a result of the per-burger tax? Why? Show this in a new graph. Label the graph as precisely as possible.

9. Jane's Juice Bar has the following cost schedules:

Quantity	Variable Cost	Total Cost
0 vats of juice	$0	$30
1	10	40
2	25	55
3	45	75
4	70	100
5	100	130
6	135	165

   a. Calculate average variable cost, average total cost, and marginal cost for each quantity.
   b. Graph all three curves. What is the relationship between the marginal-cost curve and the average-total-cost curve? Between the marginal-cost curve and the average-variable-cost curve? Explain.

10. Consider the following table of long-run total costs for three different firms:

Quantity	1	2	3	4	5	6	7
Firm A	$60	$70	$80	$90	$100	$110	$120
Firm B	11	24	39	56	75	96	119
Firm C	21	34	49	66	85	106	129

    Does each of these firms experience economies of scale or diseconomies of scale?

QuickQuiz Answers

1. **c** 2. **a** 3. **a** 4. **d** 5. **d** 6. **c** 7. **b** 8. **a** 9. **d**

Chapter

15

Firms in Competitive Markets

If a local gas station raised its price for gasoline by 20 percent and others didn't, its customers would quickly start buying gasoline elsewhere. By contrast, if a local water company raised the price of water by 20 percent, it wouldn't lose much business. People might buy more water-efficient showerheads and, in suburban areas, water their lawns less often, but they would be hard-pressed to find another source of water. The difference between the gasoline market and the water market is that most cities and towns have several stations supplying gasoline but only one company supplying tap water. This difference in market structure shapes the pricing and production decisions of the firms that operate in these markets.

This chapter examines the behavior of competitive firms, such as a local gas station. Recall that a market is competitive if each buyer and seller is small compared with the size of the market and, therefore, has little ability to influence market prices. By contrast, a firm that can influence the market price of the good it sells, such as a local water company, is said to have **market power**. The next chapter takes up that topic.

AZMANJAKA/E+/GETTY IMAGES, CARLOS FLORES/EYEEM/GETTY IMAGES

Competitive firms are the natural place to begin the study of firm behavior for two reasons. First, because competitive firms have negligible influence on market prices, they are simpler to understand than those with market power. Second, because competitive markets allocate resources efficiently (as Chapter 7 showed), they provide a benchmark for comparison with other market structures.

The analysis of competitive firms in this chapter helps explain the decisions behind market supply curves. Not surprisingly, the supply curve in a market is closely linked to firms' costs of production. Less obvious is which among the different types of cost—fixed, variable, average, and marginal—are most relevant for its supply decisions. As we will see, all of these cost measures play important and interrelated roles in supply decisions.

15-1 What Is a Competitive Market?

Our goal in this chapter is to examine how firms make production decisions in competitive markets. Let's begin by reviewing what a competitive market is.

15-1a The Meaning of Competition

competitive market
a market with many buyers and sellers trading identical products so that each buyer and seller is a price taker

A **competitive market**, sometimes called a **perfectly competitive market**, has two characteristics:

- The market has many buyers and many sellers.
- The goods offered by the various sellers are largely the same.

Under these conditions, the actions of any single buyer or seller have a negligible impact on the market price. Each buyer and seller takes the market price as given.

Consider the market for milk. No single consumer can influence the price of milk because each buys a small amount relative to the size of the market. Similarly, each dairy farmer has limited control over the price because many other farmers offer milk that is essentially identical. Because sellers can sell all they want at the going price, none of them has any reason to charge less, and if one of them charges more, buyers will go elsewhere. Buyers and sellers in competitive markets must accept the price the market determines and are said to be **price takers**.

In addition to the previous two conditions for competition, a third condition is sometimes said to characterize perfectly competitive markets:

- Firms can freely enter or exit the market.

If, for instance, anyone can start a new dairy farm and any existing dairy farmer can leave the business, then the dairy industry satisfies this condition. Much of the analysis of competitive firms does not require the assumption of free entry and exit because firms can be price takers without it. Yet, as we will see, free entry and exit is a powerful force that shapes the long-run equilibrium in competitive markets.

15-1b The Revenue of a Competitive Firm

The standard model of competitive markets assumes that firms aim to maximize profit (total revenue minus total cost). To see how they do this, let's start by considering the revenue of a typical competitive firm: the Vaca Family Dairy Farm.

The Vaca Farm produces a quantity of milk, Q, and sells each unit at the market price, P. The farm's total revenue is $P \times Q$. For example, if a gallon of milk sells for \$6 and the farm sells 1,000 gallons, its total revenue is \$6,000.

Because the Vaca Farm is small compared with the world market for milk, it takes the price given by the market. This means that the price does not depend on how much milk the Vaca Farm produces and sells. If the Vacas double what they produce to 2,000 gallons, the price remains the same, and their total revenue doubles to \$12,000. Total revenue is proportional to output.

Table 1 shows the revenue for the Vaca Family Dairy Farm. Columns (1) and (2) show the amount of output the farm produces and the price at which it sells its output. Column (3) is the farm's total revenue. The table assumes that the price of milk is \$6 a gallon, so total revenue is \$6 times the number of gallons.

Just as the concepts of average and marginal were useful in the preceding chapter when analyzing costs, they are also useful when discussing revenue. To see what these concepts reveal, consider two questions:

- How much revenue does the farm receive for the typical gallon of milk?
- How much additional revenue does the farm receive if it increases the production of milk by 1 gallon?

Columns (4) and (5) in Table 1 answer these questions.

Table 1

Total, Average, and Marginal Revenue for a Competitive Firm

(1) Quantity (Q)	(2) Price (P)	(3) Total Revenue ($TR = P \times Q$)	(4) Average Revenue ($AR = TR / Q$)	(5) Marginal Revenue ($MR = \Delta TR / \Delta Q$)
1 gallon	\$6	\$6	\$6	
				\$6
2	6	12	6	
				6
3	6	18	6	
				6
4	6	24	6	
				6
5	6	30	6	
				6
6	6	36	6	
				6
7	6	42	6	
				6
8	6	48	6	

average revenue
total revenue divided by the quantity sold

Column (4) in the table shows **average revenue**, which is total revenue [from column (3)] divided by the amount of output [from column (1)]. Average revenue is what a firm receives for the typical unit sold. Table 1 shows that average revenue equals $6, the price of a gallon of milk. This illustrates a general lesson that applies to all firms, competitive or not. Average revenue is total revenue ($P \times Q$) divided by quantity (Q). **Therefore, for all types of firms, average revenue equals the price of the good.**

marginal revenue
the change in total revenue from an additional unit sold

Column (5) shows **marginal revenue**, which is the change in total revenue from the sale of each additional unit of output. In Table 1, marginal revenue equals $6, the price of a gallon of milk. This illustrates a lesson that applies only to firms in competitive markets. Because total revenue is $P \times Q$ and P is fixed for a competitive firm, when Q rises by 1 unit, total revenue rises by P dollars. **Therefore, for competitive firms, marginal revenue equals the price of the good.**

QuickQuiz

1. A perfectly competitive firm
   a. chooses its price to maximize profits.
   b. sets its price to undercut other firms selling similar products.
   c. takes its price as given by market conditions.
   d. picks the price that yields the largest market share.

2. When a perfectly competitive firm increases the quantity it produces and sells by 10 percent, its marginal revenue _____ and its total revenue rises by _____.
   a. falls; less than 10 percent
   b. falls; exactly 10 percent
   c. stays the same; less than 10 percent
   d. stays the same; exactly 10 percent

Answers are at the end of the chapter.

15-2 Profit Maximization and the Competitive Firm's Supply Curve

The goal of a firm is to maximize profit, which equals total revenue minus total cost. The previous section discussed the competitive firm's revenue, and the previous chapter analyzed the firm's costs. Let's now consider how a competitive firm maximizes profit and how that decision determines its supply curve.

15-2a A Simple Example of Profit Maximization

Table 2 presents more information about the Vaca Family Dairy Farm. Column (1) in the table shows the number of gallons of milk the farm produces. Column (2) shows the farm's total revenue, which is $6 times the number of gallons. Column (3) shows the farm's total cost. Total cost includes fixed costs, which are $3 in this example, and variable costs, which depend on the quantity produced.

Column (4) shows the farm's profit, which is computed by subtracting total cost from total revenue. If the farm produces nothing, it has a loss of $3 (its fixed cost). If it produces 1 gallon, it has a profit of $1. If it produces 2 gallons, it has a profit of $4, and so on. The Vaca family wants to produce the quantity of milk that makes its profit as large as possible. In this example, the farm maximizes profit by producing either 4 or 5 gallons of milk for a profit of $7.

There is another way to look at Vaca Farm's decision: The Vacas can find the profit-maximizing quantity by comparing the marginal revenue and marginal cost of each unit produced. Columns (5) and (6) in Table 2 compute marginal revenue

Table 2

Profit Maximization: A Numerical Example

(1) Quantity (Q)	(2) Total Revenue (TR)	(3) Total Cost (TC)	(4) Profit ($TR - TC$)	(5) Marginal Revenue ($MR = \Delta TR / \Delta Q$)	(6) Marginal Cost ($MC = \Delta TC / \Delta Q$)	(7) Change in Profit ($MR - MC$)
0 gallons	\$0	\$3	−\$3			
				\$6	\$2	\$4
1	6	5	1			
				6	3	3
2	12	8	4			
				6	4	2
3	18	12	6			
				6	5	1
4	24	17	7			
				6	6	0
5	30	23	7			
				6	7	−1
6	36	30	6			
				6	8	−2
7	42	38	4			
				6	9	−3
8	48	47	1			

and marginal cost from the changes in total revenue and total cost, and column (7) shows the change in profit for each additional gallon produced. The first gallon of milk the farm produces has a marginal revenue of \$6 and a marginal cost of \$2, so producing that gallon increases profit by \$4 (from −\$3 to \$1). The second gallon produced has a marginal revenue of \$6 and a marginal cost of \$3, so that gallon increases profit by \$3 (from \$1 to \$4). As long as marginal revenue exceeds marginal cost, increasing the quantity produced raises profit. Once the farm's output reaches 5 gallons of milk, however, the situation changes. The sixth gallon would have a marginal revenue of \$6 and a marginal cost of \$7, so producing it would reduce profit by \$1 (from \$7 to \$6). As a result, the Vacas do not produce more than 5 gallons.

One of the **Ten Principles of Economics** in Chapter 1 is that rational people think at the margin. The Vacas can apply this principle. If marginal revenue is greater than marginal cost—as it is at 1, 2, and 3 gallons—they should increase the production of milk because it will put more money in their pockets (marginal revenue) than it takes out (marginal cost). If marginal revenue is less than marginal cost—as it is at 6, 7, and 8 gallons—the Vacas should decrease production. By thinking at the margin and making incremental adjustments to the level of production, the Vacas end up producing the profit-maximizing quantity.

15-2b The Marginal-Cost Curve and the Firm's Supply Decision

To extend this analysis, consider the cost curves in Figure 1. These cost curves exhibit the three features that, as the preceding chapter explained, are thought to describe most firms: The marginal-cost curve (*MC*) slopes upward, the average-total-cost curve (*ATC*) is U-shaped, and the marginal-cost curve crosses the average-total-cost curve at the minimum of average total cost. The figure also shows a horizontal line at the market price (*P*). The price line is horizontal because a competitive firm is a price taker: The price of the firm's output is the same regardless of how much it produces. Keep in mind that, for a competitive firm, the price equals both the firm's average revenue (*AR*) and its marginal revenue (*MR*).

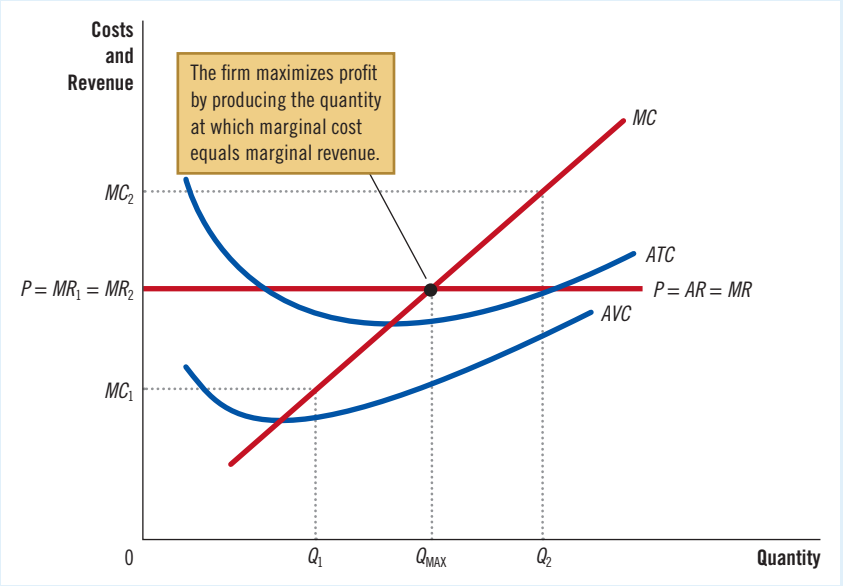
We can use Figure 1 to find the quantity of output that maximizes profit. Imagine that the firm is producing at Q_1. At this level of output, the marginal-revenue curve is above the marginal-cost curve, indicating that marginal revenue is greater than marginal cost. This means that if the firm were to raise production by 1 unit, the additional revenue (MR_1) would exceed the additional cost (MC_1). Profit, which equals total revenue minus total cost, would increase. Hence, if marginal revenue is greater than marginal cost, as it is at Q_1, the firm can increase profit by increasing production.

A similar argument applies when output is at Q_2. In this case, the marginal-cost curve is above the marginal-revenue curve, showing that marginal cost is greater than marginal revenue. If the firm were to reduce production by 1 unit, the costs saved (MC_2) would exceed the revenue lost (MR_2). Therefore, if marginal cost is greater than marginal revenue, as it is at Q_2, the firm can increase profit by reducing production.

Figure 1

Profit Maximization for a Competitive Firm

This figure shows the marginal-cost curve (*MC*), the average-total-cost curve (*ATC*), and the average-variable-cost curve (*AVC*). It also shows the market price (*P*), which for a competitive firm equals both marginal revenue (*MR*) and average revenue (*AR*). At the quantity Q_1, marginal revenue MR_1 exceeds marginal cost MC_1, so raising production increases profit. At the quantity Q_2, marginal cost MC_2 is above marginal revenue MR_2, so reducing production increases profit. The profit-maximizing quantity, Q_{MAX}, is found where the horizontal line representing the price intersects the marginal-cost curve.



Where do these marginal adjustments to production end? Regardless of whether the firm begins with production at a low level (such as Q_1) or at a high level (such as Q_2), it will eventually adjust production until the quantity produced reaches the profit-maximizing quantity, Q_{MAX}. This analysis yields three rules for profit maximization:

- If marginal revenue exceeds marginal cost, the firm should increase its output.
- If marginal cost exceeds marginal revenue, the firm should decrease its output.
- At the profit-maximizing level of output, marginal revenue equals marginal cost.

These rules are the key to rational decision making by any profit-maximizing firm. They apply not only to competitive firms but, as the next chapter shows, to other types of firms as well.

We can now see how the competitive firm decides what quantity of its good to supply to the market. Because a competitive firm is a price taker, its marginal revenue equals the market price. For any price, the competitive firm's profit-maximizing quantity of output is found by looking at the intersection of the price with the marginal-cost curve. In Figure 1, that quantity is Q_{MAX}.

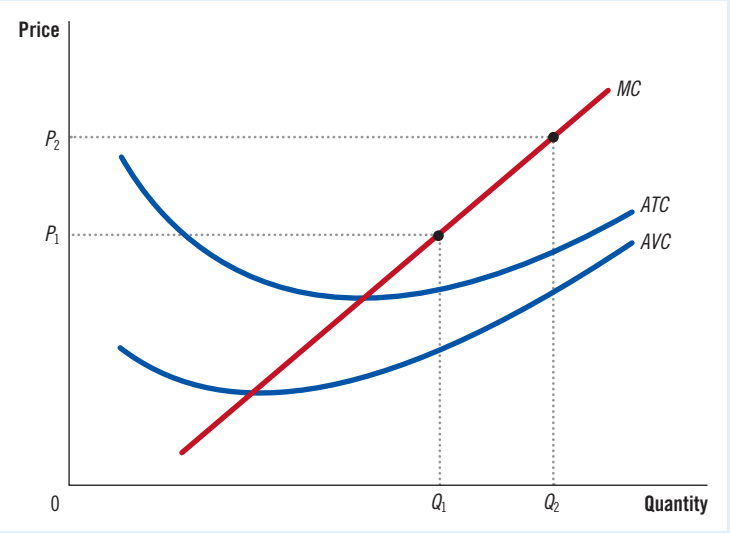
Suppose that the price prevailing in this market rises, perhaps because of an increase in market demand. Figure 2 shows how a competitive firm responds to the price increase. When the price is P_1, the firm produces quantity Q_1, the quantity that equates marginal cost to the price. When the price rises to P_2, the firm finds that marginal revenue is higher than marginal cost at the previous level of output, so it increases production. The new profit-maximizing quantity is Q_2, at which marginal cost equals the new, higher price. **Because the firm's marginal-cost curve determines the quantity of the good the firm is willing to supply at any price, the marginal-cost curve is also the competitive firm's supply curve.**

There are, however, some caveats to this conclusion, which we examine next.

Figure 2

Marginal Cost as the Competitive Firm's Supply Curve

An increase in the price from P_1 to P_2 leads to an increase in the firm's profit-maximizing quantity from Q_1 to Q_2. Because the marginal-cost curve shows the quantity supplied at any price, it is the firm's supply curve.



15-2c The Firm's Short-Run Decision to Shut Down

So far, we have been analyzing the question of how much a competitive firm will produce. In some circumstances, however, the firm will decide to shut down and not produce anything at all.

It's important to distinguish between a firm's temporary shutdown and its permanent exit from the market. A **shutdown** refers to a short-run decision not to produce anything during a specific period because of current market conditions. **Exit** refers to a long-run decision to leave the market. The short-run and long-run decisions differ because most firms cannot avoid their fixed costs in the short run but can do so in the long run. That is, a firm that shuts down temporarily still must pay its fixed costs, while a firm that exits the market doesn't pay any costs at all, fixed or variable.

For example, consider production decisions at a farm. The cost of the land is a fixed cost. If the farm owners decide not to produce any crops one season, the land lies fallow, and they cannot recover this cost. When making the short-run decision of whether to shut down for a season, the fixed cost of land is said to be a **sunk cost**. By contrast, if the farm owners decide to leave farming altogether, they can sell the land. When making the long-run decision of whether to exit the market, the cost of land is not sunk. (Sunk costs will come up again shortly.)

What determines a firm's shutdown decision? If the firm shuts down, it loses all revenue from the sale of its product. At the same time, it saves the variable costs of making the product (but must still pay the fixed costs). **Therefore, the firm shuts down if the revenue that it would earn from producing is less than its variable costs of production.**

A bit of mathematics can make this shutdown rule more useful. If TR stands for total revenue and VC stands for variable cost, then the firm's decision can be written as

$$\text{Shut down if } TR < VC.$$

The firm shuts down if total revenue is less than variable cost. By dividing both sides of this inequality by the quantity Q, we can write it as

$$\text{Shut down if } TR/Q < VC/Q.$$

The left side of the inequality, TR/Q, is total revenue $P \times Q$ divided by quantity Q, which is average revenue, most simply expressed as the good's price, P. The right side of the inequality, VC/Q, is average variable cost, AVC. The firm's shutdown rule can be restated as

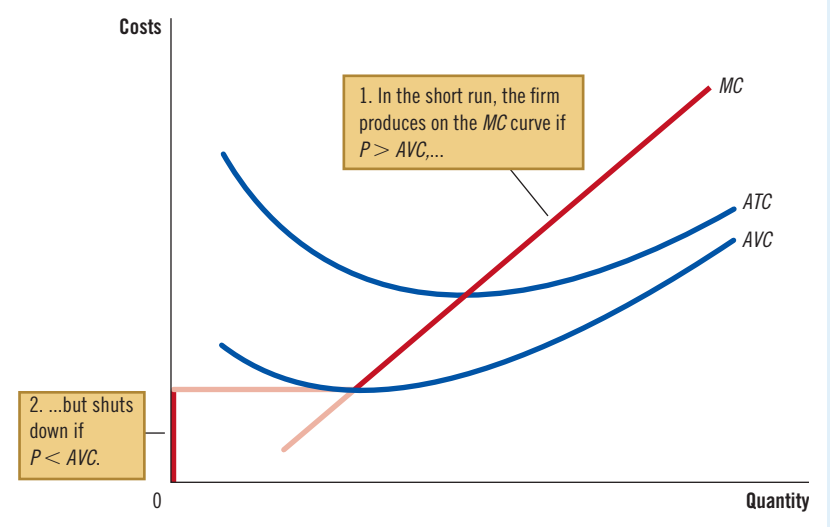
$$\text{Shut down if } P < AVC.$$

That is, a firm shuts down if the price of the good is less than the average variable cost of production. This rule is intuitive: When deciding whether to produce, the firm compares the price it receives for the typical unit to the average variable cost that it must incur to produce it. If the price doesn't cover the average variable cost, the firm is better off stopping production altogether. The firm still loses money (because it incurs fixed costs), but it would lose even more money by staying open. The firm can reopen in the future if conditions change so that price exceeds average variable cost.

Figure 3

The Competitive Firm's Short-Run Supply Curve

In the short run, the competitive firm's supply curve is the portion of its marginal-cost curve (*MC*) that lies above its average-variable-cost curve (*AVC*). If the price falls below average variable cost, the firm is better off shutting down temporarily.



We now have a full description of a competitive firm's profit-maximizing strategy. If the firm produces anything, it produces the quantity at which marginal cost equals the good's price, which the firm takes as given. Yet if the price is less than average variable cost at that quantity, the firm is better off shutting down temporarily and not producing anything. Figure 3 illustrates these results. **The competitive firm's short-run supply curve is the portion of its marginal-cost curve that lies above the average-variable-cost curve.**

15-2d Spilt Milk and Other Sunk Costs

At some point in your life, you may have been told, "Don't cry over spilt milk," or "Let bygones be bygones." These adages were not written by economists, but they could have been. They express an important truth about rational decision making. A **sunk cost** is a cost that has already been committed and cannot be recovered. Because nothing can be done about sunk costs, it's rational to ignore them when making many decisions in life, including those regarding business strategy.

sunk cost
a cost that has already been committed and cannot be recovered

Our analysis of the firm's shutdown decision is an example of the irrelevance of sunk costs. We assume that the firm can't recover its fixed costs by temporarily stopping production. That is, regardless of the quantity of output supplied (even if it is zero), the firm still must pay its fixed costs. In the short run, the fixed costs are sunk, and the firm should ignore them when deciding how much to produce. The firm's short-run supply curve is the part of the marginal-cost curve that lies above average variable cost. Because the fixed costs are sunk, their size doesn't matter for this supply decision.

The irrelevance of sunk costs is worth remembering in your personal life. Imagine, for instance, that you want to see a new movie. You're willing to pay $15 to see it, and a ticket only costs $10, so you decide to go. Before entering the theater, however, you lose the ticket. The theater management doesn't believe that you lost it. What should you do?

You might be so angry and disappointed that you're inclined to go home and forget about the movie. After all, buying another ticket now would bring your total cost to $20, which might seem to be too much. But that would be a mistake. Rationally, you should buy another ticket for $10. Why? The benefit of seeing the movie ($15) exceeds the opportunity cost (the $10 for the second ticket). The $10 you paid for the lost ticket is a sunk cost. As with spilt milk, there is no point in crying about it.

Case Study

Near-Empty Restaurants and Off-Season Miniature Golf

Have you ever walked into a restaurant for lunch and found it almost empty? Why, you might ask, does such a restaurant even bother to stay open? It might seem that the revenue from so few customers could not possibly cover the cost of running the restaurant.

When deciding whether to open for lunch, a restaurant owner must keep in mind the distinction between fixed and variable costs. Many of a restaurant's costs—the rent, kitchen equipment, tables, plates, silverware, and so on—are fixed. Shutting down during lunch would not reduce these costs, which are sunk in the short run. When the owner is deciding whether to serve lunch, only the variable costs—the price of the additional food and the wages of the extra staff—are relevant. The owner shuts down the restaurant at lunchtime only if the revenue from the few lunchtime customers would fail to cover the variable costs.

An operator of a miniature golf course in a summer resort community faces a similar decision. Because revenue varies substantially from season to season, the firm must decide what date to open and what date to close. Once again, the fixed costs—the costs of buying the land and building the course—are irrelevant to this short-run decision. The miniature golf course should open for business only during those times of the year when its revenue exceeds its variable costs. ●

ADRIAN SHERRATT/ALAMY STOCK PHOTO

Staying open can be profitable, even with many tables empty.

15-2e The Firm's Long-Run Decision to Exit or Enter a Market

A firm's long-run decision to leave a market is similar to its shutdown decision. If it exits, it will again lose all revenue from the sale of its product, but now it will save not only its variable costs of production but also its fixed costs. **The firm exits the market if the revenue it would get from producing is less than its total cost of production.**

We can again make this rule more useful by writing it mathematically. If TR stands for total revenue and TC stands for total cost, then the firm's exit rule can be written as

$$\text{Exit if } TR < TC.$$

The firm exits if total revenue is less than total cost. By dividing both sides of this inequality by quantity Q, we can write it as

$$\text{Exit if } TR/Q < TC/Q.$$

This can be simplified by noting that TR/Q is average revenue, which equals the price P, and that TC/Q is average total cost, ATC. The firm's exit rule is

$$\text{Exit if } P < ATC.$$

That is, a firm exits if the price of its good is less than the average total cost of production.

A parallel analysis applies to entrepreneurs who could establish new firms. They have an incentive to enter the market if doing so would be profitable, which occurs if the price exceeds average total cost. The entry rule is

$$\text{Enter if } P > ATC.$$

The rule for entry is exactly the opposite of the rule for exit.

We can now describe a competitive firm's long-run profit-maximizing strategy. If it produces anything, it chooses the quantity at which marginal cost equals the price of the good. Yet if the price is less than the average total cost at that quantity, the firm decides to exit (or not enter) the market. Figure 4 illustrates these results. **The competitive firm's long-run supply curve is the portion of its marginal-cost curve that lies above the average-total-cost curve.**

15-2f Measuring Profit in Our Graph for the Competitive Firm

As we study exit and entry, it is useful to analyze the firm's profit in more detail. Recall that profit equals total revenue (TR) minus total cost (TC):

$$\text{Profit} = TR - TC.$$

We can rewrite this definition by multiplying and dividing the right side by Q:

$$\text{Profit} = (TR/Q - TC/Q) \times Q.$$

Note that TR/Q is average revenue, which is the price, P, and TC/Q is average total cost, ATC. Therefore,

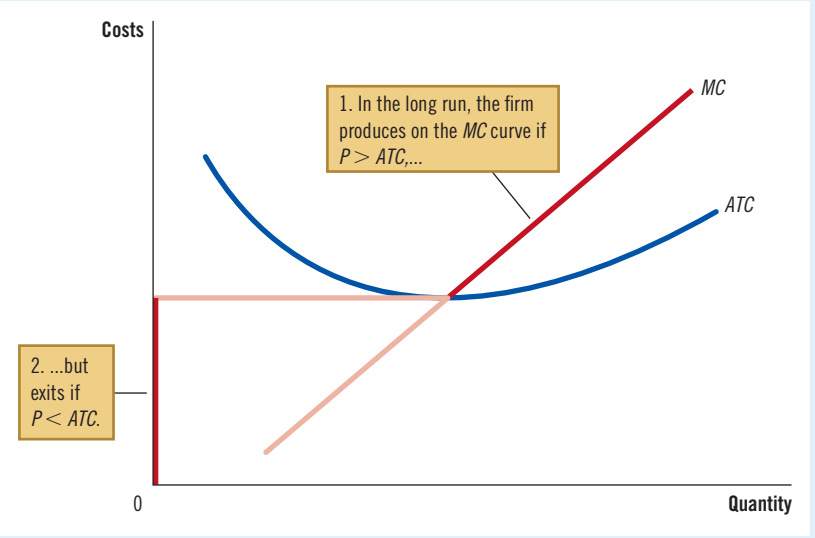
$$\text{Profit} = (P - ATC) \times Q.$$

This way of expressing the firm's profit allows us to measure profit on our graphs.

Figure 4

The Competitive Firm's Long-Run Supply Curve

In the long run, the competitive firm's supply curve is the portion of its marginal-cost curve (*MC*) that lies above its average-total-cost curve (*ATC*). If the price falls below average total cost, the firm is better off exiting the market.

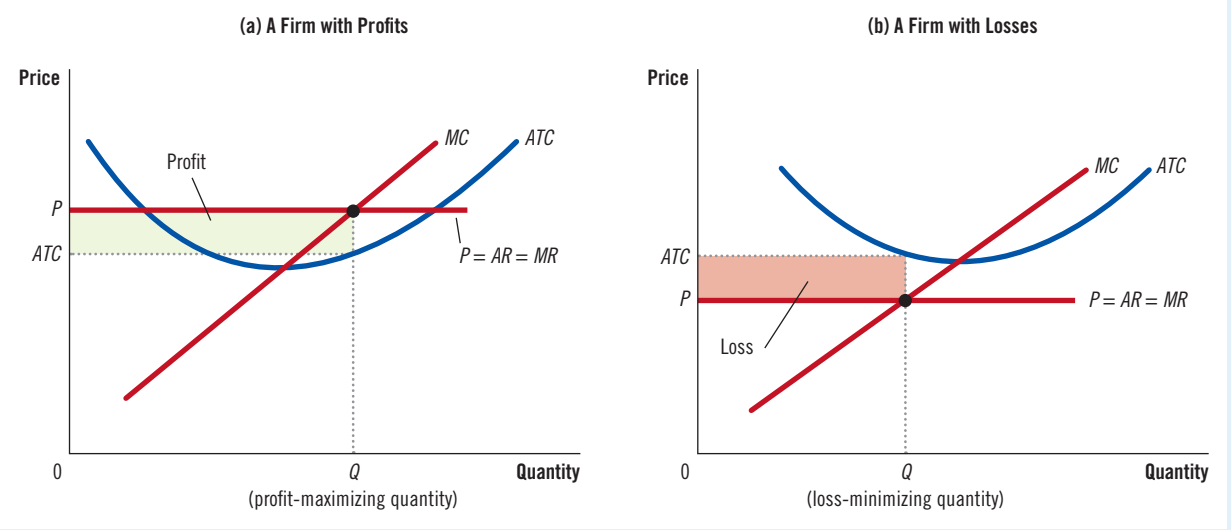


Panel (a) of Figure 5 shows a firm earning positive profit. As we have already discussed, the firm maximizes profit by producing the quantity at which price equals marginal cost. Now look at the shaded rectangle. The height of the rectangle is $P - ATC$, the difference between price and average total cost. The width of the rectangle is Q, the quantity produced. Therefore, the area of the rectangle is $(P - ATC) \times Q$, which is the firm's profit.

Figure 5

Profit as the Area between Price and Average Total Cost

The area of the shaded box between price and average total cost represents the firm's profit. The height of this box is price minus average total cost ($P - ATC$), and the width of the box is the quantity of output (Q). In panel (a), price is greater than average total cost, so the firm has positive profit. In panel (b), price is less than average total cost, so the firm incurs a loss.



Similarly, panel (b) of this figure shows a firm with losses (negative profit). In this case, maximizing profit means minimizing losses, a task accomplished once again by producing the quantity at which price equals marginal cost. Now consider the shaded rectangle. The height of the rectangle is $ATC - P$, and the width is Q. The area is $(ATC - P) \times Q$, which is the firm's loss. Because a firm in this situation is not making enough revenue on each unit to cover its average total cost, it would exit the market in the long run.

15-2g A Brief Recap

We can sum up our analysis of the competitive firm with a dialogue between two business partners. Fred and Wilma have just bought a granite quarry, which produces material for kitchen countertops. Because they compete with many other quarries, they take the price of granite as given by market conditions. Wilma, an economics major, is explaining to Fred how they should make supply decisions.

Fred: How much output should we produce to maximize profit?
Wilma: If we produce anything, we should pick the level of output at which $P = MC$.
Fred: Will we make a profit?
Wilma: We will if, at that level of output, $P > ATC$. If $P < ATC$, we will make a loss.
Fred: What should we do if that output makes a profit?
Wilma: Be happy and stay in business.
Fred: And if that output makes a loss?
Wilma: Plan on exiting in the long run.
Fred: In that case, should we keep operating in the short run?
Wilma: We should if $P > AVC$. Staying open minimizes our losses.
Fred: What if $P < AVC$?
Wilma: Then we should shut down as quickly as possible and plan our exit.
Fred: So our long-run supply curve is the MC curve above the ATC curve, and our short-run supply curve is the MC curve above the AVC curve.
Wilma: Yes, Fred, that's the plan. Table 3 summarizes everything you need to know. (Theirs is carved in granite.)

Table 3

Profit-Maximizing Rules for a Competitive Firm

1. Find Q at which $P = MC$.
2. If $P < AVC$, shut down immediately and remain out of business.
3. If $AVC < P < ATC$, operate in the short run but exit in the long run.
4. If $ATC < P$, stay in business and enjoy your profits!

QuickQuiz

3. A competitive firm maximizes profit by choosing the quantity at which
   a. average total cost is at its minimum.
   b. marginal cost equals the price.
   c. average total cost equals the price.
   d. marginal cost equals average total cost.
4. A competitive firm's short-run supply curve is its _____ cost curve above its _____ cost curve.
   a. average-total-; marginal-
   b. average-variable-; marginal-
   c. marginal-; average-total-
   d. marginal-; average-variable-
5. If a profit-maximizing competitive firm is producing a quantity at which marginal cost is between average variable cost and average total cost, it will
   a. keep producing in the short run but exit the market in the long run.
   b. shut down in the short run but return to production in the long run.
   c. shut down in the short run and exit the market in the long run.
   d. keep producing both in the short run and in the long run.

Answers are at the end of the chapter.

15-3 The Supply Curve in a Competitive Market

The supply curves for entire markets are built on the supply decisions of individual firms. There are two cases to consider: (1) markets with a fixed number of firms and (2) markets in which firms can enter and exit. Both cases are important, for each applies to a specific time horizon. Over short periods, entry and exit are often difficult, making it reasonable to assume a fixed number of firms. But over long stretches, entry and exit become easier, and so the number of firms can adjust to changing market conditions.

15-3a The Short Run: Market Supply with a Fixed Number of Firms

Imagine a market with 1,000 identical firms. Each firm acts according to our standard model: For any price, it supplies the quantity of output at which its marginal cost equals the price. Panel (a) of Figure 6 shows this. As long as price exceeds average variable cost, each firm's marginal-cost curve is its supply curve. The quantity of output supplied to the market equals the sum of the quantities supplied by each of the 1,000 firms. The market supply curve is derived by horizontally adding the supply curves of all the firms (as we did with Ben and Jerry in Chapter 4). As panel (b) of Figure 6 shows, the quantity supplied to the market is 1,000 times the quantity supplied by each of these identical firms.

15-3b The Long Run: Market Supply with Entry and Exit

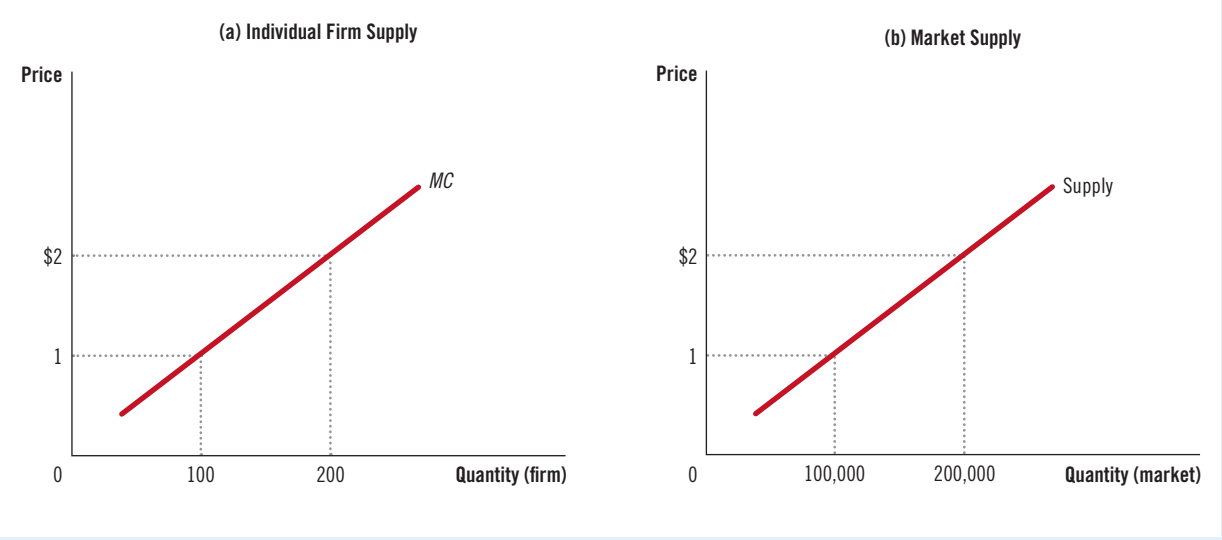
Now consider what happens when firms can enter and exit the market. Let's suppose that everyone has access to the same production technology and access to the same markets to buy the inputs for production. As a result, all current and potential firms have the same cost curves.

Decisions about entry and exit in a market of this type depend on the incentives facing the owners of existing firms and the entrepreneurs who could start new firms. If firms already in the market are profitable, new firms will enter. This entry will expand the number of firms, increase the quantity of the good supplied, and drive down prices and profits. Conversely, if firms in the market are making losses, some existing firms will exit. Their exit will reduce the number of firms, decrease the quantity of the good supplied, and drive up prices and profits. **At the end of**

Figure 6

Short-Run Market Supply

In the short run, the number of firms in the market is fixed. As a result, the market supply curve, shown in panel (b), reflects the sum of individual firms' marginal-cost curves, shown in panel (a). Here, in a market of 1,000 identical firms, the quantity of output supplied to the market is 1,000 times the quantity supplied by each firm.



this process of entry and exit, firms that remain in the market must be making zero economic profit. That conclusion may seem strange but have no fear: It will be explained shortly.

Recall that we can write a firm's profit as

$$\text{Profit} = (P - ATC) \times Q.$$

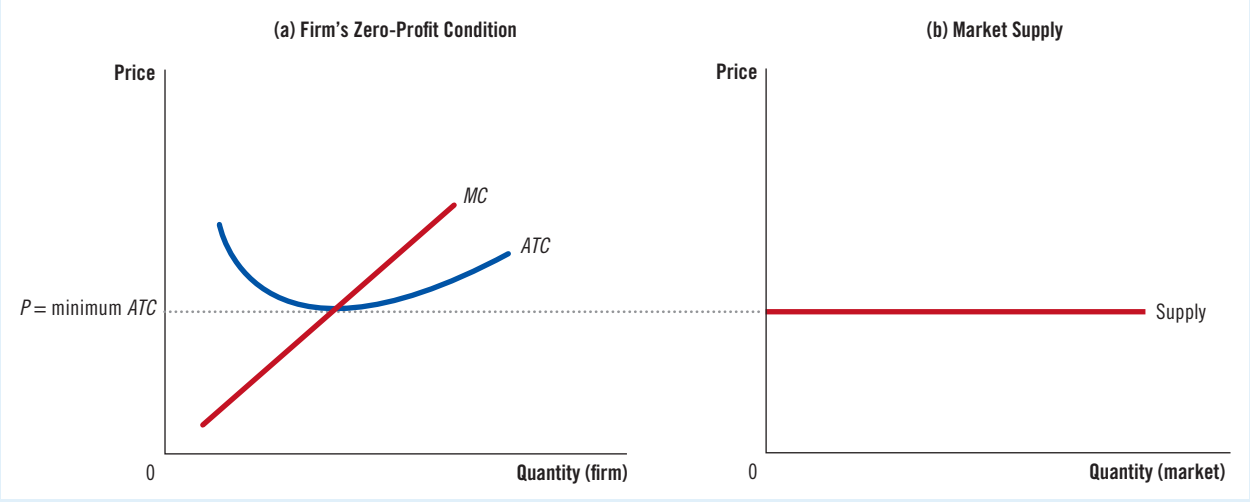
This equation shows that an operating firm has zero profit if and only if the price of the good equals the average total cost of production. If price is above average total cost, profit is positive, which encourages new firms to enter. If price is less than average total cost, profit is negative, which encourages some firms to exit. **The process of entry and exit ends only when price and average total cost are driven to equality.**

This line of reasoning leads to a surprising implication. We noted earlier in the chapter that competitive firms maximize profits by choosing a quantity at which price equals marginal cost. We just noted that free entry and exit force price to equal average total cost. But if price is to equal both marginal cost and average total cost, these two measures of cost must equal each other. Marginal cost and average total cost are equal, however, only when the firm is operating at the minimum of average total cost. Recall from the preceding chapter that economists use the term **efficient scale** to describe the level of production with the lowest average total cost. **Therefore, in the long-run equilibrium of a competitive market with free entry and exit, firms operate at their efficient scale.**

Figure 7

Long-Run Market Supply

In the long run, firms will enter or exit the market until profit is driven to zero. As a result, price equals the minimum of average total cost, as shown in panel (a). The number of firms adjusts to ensure that all demand is satisfied at this price. The long-run market supply curve is horizontal at this price, as shown in panel (b).



Panel (a) of Figure 7 shows a firm in such a long-run equilibrium. In this figure, price P equals marginal cost MC, so the firm is maximizing profit. Price also equals average total cost ATC, so profit is zero. New firms have no incentive to enter the market, and existing firms have no incentive to leave the market.

From this analysis of firm behavior, we can determine the long-run supply curve for the market. In a market with free entry and exit, there is only one price consistent with zero profit—the minimum of average total cost. As a result, the long-run market supply curve must be horizontal at this price, as illustrated by the perfectly elastic supply curve in panel (b) of Figure 7. Any price above this level would generate profits, leading to entry and an increase in the total quantity supplied. Any price below this level would generate losses, leading to exit and a decrease in the total quantity supplied. Eventually, the number of firms in the market adjusts so that price equals the minimum of average total cost, and there are enough firms to satisfy all the demand at this price.

"We're a nonprofit organization—we don't intend to be, but we are!"

15-3c Why Do Competitive Firms Stay in Business If They Make Zero Profit?

At first, it might seem odd that competitive firms earn zero profit in the long run. After all, people start businesses to make money. If entry eventually drives profit to zero, there might seem to be little reason to stay in business.

To understand the zero-profit condition more fully, recall that profit equals total revenue minus total cost and that total cost includes all the opportunity costs of the firm. In particular, total cost includes the time and money that the firm owners devote to the business. What is crucial is that in the zero-profit equilibrium, the firm's revenue must compensate the owners for these opportunity costs.

Consider an example. Suppose that, to start a farm, a farmer had to invest $1 million, which otherwise could have earned $40,000 a year in interest in a bank account. In addition, the farmer had to give up another job that would have paid $60,000 a year. Then the opportunity cost of farming includes both the forgone interest and the forgone wages—a total of $100,000. Even if the farm's profit is driven to zero, its revenue compensates the farmer for these opportunity costs.

Recall that accountants and economists measure costs differently. As the previous chapter noted, accountants keep track of explicit costs but not implicit ones. They measure costs that require an outflow of money but ignore the opportunity costs for which no money leaves the firm. As a result, in the zero-profit equilibrium, economic profit is zero, but accounting profit is positive. The farmer's accountant, for instance, would conclude that the farm earned a profit of $100,000, which is why the farmer stays in business.

15-3d A Shift in Demand in the Short Run and Long Run

Now let's turn to how markets respond to changes in demand. Because firms can enter and exit in the long run but not in the short run, the response of a market to a change in demand depends on the time horizon. To see this, let's trace the effects of a shift in demand over time.

Suppose the market for milk begins in a long-run equilibrium. Firms are earning zero profit, so price equals the minimum of average total cost. Panel (a) of Figure 8 shows this situation. The long-run equilibrium is point A, the quantity sold in the market is Q_1, and the price is P_1.

Now suppose scientists discover that milk has miraculous health benefits, causing a surge in demand. That is, the quantity of milk demanded at every price increases, and the demand curve for milk shifts outward from D_1 to D_2, as in panel (b). The short-run equilibrium moves from point A to point B; the quantity rises from Q_1 to Q_2, and the price rises from P_1 to P_2. All the firms in the market respond to the higher price by producing more milk. Because each firm's supply curve reflects its marginal-cost curve, how much each firm increases production depends on the marginal-cost curve. In the new short-run equilibrium, the price of milk exceeds average total cost, so the firms are making positive profit.

Over time, this profit encourages new firms to enter. For example, some farmers supplying other products may switch to producing milk. As the number of suppliers grows, the quantity supplied at every price increases, the short-run supply curve shifts to the right from S_1 to S_2, as in panel (c), and this shift causes the price to fall. Eventually, the price is driven back down to the minimum of average total cost, profits are zero, and firms stop entering. The market reaches a new long-run equilibrium, point C. The price of milk has returned to P_1, but the quantity produced has risen to Q_3. Each firm is again producing at its efficient scale, but because more firms are in the dairy business, the quantity of milk produced and sold is higher.

15-3e Why the Long-Run Supply Curve Might Slope Upward

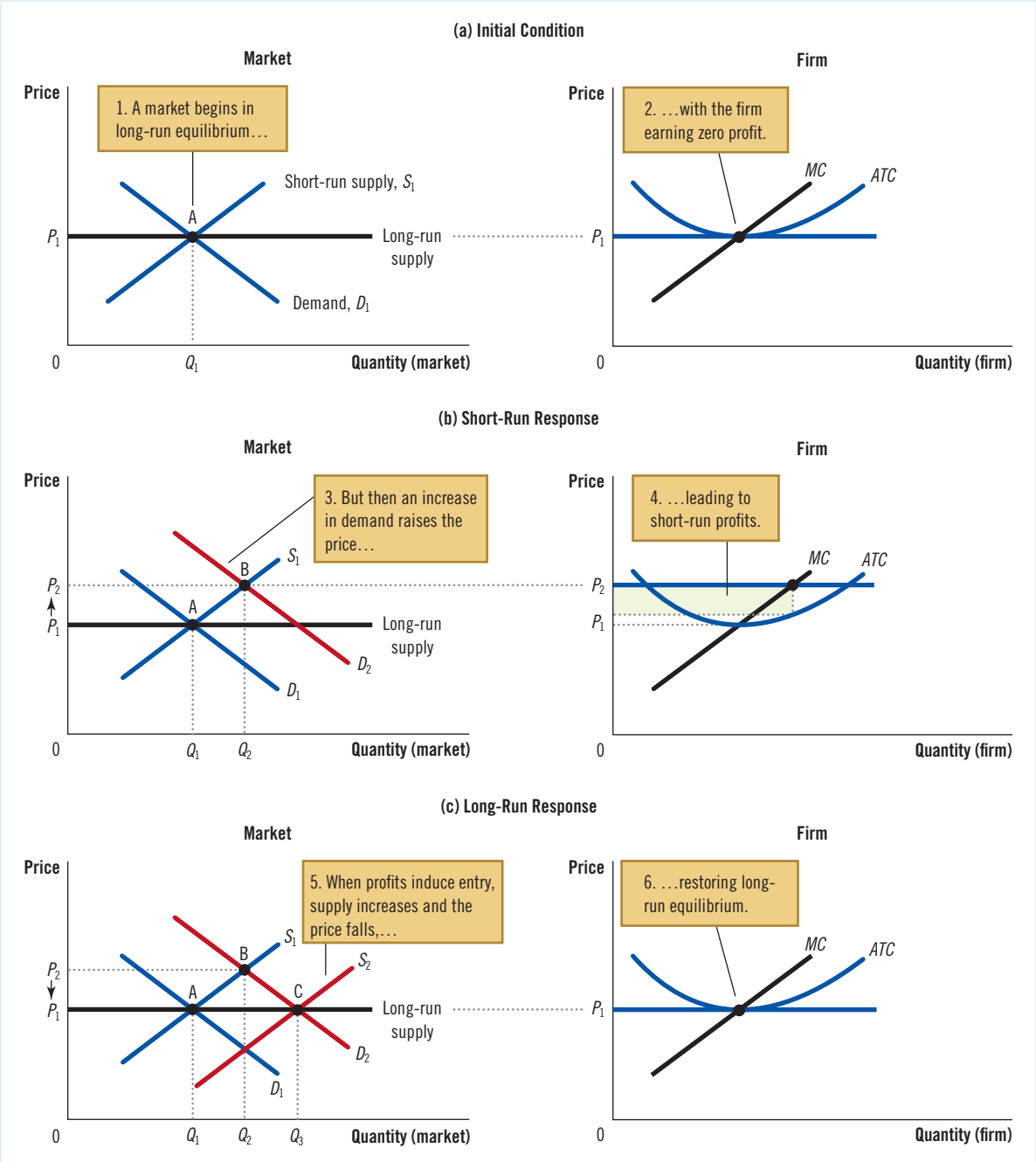
We have seen that entry and exit can cause the long-run market supply curve to be perfectly elastic. In essence, there are many potential entrants, each of which faces the same costs. As a result, the long-run market supply curve is horizontal at the minimum of average total cost. When demand increases, the long-run result is an increase in the number of firms and in the total quantity supplied, without any change in the price.

There are, however, two reasons that the long-run market supply curve might slope upward. The first is that some resources used in production may be available only

Figure 8

An Increase in Demand in the Short Run and Long Run

Panel (a) shows a market in a long-run equilibrium at point A. In this equilibrium, each firm makes zero profit, and the price equals the minimum average total cost. Panel (b) shows what happens in the short run when demand rises from D_1 to D_2. The equilibrium goes from point A to point B, price rises from P_1 to P_2, and the quantity sold in the market rises from Q_1 to Q_2. Because price now exceeds average total cost, each firm now makes a profit, which, over time, encourages new firms to enter the market. Panel (c) shows how this entry shifts the short-run supply curve to the right from S_1 to S_2. In the new long-run equilibrium, point C, price has returned to P_1, but the quantity sold has increased to Q_3. Profits are again zero, and price is back to the minimum of average total cost, but the market has more firms to satisfy the greater demand.



in limited quantities. Consider the market for farm products. Anyone can buy land and start a farm, but the quantity of land is limited. As more people become farmers, they bid up the price of farmland, increasing the costs of all farmers in the market. An increase in demand for farm products can't increase the quantity supplied without also inducing a rise in farmers' costs, which means a rise in price. The result is a long-run market supply curve that slopes upward, even with free entry into farming.

A second reason for an upward-sloping long-run supply curve is that firms may have different costs. Take the market for painters. Anyone can enter it, but not everyone has the same costs. People differ in how fast they work, and they differ in the alternative uses of their time. For any price, those with lower costs are more likely to enter than those with higher costs. To increase the quantity of painting services supplied, additional entrants must be induced to enter the market. Because these new entrants have higher costs, the price must rise to make entry profitable for them. As a result, the long-run market supply curve for painting services slopes upward even with free entry into the market.

Notice that when firms have different costs, some firms earn profit even in the long run. In this case, the market price reflects the average total cost of the **marginal firm**—the firm that would exit if the price were any lower. This firm earns zero profit, but firms with lower costs earn positive profit. Entry does not eliminate this profit because would-be entrants have higher costs than firms already in the market. Higher-cost firms will enter only if the price rises, making the market profitable for them.

For these two reasons, a higher price may be necessary to induce a larger quantity supplied, in which case the long-run supply curve is upward-sloping rather than horizontal. Nonetheless, the basic lesson about entry and exit remains true. **Because firms can enter and exit more easily in the long run than in the short run, the long-run supply curve is typically more elastic than the short-run supply curve.**

QuickQuiz

6. In the long-run equilibrium of a competitive market with identical firms, what are the relationships among price P, marginal cost MC, and average total cost ATC?
   a. $P > MC$ and $P > ATC$.
   b. $P > MC$ and $P = ATC$.
   c. $P = MC$ and $P > ATC$.
   d. $P = MC$ and $P = ATC$.

7. In the short-run equilibrium of a competitive market with identical firms, if new firms are getting ready to enter, what are the relationships among price P, marginal cost MC, and average total cost ATC?
   a. $P > MC$ and $P > ATC$.
   b. $P > MC$ and $P = ATC$.
   c. $P = MC$ and $P > ATC$.
   d. $P = MC$ and $P = ATC$.

8. Suppose pretzel stands in New York City are a perfectly competitive market in long-run equilibrium. One day, the city starts imposing a $100 per month tax on each stand. How does this policy affect the number of pretzels consumed in the short run and the long run?
   a. down in the short run, no change in the long run
   b. up in the short run, no change in the long run
   c. no change in the short run, down in the long run
   d. no change in the short run, up in the long run

Answers are at the end of the chapter.

15-4 Conclusion: Behind the Supply Curve

We have been discussing the behavior of profit-maximizing firms that supply goods in perfectly competitive markets. You may recall from Chapter 1 that one of the **Ten Principles of Economics** is that rational people think at the margin. This chapter has applied this idea to the competitive firm. Marginal analysis has given us a

theory of the supply curve in a competitive market and a deeper understanding of market outcomes.

We have learned that when you buy a good from a firm in a competitive market, the price you pay is close to the cost of producing it. In particular, if firms are competitive and profit-maximizing, the price of a good equals the marginal cost of making that good. And if firms can freely enter and exit the market, the price also equals the lowest possible average total cost of production.

We have assumed throughout this chapter that firms are price takers, but many of the tools developed here are also useful for studying firms in less competitive markets. The next chapter examines the behavior of firms with market power. Marginal analysis will again be useful, but it will have very different implications for a firm's production decisions and for the nature of market outcomes.

Chapter in a Nutshell

- Because a competitive firm is a price taker, its revenue is proportional to the amount of output it produces. The price of the good equals both the firm's average revenue and its marginal revenue.
- To maximize profit, a firm chooses a quantity of output such that marginal revenue equals marginal cost. Because marginal revenue for a competitive firm equals the market price, the firm chooses the quantity at which price equals marginal cost. Thus, the firm's marginal-cost curve is its supply curve.
- In the short run, when fixed costs are sunk, the firm will shut down temporarily if the price of the good is less than average variable cost. In the long run, when the firm can recover both fixed and variable costs, it will exit if the price is less than average total cost.
- In a market with free entry and exit, economic profit is driven to zero in the long run. In this long-run equilibrium, all firms produce at the efficient scale, price equals the minimum of average total cost, and the number of firms adjusts to satisfy the quantity demanded at this price.
- Changes in demand have different effects over different time horizons. In the short run, an increase in demand raises prices and leads to profits, and a decrease in demand lowers prices and leads to losses. But if firms can freely enter and exit the market, then, in the long run, the number of firms adjusts to drive the market back to the zero-profit equilibrium.

Key Concepts

competitive market, p. 288
average revenue, p. 290
marginal revenue, p. 290
sunk cost, p. 295

Questions for Review

1. What are the main characteristics of a competitive market?
2. Explain the difference between a firm's revenue and its profit. Which do firms maximize?
3. Draw the cost curves for a typical firm. Explain how a competitive firm chooses the level of output that maximizes profit. At that level of output, show on your graph the firm's total revenue and total cost.
4. Under what conditions will a firm shut down temporarily? Explain.
5. Under what conditions will a firm exit a market? Explain.
6. Does a competitive firm's price equal its marginal cost in the short run, in the long run, or both? Explain.
7. Does a competitive firm's price equal the minimum of its average total cost in the short run, in the long run, or both? Explain.
8. Are market supply curves typically more elastic in the short run or in the long run? Explain.

Problems and Applications

1. Many small boats are made of fiberglass and a resin derived from crude oil. Suppose that the price of oil rises.
   a. Using diagrams, show what happens to the cost curves of an individual boat-making firm and to the market supply curve.
   b. What happens to the profits of boat makers in the short run? What happens to the number of boat makers in the long run?
2. Leah's lawn-mowing service is a profit-maximizing competitive firm. Leah mows lawns for $27 each. Her total cost each day is $280, of which $30 is a fixed cost. She mows 10 lawns a day. What can you say about Leah's short-run decision regarding shutdown and her long-run decision regarding exit?
3. Consider total cost and total revenue given in the following table:

Quantity	0	1	2	3	4	5	6	7
Total cost	$8	9	10	11	13	19	27	37
Total revenue	$0	8	16	24	32	40	48	56

   a. Calculate profit for each quantity. How much should the firm produce to maximize profit?
   b. Calculate marginal revenue and marginal cost for each quantity. Graph them. (Hint: Put the points between whole numbers. For example, the marginal cost between 2 and 3 should be graphed at 2½.) At what quantity do these curves cross? How does this relate to your answer to part (a)?
   c. Can you tell whether this firm is in a competitive industry? If so, can you tell whether the industry is in a long-run equilibrium?
4. Ball Bearings, Inc., faces costs of production as follows:

Quantity (cases)	Total Fixed Cost	Total Variable Cost
0	$100	$0
1	100	50
2	100	70
3	100	90
4	100	140
5	100	200
6	100	360

   a. Calculate the company's average fixed cost, average variable cost, average total cost, and marginal cost at each level of production.
   b. The price of a case of ball bearings is $50. Seeing that the company can't make a profit, the chief executive officer (CEO) decides to shut down operations. What is the firm's profit or loss? Is shutting down a wise decision? Explain.
   c. Vaguely remembering an introductory economics course, the chief financial officer tells the CEO it is better to produce 1 case of ball bearings because marginal revenue equals marginal cost at that quantity. What is the firm's profit or loss at that level of production? Is producing 1 case the best decision? Explain.
5. Suppose the book-printing industry is competitive and begins in a long-run equilibrium.
   a. Draw a diagram showing the average total cost, marginal cost, marginal revenue, and supply curve of the typical firm in the industry.
   b. Hi-Tech Printing Company invents a new process that sharply reduces the cost of printing books. What happens to Hi-Tech's profits and to the price of books in the short run when Hi-Tech's patent prevents other firms from using the new technology?
   c. What happens in the long run when the patent expires and other firms are free to use the technology?
6. A firm in a competitive market receives $500 in total revenue and has marginal revenue of $10. What is the average revenue, and how many units were sold?
7. A profit-maximizing firm in a competitive market is currently producing 100 units of output. It has average revenue of $10, average total cost of $8, and fixed costs of $200.
   a. What is its profit?
   b. What is its marginal cost?
   c. What is its average variable cost?
   d. Is the efficient scale of the firm more than, less than, or exactly 100 units?
8. The market for fertilizer is perfectly competitive. Firms in the market are producing output but are currently incurring economic losses.
   a. How does the price of fertilizer compare to the average total cost, the average variable cost, and the marginal cost of producing fertilizer?
   b. Draw two graphs, side by side, illustrating the present situation for the typical firm and for the market.
   c. Assuming there is no change in either demand or the firms' cost curves, explain what will happen in the long run to the price of fertilizer, marginal

cost, average total cost, the quantity supplied by each firm, and the total quantity supplied to the market.

9. The market for apple pies in the city of Ectenia is competitive and has the following demand schedule:

Price	Quantity Demanded
$1	1,200 pies
2	1,100
3	1,000
4	900
5	800
6	700
7	600
8	500
9	400
10	300
11	200
12	100
13	0

Each producer in the market has fixed costs of $9 and the following marginal cost schedule:

Quantity	Marginal Cost
1 pie	$2
2	4
3	6
4	8
5	10
6	12

a. Compute each producer's total cost and average total cost for each quantity from 1 to 6 pies.
b. The price of a pie is now $11. How many pies are sold? How many pies does each producer make? How many producers are there? How much profit does each producer earn?
c. Is the situation described in part (b) a long-run equilibrium? Why or why not?
d. Suppose that in the long run, there is free entry and exit. How much profit does each producer earn in the long-run equilibrium? What is the market price? How many pies does each producer make? How many pies are sold in the market? How many pie producers are operating?

10. An industry currently has 100 firms, each of which has fixed costs of $16 and average variable cost as follows:

Quantity	Average Variable Cost
1	$1
2	2
3	3
4	4
5	5
6	6

a. Compute a firm's marginal cost and average total cost for each quantity from 1 to 6.
b. The equilibrium price is currently $10. How much does each firm produce? What is the total quantity supplied in the market?
c. In the long run, firms can enter and exit the market, and all entrants have the same costs as above. As this market makes the transition to its long-run equilibrium, will the price rise or fall? Will the quantity demanded rise or fall? Will the quantity supplied by each firm rise or fall? Explain your answers.
d. Graph the long-run supply curve for this market with specific numbers on the axes as relevant.

11. Suppose that each firm in a competitive industry has the following costs:

Total cost: $TC = 50 + \frac{1}{2} q^2$

Marginal cost: $MC = q$

where q is an individual firm's quantity produced. The market demand curve for this product is:

Demand: $Q^D = 120 - P$

where P is the price and Q is the total quantity of the good. Currently, there are 9 firms in the market.

a. What is each firm's fixed cost? What is its variable cost? Give the equation for average total cost.
b. Graph the average-total-cost curve and the marginal-cost curve for q from 5 to 15. At what quantity is the average-total-cost curve at its minimum? What is marginal cost and average total cost at that quantity?
c. Give the equation for each firm's supply curve.

d. Give the equation for the market supply curve for the short run in which the number of firms is fixed.
e. What is the equilibrium price and quantity for this market in the short run?
f. In this equilibrium, how much does each firm produce? Calculate each firm's profit or loss. Do firms have an incentive to enter or exit?
g. In the long run, with free entry and exit, what is the equilibrium price and quantity in this market?
h. In this long-run equilibrium, how much does each firm produce? How many firms are in the market?

Quick**Quiz Answers**

1. **c** 2. **d** 3. **b** 4. **d** 5. **a** 6. **d** 7. **c** 8. **c**

Chapter

16

Monopoly

In the 1990s, if you owned a personal computer, it most likely used some version of Windows, the operating system sold by the Microsoft Corporation, and even today, Windows computers remain popular. When Microsoft originally designed Windows, it applied for and received a copyright from the government. The copyright gives Microsoft the exclusive right to make and sell copies of the Windows operating system. Someone who wants to buy a copy has little choice but to fork over the roughly $100 that Microsoft charges for its product. These days, other operating systems are available, but they are often very different. Microsoft is said to have a **monopoly** in the market for Windows.

Microsoft's business decisions aren't well described by the model of firm behavior developed in the previous chapter. That chapter analyzed competitive markets, in which many firms offer largely identical products, so each firm has little influence over the price it receives. By contrast, because a monopoly has no close competitors, it has the power to influence the market price of its product. While a competitive firm is a **price taker**, a monopoly firm is a **price maker**.

ROMAN SAMBORSKYI/SHUTTERSTOCK.COM; ARNO VAN ENGELAND/EYEEM/GETTY IMAGES

This chapter examines the implications of this market power. We will see that market power alters the relationship between the costs a firm incurs producing a good and the price at which it sells that good. So far, we have seen that a competitive firm takes the price of its output as given by the market and then chooses the quantity it will supply so that price equals marginal cost. A monopoly is different. It charges a price that exceeds marginal cost. Sure enough, this practice is evident in the case of Microsoft's Windows. The marginal cost of Windows—the extra cost that Microsoft incurs when a customer downloads one more copy—is trivial. The market price of Windows is many times its marginal cost.

It is not surprising that monopolies charge high prices for their products. Customers of a monopoly might seem to have little choice but to pay whatever the monopoly charges. But if so, why doesn't Microsoft charge $1,000 for a copy of Windows? Or $10,000? The reason is that if the price were that high, fewer people would buy it. People would buy fewer computers, switch to other operating systems, or make illegal copies. A monopoly can control the price of what it sells, but because a high price reduces the quantity demanded, the monopoly's profits are not unlimited.

In examining how monopolies make production and pricing decisions, this chapter considers the implications of monopoly for society as a whole. Monopolies, like competitive firms, aim to maximize profit, but the pursuit of this goal has very different ramifications. In a competitive market, self-interested consumers and producers reach an equilibrium that promotes general economic well-being, as if guided by an invisible hand. But because monopolies are unchecked by competition, the outcome in a monopolized market is often not in the best interest of society.

One of the **Ten Principles of Economics** in Chapter 1 is that governments can sometimes improve market outcomes. This chapter sheds more light on this principle by examining the inefficiencies that monopolies cause and discussing how government policymakers can respond to these problems. The U.S. government keeps a close eye on Microsoft's business decisions, for example. In 1994, the government blocked Microsoft from acquiring Intuit, a leading seller of personal finance software, on the grounds that a merger between the two firms would concentrate too much market power. Similarly, in 1998, the U.S. Department of Justice objected when Microsoft started integrating its Internet Explorer browser into its Windows operating system, claiming that this practice would extend the firm's market power into new areas. In recent years, regulators in the United States and abroad have shifted their focus to firms with growing market power, such as Apple, Google, and Amazon, but they continue to monitor Microsoft's compliance with antitrust laws.

16-1 Why Monopolies Arise

monopoly
a firm that is the sole seller of a product without close substitutes

A firm is a **monopoly** if it is the sole seller of a product that doesn't have close substitutes. The fundamental cause of monopoly is **barriers to entry**: A monopoly remains the only seller in its market because other firms can't enter and compete with it. Barriers to entry, in turn, have three main sources:

- **Monopoly resources:** A single firm owns a key resource required for production.
- **Government regulation:** The government gives a single firm the exclusive right to produce a good or service.
- **The production process:** A single firm can produce output at a lower cost than a larger number of firms can.

Here's more on these barriers to entry.

16-1a Monopoly Resources

The simplest way for a monopoly to arise is for a single firm to own a key resource. Consider the market for water in a small town. If dozens of residents have working wells, the model of competitive markets in the preceding chapter describes sellers' behavior. Competition among suppliers drives the price of a gallon of water to equal the marginal cost of pumping an extra gallon. But if there is only one well in town and it is impossible to get water from anywhere else, then the owner of the well has a monopoly. Not surprisingly, the monopolist has much greater market power than any single firm in a competitive market. For a necessity like water, the monopolist can command quite a high price, even if the marginal cost of pumping an extra gallon is low.

DAVE CARPENTER/CARTOONSTOCK.COM

"Rather than a monopoly, we like to consider ourselves 'the only game in town.'"

A classic example of market power arising from the ownership of a key resource is DeBeers, the diamond company. Founded in South Africa in 1888 by Cecil Rhodes, an English businessman (and benefactor of the Rhodes scholarship), DeBeers has at times controlled up to 80 percent of the production from the world's diamond mines. Because its market share is less than 100 percent, DeBeers is not exactly a monopoly, but the company has nonetheless exerted substantial influence over the market price of diamonds.

Although exclusive ownership of a key resource can create a monopoly, this is relatively rare in practice. Economies are large, and resources are owned by many people. The natural scope of many markets is worldwide because goods are often traded internationally. There are few examples of firms that own resources for which there are no close substitutes.

16-1b Government-Created Monopolies

In many cases, monopolies arise when the government gives one person or firm the exclusive right to sell a good or service. Sometimes, a would-be monopolist receives the right out of sheer political clout. Kings once granted exclusive business licenses to their friends and allies. Autocrats continue to do so. Sometimes, the government grants a monopoly because doing so is in the public interest.

Patent and copyright laws are two important examples of how a government can create a monopoly. When a pharmaceutical company discovers a new drug, it can apply to the government for a patent. If the government deems the drug to be original, it approves the patent, which grants the company the exclusive right to manufacture and sell the drug for 20 years. Similarly, after finishing a book, a novelist can copyright it. The copyright is a government guarantee that no one can sell the work without the author's permission. Writers who can barely support themselves may not think of themselves as monopolists. But this only goes to show that being the sole seller of a product does not guarantee a large number of buyers.

The effects of patent and copyright laws are easy to see. Because these laws give one producer a monopoly, they lead to higher prices and higher profits than would occur under competition. But the laws also encourage some desirable behavior. By allowing drug companies to be monopolists in the drugs they discover, the patent laws encourage their research. By allowing authors to be monopolists in the sale of their books, the copyright laws encourage them to write more and better books.

The laws governing patents and copyrights have both benefits and costs. The benefits are the increased incentives for creative activity. They are offset, to some extent, by the costs of monopoly pricing, which we examine later in this chapter.

16-1c Natural Monopolies

natural monopoly a type of monopoly that arises because a single firm can supply a good or service to an entire market at a lower cost than could two or more firms

An industry is a **natural monopoly** when a single firm can supply a market with a good or service at a lower cost than two or more firms could. This happens when there are economies of scale over the relevant range of output. Figure 1 shows the average total costs of a firm with economies of scale. In this case, a single firm can produce any amount of output at the lowest cost. That is, for any given amount of output, a larger number of firms leads to less output per firm and higher average total cost.

The distribution of tap water is an example of a natural monopoly. To provide water to town residents, a firm must build a network of pipes. If two or more firms were to compete, each would have to incur the fixed cost of building a network. The average total cost is lowest if a single firm provides water to the entire market.

Other examples of natural monopolies appeared in Chapter 11, which noted that **club goods** are excludable but not rival in consumption. An example is a bridge used so rarely that it is never congested. The bridge is excludable because a toll collector can prevent someone from using it. The bridge is not rival in consumption because one person's use of the bridge does not hinder others' use of it. There is a large fixed cost of building the bridge but a negligible marginal cost of additional users, so the average total cost (the total cost divided by the number of trips) declines as the number of trips rises, making the bridge a natural monopoly.

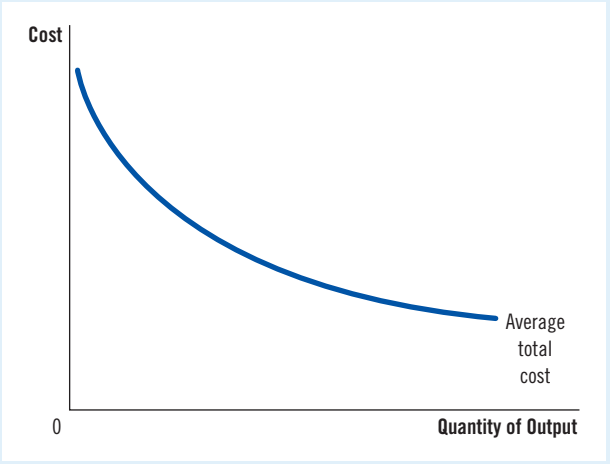
When a firm is a natural monopoly, it is less concerned about new entrants eroding its monopoly power. Normally, a firm has trouble maintaining a monopoly position without government protection or ownership of a key resource. The monopolist's profit attracts entrants into the market, and these entrants make the market more competitive. By contrast, entering a market in which another firm has a natural monopoly is unattractive. Would-be entrants know that they cannot achieve the same low costs that the monopolist enjoys because, after entry, each firm would have a smaller piece of the market.

In some cases, the size of the market determines whether an industry is a natural monopoly. Again, consider a bridge across a river. When the population is small, the bridge may be a natural monopoly. A single bridge can meet the entire demand

Figure 1

Economies of Scale as a Cause of Monopoly

When a firm's average-total-cost curve continually declines, the firm has what is called a natural monopoly. In this case, when production is divided among more firms, each firm produces less, and average total cost rises. As a result, a single firm can produce any given amount at the lowest cost.



for trips across the river at the lowest cost. Yet as the population grows and the bridge becomes congested, meeting demand may require multiple bridges. As the market expands, the natural monopoly can evolve into a more competitive market.

QuickQuiz

1. Some government grants of monopoly power may be desirable if they
   a. curtail the adverse effects of cut-throat competition.
   b. make industries more profitable.
   c. provide incentives for invention and artistic creation.
   d. save consumers from having to choose among alternative suppliers.
2. A firm is a natural monopoly if it exhibits _____ as its output increases.
   a. increasing total revenue
   b. increasing marginal cost
   c. decreasing marginal revenue
   d. decreasing average total cost

Answers are at the end of the chapter.

16-2 How Monopolies Make Production and Pricing Decisions

Having seen how monopolies arise, let's consider how a monopoly firm decides how much to produce and what price to charge. The analysis of monopoly behavior in this section is the starting point for evaluating whether monopolies are desirable and what policies a government might pursue in monopoly markets.

16-2a Monopoly versus Competition

The key difference between a competitive firm and a monopoly is the monopoly's ability to influence the price of its output. A competitive firm is small relative to the market in which it operates and, therefore, has no power to influence the price of its output. It takes the price as given by market conditions. By contrast, because a monopoly is the sole producer in its market, it can alter the price of its good by adjusting the quantity it supplies.

One way to view this difference between a competitive firm and a monopoly is to consider the demand curve that each faces. In the analysis of competitive firms in the preceding chapter, we drew the market price as a horizontal line. Because a competitive firm can sell as much or as little as it wants at this price, the competitive firm faces a horizontal demand curve, as in panel (a) of Figure 2. In effect, because the competitive firm sells a product with many perfect substitutes (the products of all the other firms in its market), the demand curve for any one firm is perfectly elastic.

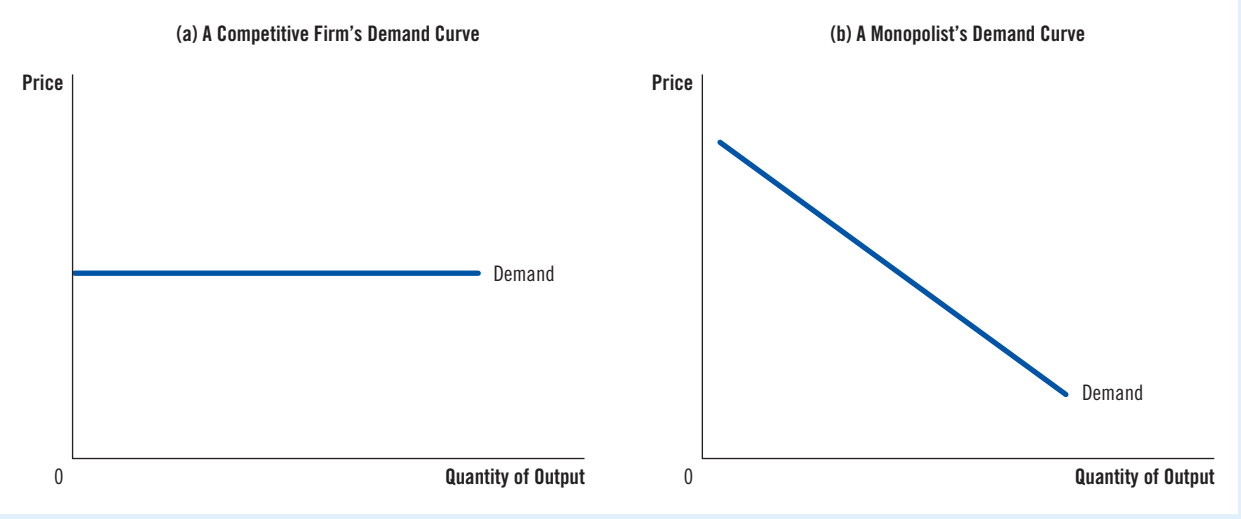
By contrast, because a monopoly is the sole producer in its market, its demand curve is simply the market demand curve, which slopes downward, as in panel (b) of Figure 2. If the monopolist raises the price of its good, consumers buy less of it. Put another way, if the monopolist reduces the quantity of output it produces and sells, the price of its output increases.

The market demand curve provides a constraint on a monopoly's ability to profit from its market power. A monopolist would prefer to charge a high price and sell a large quantity at that high price. But its demand curve makes that outcome impossible. The market demand curve describes the combinations of price and quantity

Figure 2

Demand Curves for Competitive and Monopoly Firms

As a price taker, a competitive firm faces a horizontal demand curve, as in panel (a). It can sell all it wants at the going price. But a monopoly is the sole producer in its market, so it faces the downward-sloping market demand curve, as in panel (b). If it wants to sell more output, it has to accept a lower price.



available to a monopoly firm. By adjusting the quantity produced (or equivalently, the price charged), the monopolist can choose any point on the demand curve, but it can't choose a point above the demand curve.

What price and quantity of output will the monopolist choose? As with competitive firms, we assume that the monopolist's goal is to maximize profit. Because the firm's profit is total revenue minus total costs, our next task in explaining monopoly behavior is to examine a monopolist's revenue.

16-2b A Monopoly's Revenue

Consider a town with a single water producer. Table 1 shows how the monopoly's revenue might depend on the amount of water produced.

Columns (1) and (2) show the monopolist's demand schedule. If the monopolist produces 1 gallon of water, it can sell that gallon for $10. If it produces 2 gallons, it must lower the price to $9 to sell both gallons. If it produces 3 gallons, it must lower the price to $8. And so on. If you graphed these two columns of numbers, you would get a typical downward-sloping demand curve.

Column (3) of the table presents the monopolist's **total revenue**. It equals the quantity sold [from column (1)] times the price [from column (2)]. Column (4) computes the firm's **average revenue**, the amount the firm receives per unit sold. Average revenue is calculated by taking the number for total revenue in column (3) and dividing it by the quantity of output in column (1). As the previous chapter explained, average revenue always equals the price of the good. This is true for monopolists as well as for competitive firms.

Column (5) of Table 1 computes the firm's **marginal revenue**, the amount of revenue that the firm receives for each additional unit of output. Marginal revenue

Table 1

A Monopoly's Total, Average, and Marginal Revenue

(1) Quantity of Water (*Q*)	(2) Price (*P*)	(3) Total Revenue ($TR = P \times Q$)	(4) Average Revenue ($AR = TR/Q$)	(5) Marginal Revenue ($MR = \Delta TR/\Delta Q$)
0 gallons	$11	$0	—	
				$10
1	10	10	$10	
				8
2	9	18	9	
				6
3	8	24	8	
				4
4	7	28	7	
				2
5	6	30	6	
				0
6	5	30	5	
				−2
7	4	28	4	
				−4
8	3	24	3	

is calculated by taking the change in total revenue when output increases by 1 unit. For example, when the firm increases production from 3 to 4 gallons of water, the total revenue it receives increases from $24 to $28. Marginal revenue from the sale of the fourth gallon is $28 minus $24, or $4.

Table 1 shows an important result in the basic model of monopoly behavior: **A monopolist's marginal revenue is less than the price of its good.** For example, if the firm raises the production of water from 3 to 4 gallons, it increases total revenue by only $4, even though it sells each gallon for $7. For a monopoly, marginal revenue is lower than price because a monopoly faces a downward-sloping demand curve. To increase the amount sold, a monopoly firm must lower the price it charges to all customers. To sell the fourth gallon of water, the monopolist must earn $1 less revenue for each of the first 3 gallons. This $3 loss accounts for the difference between the price of the fourth gallon ($7) and the marginal revenue of that fourth gallon ($4).

Marginal revenue for monopolies is very different from marginal revenue for competitive firms. When a monopoly increases the amount it sells, there are two effects on total revenue ($P \times Q$):

- **The output effect:** More output is sold, so Q is higher, which increases total revenue.
- **The price effect:** The price falls, so P is lower, which decreases total revenue.

Because a competitive firm can sell all it wants at the market price, there is no price effect. When it increases production by 1 unit, it receives the market price for that unit, and it does not receive any less for the units it was already selling. That is, because the competitive firm is a price taker, its marginal revenue equals the price of its good. By contrast, when a monopoly increases production by 1 unit, it must reduce the price it charges for every unit it sells, and this price cut reduces revenue from the units it was already selling. As a result, a monopoly's marginal revenue is less than its price.

Figure 3 graphs the demand curve and the marginal-revenue curve for a monopoly. (Because the monopoly's price equals its average revenue, the demand curve is also the average-revenue curve.) These two curves always start at the same point on the vertical axis because the marginal revenue of the first unit sold equals the price of the good. But for the reason just discussed, the monopolist's marginal revenue on all units after the first is less than the price. That's why a monopoly's marginal-revenue curve lies below its demand curve.

You can see in Figure 3 (as well as in Table 1) that marginal revenue can even become negative. That happens when the price effect on revenue outweighs the output effect. In this case, an additional unit of output causes the price to fall by enough that the firm, despite selling more units, receives less revenue.

16-2c Profit Maximization

Now that we have considered the revenue of a monopoly firm, we are ready to examine how such a firm maximizes profit. Recall from Chapter 1 that one of the **Ten Principles of Economics** is that rational people think at the margin. This lesson is as true for monopolists as it is for competitive firms. Here, we apply the logic of marginal analysis to the monopolist's decision about how much to produce.

Figure 3

Demand and Marginal-Revenue Curves for a Monopoly

The demand curve shows how the quantity sold affects the price. The marginal-revenue curve shows how the firm's revenue changes when the quantity increases by 1 unit. Because the price on **all** units sold must fall if the monopoly increases production, marginal revenue is less than the price.

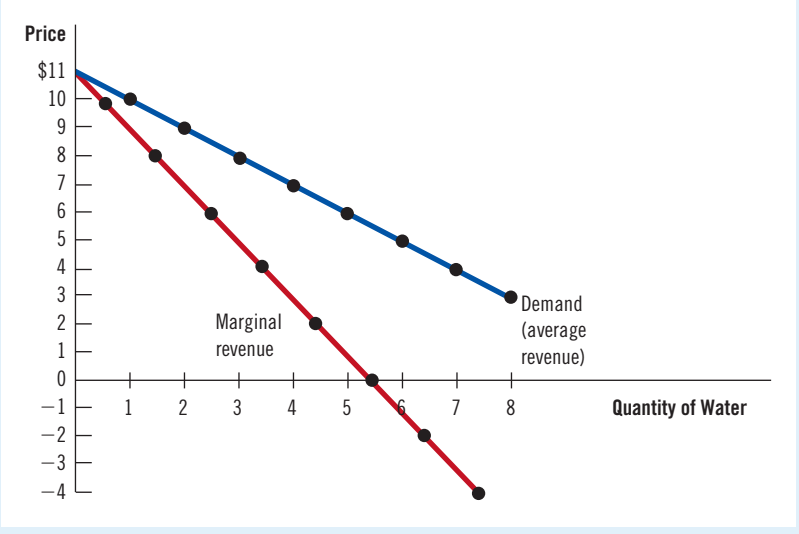


Figure 4

Profit Maximization for a Monopoly

A monopoly maximizes profit by choosing the quantity at which marginal revenue equals marginal cost (point A). It then uses the demand curve to find the price that will induce consumers to buy that quantity (point B).

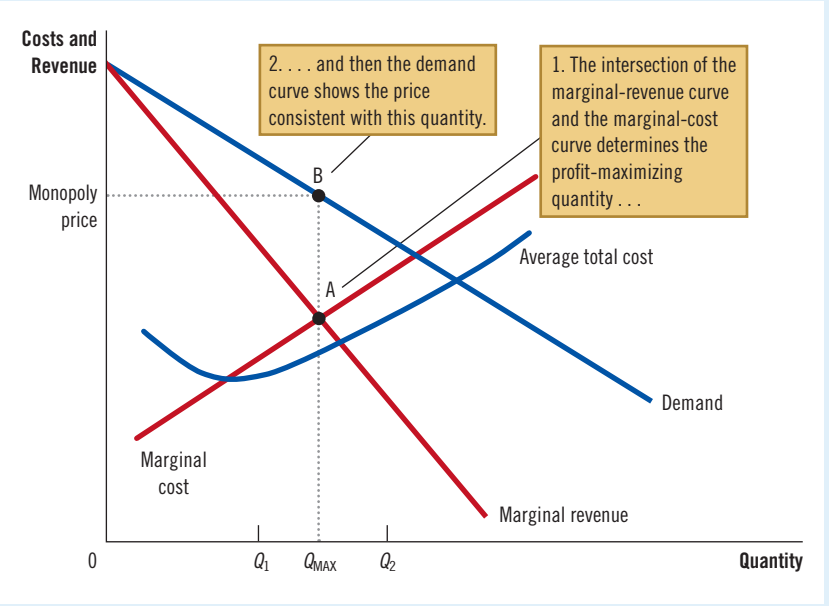


Figure 4 graphs the demand curve, the marginal-revenue curve, and the cost curves for a monopoly firm. These should all seem familiar: The demand and marginal-revenue curves are like those in Figure 3, and the cost curves are like those encountered in the last two chapters. These curves contain all the information we need to determine the level of output that a profit-maximizing monopolist will choose.

Suppose that the firm is producing at a low level of output, such as Q_1. In this case, marginal revenue exceeds marginal cost. If the firm were to increase production by 1 unit, the additional revenue would more than cover the additional costs, and profit would rise. In other words, when marginal revenue exceeds marginal cost, the firm should increase output.

Similar reasoning applies at a high level of output, such as Q_2. In this case, marginal cost exceeds marginal revenue. If the firm were to reduce production by 1 unit, the costs saved would be greater than the revenue lost, increasing profit. Thus, when marginal cost exceeds marginal revenue, the firm should reduce output.

In the end, the firm adjusts production until it reaches Q_{MAX}, the quantity at which marginal revenue equals marginal cost. **The monopolist's profit-maximizing quantity of output is determined by the intersection of the marginal-revenue curve and the marginal-cost curve.** In Figure 4, this intersection occurs at point A.

How does the monopoly find the profit-maximizing price for its product? The demand curve gives the answer. It relates the amount that customers are willing to pay to the quantity sold. After the monopoly finds the profit-maximizing quantity (at which $MR = MC$), it looks to the demand curve to find the highest price it can charge at that quantity. In Figure 4, the profit-maximizing price is found at point B.

Now compare the outcomes for a competitive firm and a monopoly. They are alike in one way: To maximize profit, both firms choose the quantity of output at which marginal revenue equals marginal cost. Yet there is an important difference:

FYI

Why a Monopoly Does Not Have a Supply Curve

We have analyzed the price in a monopoly market using the market demand curve and the firm's cost curves but haven't mentioned the market supply curve. Yet when we analyzed prices in competitive markets beginning in Chapter 4, the two most important words were always "supply" and "demand."

What happened to the supply curve? Although monopolies make decisions about what quantity to supply, a monopoly does not have a supply curve. A supply curve tells us the quantity that firms choose to supply at any given price. This concept makes sense for competitive firms, which are price takers. But a monopoly is a price maker, not a price taker. It is not meaningful to ask what amount such a firm would produce at any given price because it does not take the price as given. Instead, when the firm chooses the quantity to supply, that decision—along with the demand curve—determines the price.

The monopolist's decision about how much to supply is impossible to separate from the demand curve it faces. The shape of the demand curve determines the shape of the marginal-revenue curve, which in turn determines the monopolist's profit-maximizing quantity. In a competitive market, each firm's supply decisions can be analyzed without knowing the demand curve, but the same is not true in a monopoly market. Therefore, it doesn't make sense to talk about a monopoly's supply curve. ■

At the profit-maximizing quantity, the price equals marginal revenue for a competitive firm but exceeds marginal revenue for a monopoly. That is:

$$\begin{aligned} &\text{For a competitive firm:} \quad P = MR = MC. \\ &\text{For a monopoly firm:} \quad P > MR = MC. \end{aligned}$$

This highlights a key difference between competition and monopoly: **In competitive markets, price equals marginal cost. In monopolized markets, price exceeds marginal cost.** As we will see in a moment, this result is crucial to understanding the social cost of monopoly.

16-2d A Monopoly's Profit

How much profit does a monopoly make? To see a monopoly firm's profit in a graph, recall that profit equals total revenue (TR) minus total costs (TC):

$$\text{Profit} = TR - TC.$$

We can rewrite this as:

$$\text{Profit} = (TR/Q - TC/Q) \times Q.$$

TR/Q is average revenue, which equals the price, P, and TC/Q is average total cost, ATC. Therefore:

$$\text{Profit} = (P - ATC) \times Q.$$

This equation for profit (which also holds for competitive firms) allows us to measure the monopolist's profit in our graph.

Consider the shaded box in Figure 5. The height of the box (the segment BC) is price minus average total cost, $P - ATC$, which is the profit on the typical unit sold. The width of the box (the segment DC) is the quantity sold, Q_{MAX}. The area of this box is the monopoly's total profit.

Table 2 summarizes how a monopoly maximizes profit.

Figure 5

The Monopolist's Profit

The area of the box BCDE equals the profit of the monopoly firm. The height of the box (BC) is price minus average total cost, which equals profit per unit sold. The width of the box (DC) is the number of units sold.

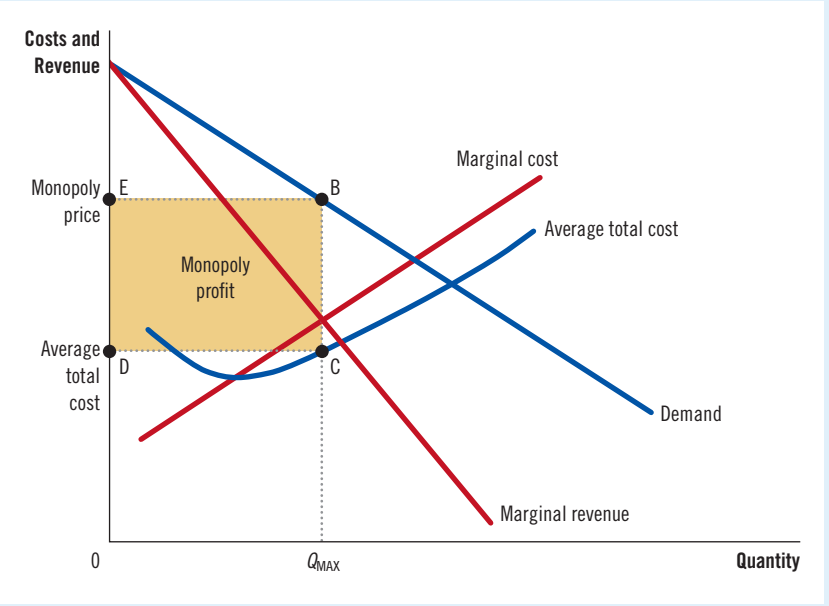


Table 2

Profit-Maximizing Rules for a Monopoly Firm

1. Derive the *MR* curve from the demand curve.
2. Find *Q* at which *MR* = *MC*.
3. On the demand curve, find *P* at which consumers will buy *Q*.
4. If *P* > *ATC*, the monopoly earns a profit.

Case Study

Monopoly Drugs versus Generic Drugs

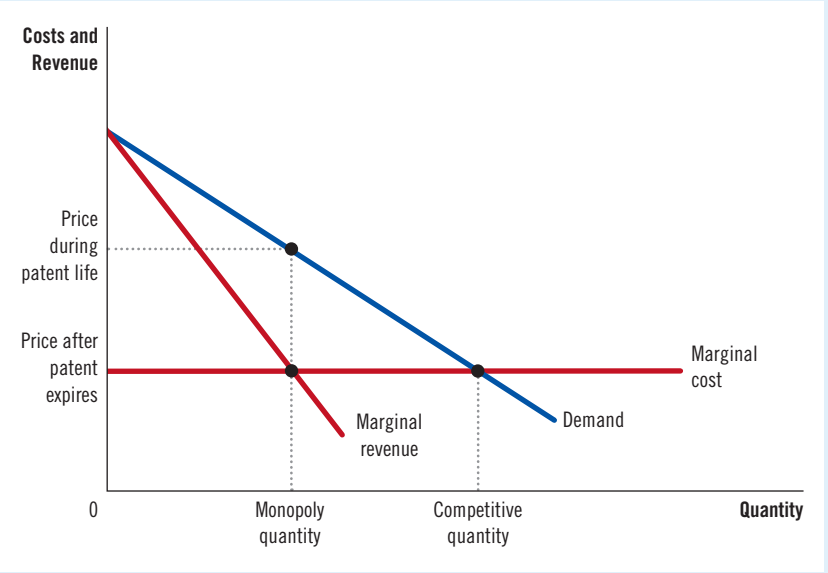
According to our analysis, prices are determined differently in monopolized and competitive markets. A natural place to test this theory is the market for pharmaceutical drugs, which takes on both market structures. When a firm discovers a drug, patent laws give it a monopoly on the sale of that drug. But when the patent expires, any company can make and sell the drug. At that point, the market becomes competitive rather than monopolistic.

What does theory predict will happen to the price of a drug when the patent expires? Consider Figure 6, which shows the market for a typical drug. The marginal cost of producing the drug is assumed to be constant here. (This is roughly true for many drugs.) During the life of the patent, the monopoly maximizes profit by producing the quantity at which marginal revenue equals marginal cost and charging a price well above marginal cost. But when the patent expires, the profit from making the drug encourages new firms to enter the market. With competition, the price should fall to equal marginal cost.

Figure 6

The Market for Drugs

When a patent gives a firm a monopoly over the sale of a drug, the firm charges the monopoly price, which is well above the marginal cost. When the patent on a drug expires and new firms enter, the market becomes competitive, and the price falls to marginal cost.



Experience supports the theory. When the patent on a drug expires, other companies quickly enter and begin selling generics that are chemically identical to the former monopolist's brand-name product. Just as theory predicts, the competitively produced generic drugs are priced well below the price that the monopolist was charging.

The expiration of a patent, however, does not cause the monopolist to lose all its market power. Some consumers remain loyal to the brand-name drug, perhaps out of fear that the new generic drugs are not the same as the drug they have been using for years. As a result, the former monopolist can charge a higher price than its new competitors.

For example, the drug fluoxetine, an antidepressant taken by millions of Americans, was originally sold under the brand name Prozac. Since the patent expired in 2001, consumers have had a choice between the original drug and generic versions. Prozac today sells for much more than generic fluoxetine. This price differential persists because some consumers doubt that the two pills are perfect substitutes. ●

QuickQuiz

3. For a profit-maximizing monopoly that charges a single price, what is the relationship between price P, marginal revenue MR, and marginal cost MC?
   a. $P = MR$ and $MR = MC$.
   b. $P > MR$ and $MR = MC$.
   c. $P = MR$ and $MR > MC$.
   d. $P > MR$ and $MR > MC$.

4. If a monopoly's fixed costs increase, its price will _____, and its profit will _____.
   a. increase; decrease
   b. decrease; increase
   c. increase; stay the same
   d. stay the same; decrease

Answers are at the end of the chapter.

16-3 The Welfare Cost of Monopolies

Is monopoly a good way to organize a market? Unlike a competitive firm, a monopoly charges a price above marginal cost. For consumers, this high price makes monopoly undesirable. But for the owners of the firm, the high price generates more profit and makes monopoly extremely attractive. Is it possible that the benefits to the firm's owners exceed the costs imposed on consumers, making monopoly desirable from the standpoint of society as a whole?

We can answer this question using the tools of welfare economics. Recall from Chapter 7 that total surplus measures the economic well-being of buyers and sellers in a market. Total surplus is the sum of consumer surplus and producer surplus. Consumer surplus is consumers' willingness to pay for a good minus the amount they actually pay for it. Producer surplus is the amount producers receive for a good minus their costs of producing it. In this case, there is a single producer—the monopolist.

You can probably guess the result of this analysis. Chapter 7 concluded that the equilibrium of supply and demand in a competitive market is not only a natural outcome but also a desirable one. The invisible hand of the market leads to an allocation of resources that makes total surplus as large as it can be. Because a monopoly leads to an allocation of resources different from that in a competitive market, the outcome must, in some way, fail to maximize total economic well-being. Let's see why this is the case.

16-3a The Deadweight Loss

Consider what the monopoly firm would do if it were run by a committee of benevolent social planners, a group introduced in Chapter 7. The planners care not only about the firm owners' profit but also about the benefits received by consumers. The planners want to maximize total surplus, which equals producer surplus (profit) plus consumer surplus. Recall that total surplus equals the value of the good to consumers minus the costs of making the good incurred by the monopoly producer.

Figure 7 analyzes how the planners would choose the monopoly's level of output. The demand curve reflects the value of the good to consumers, as measured by their willingness to pay for it. The marginal-cost curve reflects the costs of the monopolist. **The socially efficient quantity is found where the demand curve and the marginal-cost curve intersect.** Below this quantity, the value of an extra unit to consumers exceeds the cost of providing it, so increasing output would raise total surplus. Above this quantity, the cost of producing an extra unit exceeds the value of that unit to consumers, so decreasing output would raise total surplus. At the optimal quantity, the value of an extra unit to consumers exactly equals the marginal cost of production.

If the social planners were running the monopoly, the firm could achieve this efficient outcome by charging the price found at the intersection of the demand and marginal-cost curves. Like a competitive firm and unlike a profit-maximizing monopoly, a social planner would charge a price equal to marginal cost. Because this price would give consumers an accurate signal about the cost of producing the good, consumers would buy the efficient quantity.

We can evaluate the welfare effects of monopoly by comparing the output that the monopolist chooses with the output that a social planner would choose. As we have seen, the monopolist chooses to produce and sell the quantity of output at which the marginal-revenue and marginal-cost curves intersect; the social planner would choose the quantity at which the demand and marginal-cost curves intersect.

Figure 7

The Efficient Level of Output

Social planners maximize total surplus in the market by choosing the level of output where the demand curve and marginal-cost curve intersect. Below this level, the value of the good to the marginal buyer (as reflected in the demand curve) exceeds the marginal cost of making the good. Above this level, the value to the marginal buyer is less than marginal cost.

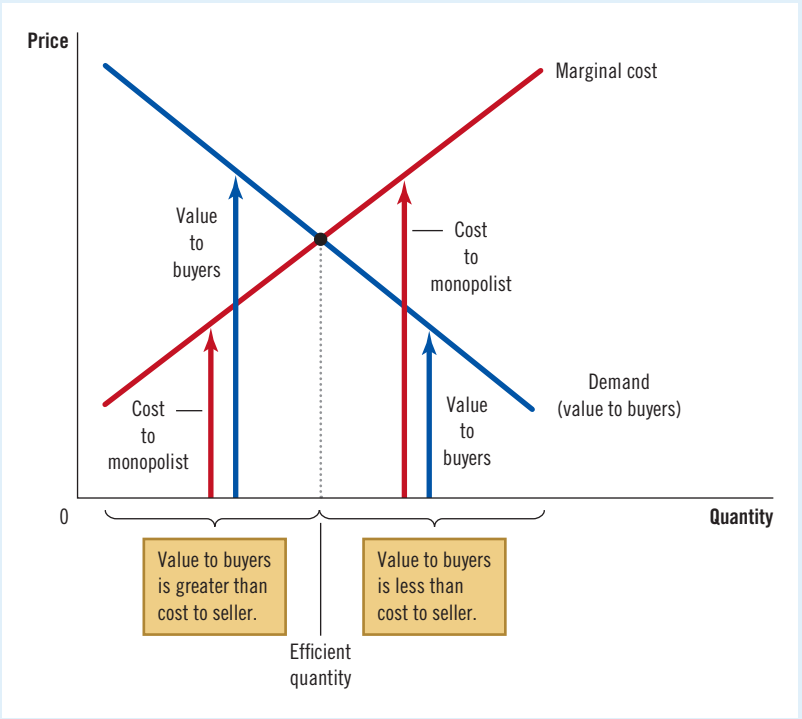


Figure 8 shows the comparison. **The monopolist produces less than the socially efficient quantity of output.**

We can also view the inefficiency of monopoly in terms of the monopolist's price. Because the market demand curve describes a negative relationship between the price and quantity of the good, producing a quantity that is inefficiently low is equivalent to charging a price that is inefficiently high. When a monopolist charges a price above marginal cost, some potential consumers value the good at more than its marginal cost but less than the monopolist's price. These consumers don't buy the good. Because the value they place on the good exceeds the firm's cost of providing it to them, this result is inefficient. Monopoly pricing prevents some mutually beneficial trades from taking place.

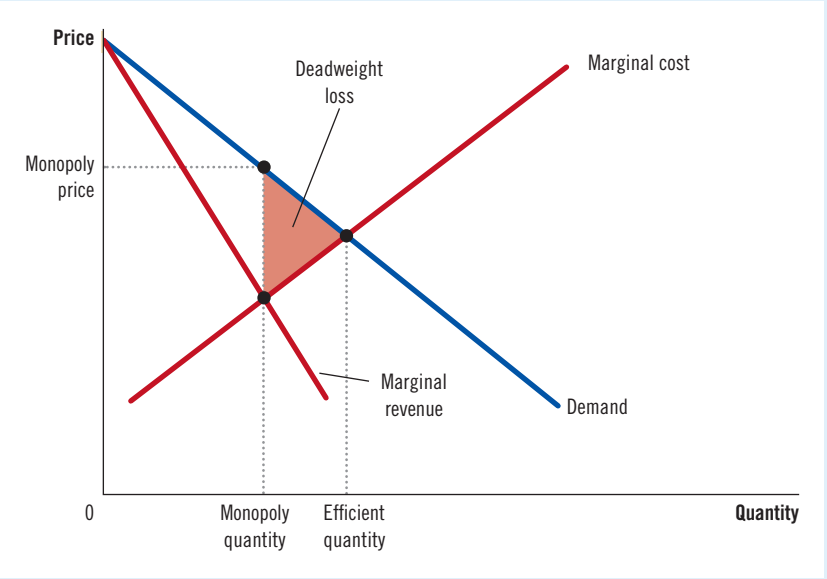
The inefficiency of monopoly can be measured with a deadweight loss triangle, as illustrated in Figure 8. Because the demand curve reflects the value to consumers and the marginal-cost curve reflects the costs to the monopoly producer, the area of the deadweight loss triangle between the demand curve and the marginal-cost curve equals the total surplus lost because of monopoly pricing. It represents the reduction in economic well-being that results from the monopoly's use of its market power.

The deadweight loss caused by a monopoly is similar to the deadweight loss caused by a tax. In a sense, a monopolist is like a private tax collector. As Chapter 8 showed, a tax on a good puts a wedge between consumers' willingness to pay (as reflected by the demand curve) and producers' costs (as reflected by the supply curve). Because a monopoly exerts its market power by charging a price above marginal cost, it creates a similar wedge. In both cases, the wedge causes the quantity sold to fall short of the social optimum. The difference between the two cases is

Figure 8

The Inefficiency of Monopoly

Because a monopoly charges a price above marginal cost, not all consumers who value the good at more than its cost buy it. That means that the quantity produced and sold by a monopoly is below the socially efficient level. The deadweight loss is represented by the area of the triangle between the demand curve (which reflects the value of the good to consumers) and the marginal-cost curve (which reflects the costs of the monopoly producer).



that a tax generates revenue for the government, while a monopoly price generates profit for the firm.

16-3b The Monopoly's Profit: A Social Cost?

It is tempting to decry monopolies for "profiteering" at the public's expense. To be sure, a monopoly earns a profit by virtue of its market power. But according to the economic analysis of monopoly, the firm's profit is not in itself necessarily a problem for society.

Welfare in a monopolized market, as in all markets, includes the welfare of both consumers and producers. When a consumer pays an extra dollar to a producer because of a monopoly price, the consumer is worse off by a dollar, and the producer is better off by the same amount. Because total surplus equals the sum of consumer and producer surplus, this transfer from consumers to the owners of the monopoly does not affect the market's total surplus. In other words, the monopoly profit itself represents not a reduction in the size of the economic pie but merely a bigger slice for producers and a smaller one for consumers. Unless consumers are for some reason more deserving than producers—a normative judgment about equity that goes beyond the realm of economic efficiency—the monopoly profit is not a social problem.

The problem, instead, is that the monopoly firm produces and sells a quantity of output below the level that maximizes total surplus. The deadweight loss measures how much the economic pie shrinks as a result. This inefficiency is connected to the monopoly's high price: Consumers buy fewer units when the firm raises its price above marginal cost. But remember that the profit earned on the units that continue to be sold is not the problem. The problem stems from the inefficiently low quantity of output. Put differently, if the high monopoly price did not discourage some consumers from buying the good, it would raise producer surplus by exactly the amount it reduced consumer surplus, leaving total surplus the same as that achieved by the social planners.

There is, however, a possible exception to this conclusion. Suppose that a monopoly has to incur additional costs to maintain its position as the sole producer in the market. For example, a firm with a government-created monopoly might need to hire lobbyists to convince lawmakers to continue its monopoly. In this case, the monopoly may use up some of its monopoly profits paying for these additional costs. If so, the social loss from monopoly includes both these costs and the deadweight loss resulting from reduced output.

QuickQuiz

5. Compared with the social optimum, a monopoly firm chooses
   a. a quantity that is too low and a price that is too high.
   b. a quantity that is too high and a price that is too low.
   c. a quantity and a price that are both too high.
   d. a quantity and a price that are both too low.

6. The deadweight loss from monopoly arises because
   a. the monopoly makes higher profits than a competitive firm would.
   b. some potential consumers who forgo buying the good value it more than its marginal cost.
   c. consumers who buy the good have to pay more than marginal cost, reducing their consumer surplus.
   d. the monopoly chooses a quantity that fails to equate price and average revenue.

Answers are at the end of the chapter.

16-4 Price Discrimination

price discrimination the business practice of selling the same good at different prices to different customers

So far, we have been assuming that the monopoly firm charges the same price to all customers. Yet in many cases, firms sell the same good to different customers for different prices, even though the costs of producing the good for the two customers are the same. This practice is called **price discrimination**. (Marketing experts sometimes call it **price customization**, perhaps because the word "discrimination" sounds negative. This book uses the standard name.)

Before discussing the behavior of a price-discriminating monopolist, we should note that price discrimination is not possible in a competitive market, where many firms are selling the same good at the market price. No firm is willing to charge a lower price to any customer because it can sell all it wants at the market price. And if any firm tried to charge a higher price to a customer, that customer would buy from another firm. For a firm to price discriminate, it must have some market power.

16-4a A Parable about Pricing

To understand why a monopolist would price discriminate, consider an example. Imagine that you are the president of Readalot Publishing Company. Your best-selling author has just written a new novel. To keep things simple, assume that you pay the author a flat $2 million for exclusive rights to the book and that the cost of printing the book is zero. Readalot's profit, therefore, is the revenue from selling the book minus the $2 million it pays the author. Given these assumptions, how would you, as Readalot's president, decide the book's price?

Your first step is to estimate the demand for the book. Readalot's marketing department tells you that the book will attract two types of readers: 100,000 die-hard fans who are willing to pay as much as $30 and 400,000 less enthusiastic readers who will pay no more than $5.

If Readalot charges everyone the same price, what price maximizes profit? There are two natural prices to consider: $30 is the most Readalot can charge and attract the 100,000 die-hard fans, and $5 is the highest price that will attract the entire market of 500,000 potential readers. Solving this problem is a matter of simple arithmetic. At $30, Readalot sells 100,000 copies, has revenue of $3 million, and makes $1 million in profit. At $5, it sells 500,000 copies, has revenue of $2.5 million, and makes $500,000. The profit-maximizing strategy is to charge $30 and say goodbye to the 400,000 less enthusiastic readers.

Readalot's decision causes a deadweight loss. There are 400,000 readers willing to pay $5 for the book, and the marginal cost of providing it to them is zero. Society loses $2 million of total surplus when Readalot charges the higher price. This deadweight loss is the inefficiency that arises whenever a monopolist charges a price above marginal cost.

Now suppose that Readalot's marketing department makes a discovery: The Pacific Ocean separates these two groups of readers. The die-hard fans live in Australia, the less hard-core readers live in the United States, and readers in one country can't easily buy books in the other.

Eureka! Readalot quickly changes its marketing strategy. It charges the 100,000 Australian readers $30 for the book, while asking only $5 of the 400,000 American readers. Now, revenue is $3 million in Australia and $2 million in the United States, for a total of $5 million. Profit is $3 million, a lot more than the $1 million the company could earn by charging the same $30 price to all customers. As Readalot's president, you embrace this strategy of price discrimination.

The story of Readalot Publishing is hypothetical, but it describes the business practice of many companies. Consider the prices of hardcover books, e-books, and paperbacks. When a publisher has a new novel, it initially releases an expensive hardcover edition and an e-book, usually at a lower price. For readers who prefer print but won't pay the high price of a hardcover, it later releases a cheaper paperback edition. The price differences among these various editions far exceed the differences in marginal production costs. The publisher is price discriminating by selling the hardcover to die-hard fans, the e-book to those who prefer the lower cost and don't mind reading on a tablet, and the paperback to price-sensitive print readers, thereby maximizing its profit.

16-4b The Moral of the Story

Like any parable, the story of Readalot Publishing is stylized, yet it contains some important truths. This parable teaches three lessons about price discrimination.

The first is that price discrimination is a rational strategy for a profit-maximizing monopolist. By charging different prices to different customers, a monopolist can increase its profit. In essence, a price-discriminating monopolist charges prices closer to each customer's willingness to pay than is possible with a single price.

The second lesson is that for price discrimination to work, the seller must be able to separate customers according to their willingness to pay. In the Readalot story, customers were separated geographically. But monopolists can use other differences, such as age or income, to distinguish among customers.

A corollary to this second lesson is that certain market forces can prevent firms from price discriminating. One such force is **arbitrage**, the process of buying a good in one market at a low price and selling it in another market at a higher price to profit from the price difference. Suppose that in our example, Australian bookstores could buy the book in the United States and resell it to Australian readers. This arbitrage

would prevent Readalot from price discriminating because no Australian would buy the book at the higher price.

The third lesson may be the most surprising: Price discrimination can raise welfare as measured by total surplus. Recall that a deadweight loss arises when Readalot charges a single $30 price because the 400,000 less enthusiastic readers don't get the book, even though they value it at more than its marginal cost of production. But when Readalot price discriminates, all readers buy the book, and the outcome is efficient. Thus, price discrimination can eliminate the inefficiency inherent in monopoly pricing.

Note that in this example, the increase in welfare from price discrimination shows up as higher producer surplus rather than higher consumer surplus. Consumers are no better off for having bought the book: The price they pay exactly equals the value they place on the book, so they receive no consumer surplus. The entire increase in total surplus from price discrimination accrues to Readalot Publishing in the form of higher profit.

16-4c The Analytics of Price Discrimination

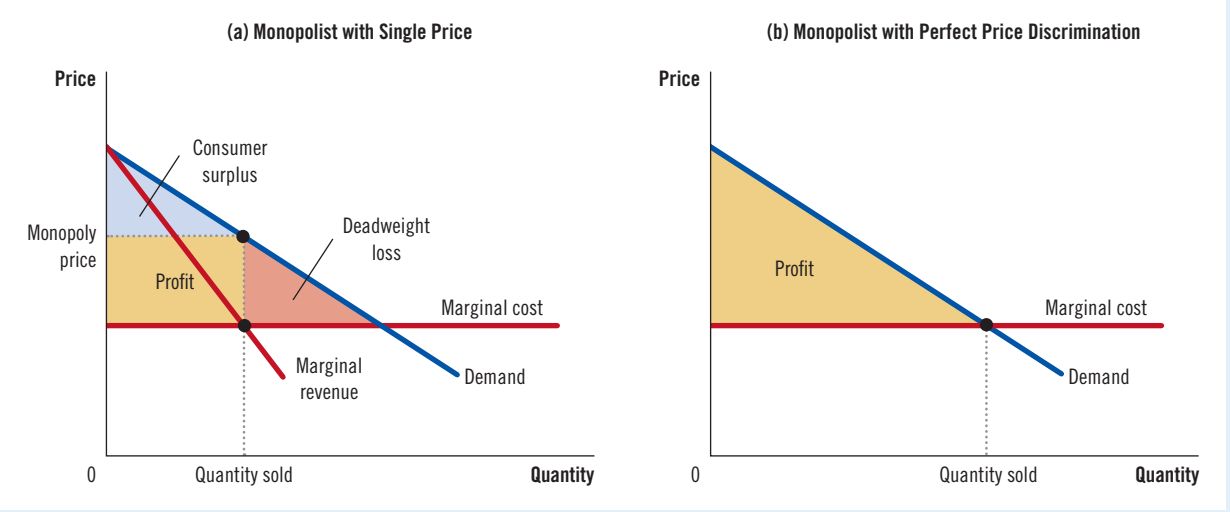
Let's consider a bit more formally how price discrimination affects welfare. We begin by assuming that the monopolist can price discriminate perfectly. **Perfect price discrimination** describes a situation in which the monopolist knows exactly each customer's willingness to pay and can charge each customer a different price. In this case, the monopolist charges customers exactly their willingness to pay, and the monopolist gets the entire surplus in every transaction.

Figure 9 shows producer and consumer surplus with and without price discrimination. To keep things simple, this figure is drawn assuming constant unit costs—that is, marginal cost and average total cost are constant and equal. Without price discrimination, the firm charges a single price above marginal cost, as in

Figure 9

Welfare with and without Price Discrimination

Panel (a) shows a monopoly that charges the same price to all customers. Total surplus in this market equals the sum of profit (producer surplus) and consumer surplus. Panel (b) shows a monopoly that can price discriminate perfectly. Because consumer surplus equals zero, total surplus now equals the firm's profit. Comparing these two panels, you can see that perfect price discrimination raises profit, raises total surplus, and lowers consumer surplus.



panel (a). Because some potential customers who value the good at more than marginal cost do not buy it at this high price, the monopoly causes a deadweight loss. Yet when a firm can price discriminate perfectly, as in panel (b), all customers who value the good at more than marginal cost buy it and are charged their willingness to pay. Every mutually beneficial trade takes place, no deadweight loss occurs, and the entire surplus derived from the market goes to the monopoly producer in the form of profit.

In reality, of course, price discrimination is not perfect. Customers do not walk into stores with signs displaying their willingness to pay. Instead, firms price discriminate by dividing customers into groups: young versus old, weekday versus weekend shoppers, Americans versus Australians, print versus e-book readers, and so on. Unlike those in the parable of Readalot Publishing, customers within each group differ in their willingness to pay for the product, making perfect price discrimination impossible.

How does this imperfect price discrimination affect welfare as measured by total surplus? The analysis of these pricing schemes is complicated, and it turns out that there is no general answer to this question. Compared with the single-price monopoly outcome, imperfect price discrimination can raise, lower, or leave unchanged the total surplus in a market. The only certain conclusion is that price discrimination raises the monopoly's profit; otherwise, the firm would choose to charge all customers the same price.

16-4d Examples of Price Discrimination

Firms in the economy use various business strategies to charge different prices to different customers. Here are some examples.

Movie Tickets Many movie theaters charge a lower price for children and people over 65 than for other patrons. This is hard to explain in a competitive market where price equals marginal cost because the marginal cost of providing a seat is the same for anyone of any age. Yet the differential pricing is easily explained if movie theaters have some local monopoly power and if children and older people have a lower willingness to pay for a ticket. In this case, movie theaters increase their profit by price discriminating.

Airline Prices Seats on airplanes are sold at many different prices. Most airlines charge a lower price for a round-trip ticket between two cities if the traveler stays over a Saturday night. At first, this seems odd. Why should it matter to the airline whether a passenger stays over a Saturday night? The reason is that this rule provides a way to separate business and leisure travelers. Passengers on business trips have a high willingness to pay and, most likely, do not want to stay over a Saturday night because business meetings are rarely held during weekends. But passengers on vacation or visiting friends and family have a lower willingness to pay and are more likely to want to spend the weekend at their destinations. For airlines, charging lower prices to passengers who stay over a Saturday night is successful price discrimination.

LUCY HAMILTON

"Would it bother you to hear how little I paid for this flight?"

Discount Opportunities Many companies offer discount coupons online and in newspapers and magazines. And some vendors offer online savings on special days that occur so frequently that they aren't really very special. To get a discount, a buyer simply has to clip a coupon or buy on the right day. Why do companies bother? Why don't they just cut the price of the product in the first place?

The answer is that these strategies allow companies to price discriminate. Companies know that not all customers are willing to spend time clipping

coupons or navigating online bargains. Moreover, the willingness to seek bargains is related to customers' willingness to pay for the good. Rich and busy executives are unlikely to spend their time doing so, and they are probably willing to pay a higher price for many goods. The unemployed are more likely to hunt for cheaper deals and have a lower willingness to pay. By charging a lower price only to those customers willing to spend the time to seek discounts, firms can successfully price discriminate.

Financial Aid Many colleges and universities give financial aid based on family income. One can view this policy as a type of price discrimination. Students from wealthy families have greater financial resources and, therefore, a higher willingness to pay than students from lower-income families. By charging high tuition and selectively offering financial aid, schools in effect charge prices to customers based on the value they place on going to that school. This behavior is like that of any price-discriminating monopolist.

Quantity Discounts So far in these examples of price discrimination, the monopolist charges different prices to different customers. Sometimes, however, monopolists price discriminate by charging different prices to the same customer for different units that the customer buys. For example, many firms offer lower prices to customers who buy large quantities. A bakery might charge $0.50 for each donut but $5 for a dozen. This is a form of price discrimination because customers pay a higher price for the first unit they buy than for the twelfth. Quantity discounts are often a successful way to price discriminate because customers' willingness to pay for an additional unit declines as they buy more units.

QuickQuiz

7. Price discrimination by a monopolist refers to charging different prices based on
   a. the consumer's willingness to pay.
   b. the consumer's racial or ethnic group.
   c. the cost of producing the good for a particular consumer.
   d. whether the consumer is likely to become a repeat buyer.

8. When a monopolist switches from charging a single price to practicing perfect price discrimination, it reduces
   a. the quantity produced.
   b. the firm's profit.
   c. consumer surplus.
   d. total surplus.

Answers are at the end of the chapter.

16-5 Public Policy toward Monopolies

Unlike competitive markets, monopolies fail to allocate resources efficiently. They produce less than the socially desirable quantity of output and charge prices above marginal cost. Government policymakers can deal with the problem of monopoly in several ways:

- By trying to make monopolized industries more competitive
- By regulating the behavior of the monopolies
- By turning some private monopolies into public enterprises
- By doing nothing at all

16-5a Increasing Competition with Antitrust Laws

SCIENCECARTOONSPLUS.COM

"But if we do merge with Amalgamated, we'll have enough resources to fight the antitrust violation caused by the merger."

If Coca-Cola and PepsiCo wanted to merge, the federal government would scrutinize the deal before it went into effect. The lawyers and economists in the Department of Justice might well decide that a merger between these two large soft-drink companies would make the U.S. soft-drink market substantially less competitive and, as a result, would reduce the well-being of the country as a whole. If so, the Department of Justice would challenge the merger in court, and, if the judge agreed, the two companies would not be allowed to merge. Traditionally, the courts are especially wary of **horizontal mergers**, those between two firms in the same market, like Coca-Cola and PepsiCo. They are less likely to block **vertical mergers**, those between firms at different stages of the production process, such as a merger between a tire company and an auto company. In other words, if a company wants to merge with a competitor, it will face closer scrutiny than if it wants to merge with one of its suppliers or one of its customers.

The government derives this power over private industry from **antitrust laws**, statutes aimed at curbing monopoly power. In the United States, the first and most important of these laws was the Sherman Antitrust Act, passed in 1890 to reduce the market power of "trusts," the dominant monopolies of that era. The Clayton Antitrust Act, passed in 1914, strengthened the government's powers and authorized private lawsuits. As the U.S. Supreme Court put it, the antitrust laws are "a comprehensive charter of economic liberty aimed at preserving free and unfettered competition as the rule of trade."

The antitrust laws give the government tools to promote competition. They allow the government to prevent mergers and, at times, to break up large companies. The antitrust laws also prevent companies from colluding to reduce competition.

Stopping mergers and breaking up companies can have costs as well as benefits. Sometimes, companies combine to lower costs through more efficient joint production. These advantages are called **synergies**. For example, many U.S. banks have merged in recent years to reduce administrative expenses. The airline industry has also experienced consolidation. If antitrust laws are to raise social welfare, the government must be able to determine which mergers are desirable and which are not. That is, it must measure and compare the social benefit from synergies with the social costs of reduced competition. It is open to debate whether the government can perform the necessary cost–benefit analysis with sufficient accuracy. In the end, the application of antitrust laws is often controversial, even among the experts.

16-5b Regulation

Another way the government deals with the problem of monopoly is by regulating monopolists' behavior. This solution is common for natural monopolies, such as water and electric companies, whose prices are often regulated by the government.

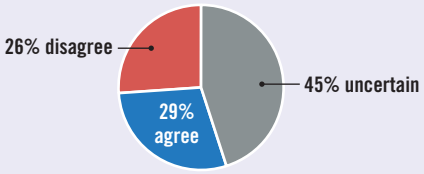
What price should the government set for a natural monopoly? This question is not as easy as it might at first appear. One might conclude that the price should equal the monopolist's marginal cost. If price equals marginal cost, customers will buy the quantity of the monopolist's output that maximizes total surplus, and the allocation of resources will be efficient.

Yet two practical problems arise with marginal-cost pricing as a regulatory system. The first arises from the logic of cost curves. By definition, natural monopolies have declining average total cost. As we discussed in a previous chapter, when average total cost is declining, marginal cost is less than average total cost. This situation is

Ask the Experts: Mergers and Competition

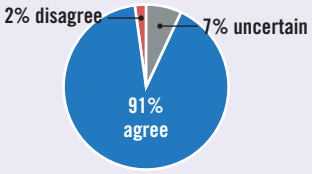
"If regulators had not approved mergers in the past decade between major networked airlines, travelers would be better off today."

What do economists say?



"Americans pay too much for broadband, cable television, and telecommunications services, in part because of a lack of adequate competition."

What do economists say?



Source: IGM Economic Experts Panel, August 28, 2013, July 20, 2021.

illustrated in Figure 10, which shows a firm with a large fixed cost and then a constant marginal cost thereafter. If regulators were to set price equal to marginal cost, that price would be less than the firm's average total cost, and the firm would lose money. Instead of charging such a low price, the monopoly firm would just exit the industry.

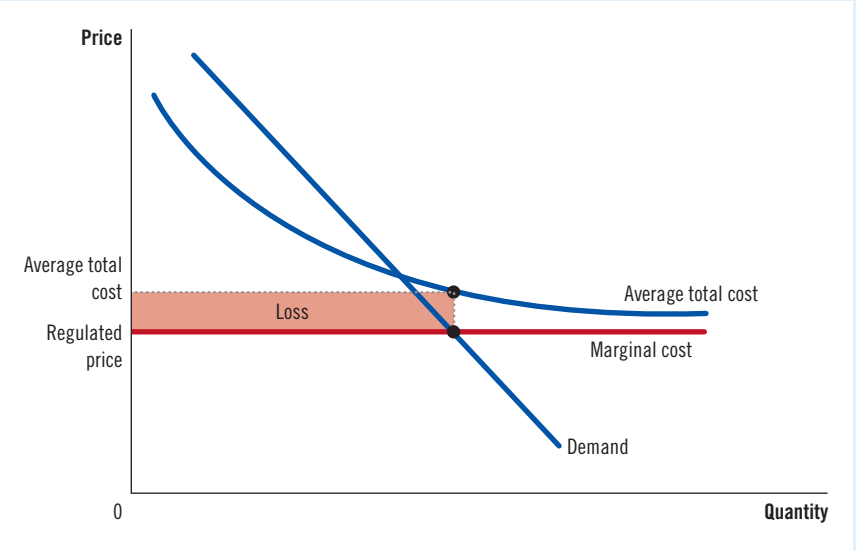
Regulators can respond to this problem in various ways, none of which is perfect. One way is to subsidize the monopolist. In essence, the government picks up the losses inherent in marginal-cost pricing. Yet to finance the subsidy, the government needs to raise money through taxation, which itself generates deadweight losses. Alternatively, the regulators can allow the monopolist to charge a price higher than marginal cost. If the regulated price equals average total cost, the monopolist earns exactly zero economic profit. Yet average-cost pricing leads to deadweight losses because the monopolist's price no longer reflects the marginal cost of producing the good. In essence, average-cost pricing is like a tax on the good the monopolist is selling.

The second problem with marginal-cost pricing as a regulatory system (and with average-cost pricing as well) is that it gives the monopolist no incentive to reduce costs. Each firm in a competitive market tries to reduce its costs because lower costs mean higher profits. But if a regulated monopolist knows that regulators will reduce prices whenever costs fall, the monopolist won't benefit from lower costs. In practice, regulators deal with this problem by allowing monopolists to keep some of the benefits from lower costs in the form of higher profit, a practice that requires some departure from marginal-cost pricing.

Figure 10

Marginal-Cost Pricing for a Natural Monopoly

Because a natural monopoly has declining average total cost, marginal cost is less than average total cost. Therefore, if regulators require a natural monopoly to charge a price equal to marginal cost, the price will be below average total cost, and the monopoly will lose money.



16-5c Public Ownership

The third policy for dealing with monopoly is public ownership. That is, rather than regulating a natural monopoly run by a private firm, a government unit can run the monopoly itself. This solution is common in European countries, where the government owns and operates utilities such as telephone, water, and electric companies. It's fairly common in the United States, too. The government runs the Postal Service, which is often considered to be a natural monopoly. And there are many publicly owned water and power utilities throughout the country.

While public ownership of natural monopolies often has popular support, many economists prefer private to public ownership. The key issue is how the ownership affects the costs of production. Private owners have an incentive to minimize costs as long as they reap part of the benefit in the form of higher profit. If the managers do a bad job of keeping costs down, the firm's owners will fire them. But if public employees do a bad job, the losers are the customers and taxpayers, whose only recourse is the political system. Public employees may become a special-interest group and seek to bend the political system to their advantage. Put simply, as a way of ensuring that firms are efficiently run, the voting booth is less reliable than the profit motive.

16-5d Above All, Do No Harm

Each of the foregoing policies aimed at reducing the problem of monopoly has drawbacks. As a result, some economists argue that the government should be careful not to make matters worse when dealing with monopoly pricing. Here is the assessment of George Stigler, who won the Nobel Prize for his work in industrial organization.

> A famous theorem in economics states that a competitive enterprise economy will produce the largest possible income from a given stock of resources. No real economy meets the exact conditions of the theorem, and all real economies will fall short of the ideal economy—a difference called "market failure." In my view, however, the degree of "market failure" for the American economy is much smaller than the "political failure" arising from the imperfections of economic policies found in real political systems.

As this quotation suggests, the political world is sometimes even less perfect than a highly imperfect market. An institute at the University of Chicago named after Stigler continues to analyze the uneasy relationship between the power of monopolies and the political system. The best solution isn't always clear. In some, though not all, cases, it may be wiser to do nothing.

QuickQuiz

9. Antitrust regulators are likely to prohibit two firms from merging if
   a. there are many other firms in the industry.
   b. there are sizable synergies to the combination.
   c. the combined firm will have a large share of the market.
   d. the combined firm will undercut competitors with lower prices.

10. If regulators impose marginal-cost pricing on a natural monopoly, a possible problem is that
    a. consumers will buy more of the good than is efficient.
    b. consumers will buy less of the good than is efficient.
    c. the firm will lose money and exit the market.
    d. the firm will make excessive profits.

Answers are at the end of the chapter.

In the News

Will the Biden Administration Expand the Scope of Antitrust Policy?

An ongoing debate is whether antitrust regulators should expand their focus from just the economic impact of large businesses to the political and cultural impact as well.

Antitrust's New Mission: Preserving Democracy, Not Efficiency

By Greg Ip

Amazon.com Inc.'s proposed acquisition of film studio MGM would ordinarily provoke little antitrust concern. MGM's share of box office receipts is tiny and Amazon's entertainment footprint is relatively small.

But Amazon of course does much more than make movies: it is the country's largest e-commerce and cloud-computing company, and a dominant seller of books, videos and music. Founder and Executive Chairman Jeff Bezos owns the Washington Post. In their totality, then, Amazon and Mr. Bezos represent a significant concentration of economic and cultural influence. For a new generation of trustbusters, that's a bigger concern than the efficiency benefits a merger might bring.

"There is sound reason to ask whether permitting Amazon to leverage its platform to integrate across business lines hands it undue economic and political power," then law student Lina Khan wrote in a now famous law journal article in 2017.

Last month, President Biden named the 32-year-old Ms. Khan chairwoman of the Federal Trade Commission. At a confirmation hearing in April, she said antitrust's historical role is to "protect our economy and our democracy from unchecked monopoly power."

Ms. Khan embodies the neo-Brandeisian movement, named for Louis Brandeis, a crusading lawyer and later Supreme Court justice who argued bigness was both inefficient and antithetical to liberty. "Size, we are told, is not a crime," he wrote in 1914. "But size may, at least, become noxious by reason of the means through which it was attained or the uses to which it is put."

For Brandeis, democracy included the freedom of a worker to negotiate with their employer, a supplier with a retailer and vice versa, and a farmer with his bank, which required a plurality of market participants. "He feared that as the corporations became large and powerful, they took on a life of their own, becoming increasingly insensitive to humanity's wants and fears," Columbia University law professor Tim Wu wrote in a 2018 book. Mr. Wu now serves on Mr. Biden's National Economic Council.

For decades that's how courts and regulators interpreted antitrust laws, regularly ruling against mergers and business practices such as exclusivity deals between suppliers and customers.

In the 1970s that approach came under attack by conservatives led by the late legal scholar Robert Bork, who had served under Richard Nixon. "The only legitimate goal of American antitrust law is the maximization of consumer welfare," he wrote in his 1978 book "The Antitrust Paradox." Companies got big by becoming more efficient, which benefited consumers, he wrote. Judges and regulators, lacking even a "rudimentary understanding of market economics," deprived consumers of these efficiency benefits when they blocked mergers in pursuit of a "grab bag" of political and social objectives.

The consumer welfare standard has come to govern antitrust. But in the last decade rising economic concentration, high profit margins, declining business startups and subdued investment all suggest monopoly power is growing again.

Meanwhile, Mr. Biden's election has coincided with a backlash against the pro-market principles that have governed economic policy for decades. Progressives blame the obsession with efficiency for aggravating inequality and racial disparities; conservative populists blame free trade for hollowing out manufacturing.

Big Tech is ground zero for this backlash. One or two companies now dominate social

16-6 Conclusion: The Prevalence of Monopolies

This chapter has discussed the behavior of firms that have control over the prices they charge. We have seen that such firms behave very differently from the competitive firms studied in the previous chapter. Table 3 summarizes some of the key similarities and differences between competitive and monopoly markets.

An important finding for public policy is that a monopolist produces less than the efficient quantity and charges a price above marginal cost. As a result, a monopoly causes deadweight losses. Price discrimination can sometimes mitigate these inefficiencies. But other times, they call for policymakers to take an active role.

media, smartphone app stores, Internet search, web advertising and electronic commerce. Judged only by consumer welfare, this doesn't seem to present a problem: their products are cheap or free, and extremely popular. Judged by concentrations of power, it's problematic: their control of essential platforms leave individual merchants, app developers, content providers and users next to no bargaining power since they have so few alternatives. Barriers to entry are high to insurmountable for potential competitors. They determine what artistic and political content billions of users share and see.

Before becoming FTC chair Ms. Khan had advocated either barring platform operators like Amazon from competing with users of that platform; or regulating them like utilities. She would ban vertical mergers—that is, between two components of the same supply chain, for example a supplier of content such as MGM and a distributor such as Amazon—once a platform operator has reached a dominant size. She hasn't commented on Amazon since becoming chair. Nonetheless, citing these views, Amazon has asked Ms. Khan to recuse herself from FTC investigations of the company.

Under Ms. Khan's vision, the risk that a business structure could enable anticompetitive behavior matters more than evidence of that behavior such as higher prices.

But that approach carries risks of its own. Lack of evidence could lead to weak cases that fail in court. A federal judge dismissed a lawsuit by the FTC and most state attorneys-general for failing to establish that Facebook Inc. is a monopoly. Bills proposed by Democrats in the House of Representatives would lower the bar for winning such suits, but their fate is unclear. Barring firms from getting big could deprive consumers of the benefits that only big firms can deliver. Millions relied on Amazon's heft when the pandemic kept them out of stores.

And while neo-Brandeisians worry about abuse of corporate power, Mr. Bork's acolytes worry about abuse of antitrust authority. In 2018, the Department of Justice tried, unsuccessfully, to stop AT&T Inc. from acquiring Time Warner Inc., in a move many saw as motivated by then President Donald Trump's personal animosity toward CNN, a unit of Time Warner.

For all its flaws, antitrust governed by the consumer welfare standard is less at risk of politicization than beliefs about what's good or bad for democracy. ■

SIPA USA/ALAMY STOCK PHOTO

FTC chair Lina Khan

Questions to Discuss

1. Do you think policymakers should be concerned with more than the economic impact of large businesses? Why or why not?
2. What risks do you see in expanding the scope of antitrust laws?

Source: *The Wall Street Journal*, July 8, 2021.

How prevalent are the problems of monopoly?

In one sense, monopolies are common. Most firms have some control over the prices they charge. They are not forced to charge the market price for their goods because their goods are not exactly the same as those offered by other firms. A Tesla is not the same as an electric Mustang. Ben and Jerry's ice cream is not the same as Breyer's. Each of these goods has a downward-sloping demand curve, which gives the producer some degree of monopoly power.

Yet firms with substantial monopoly power are rare. Few goods are truly unique. Most have substitutes that, even if not exactly the same, are similar. Ben and Jerry can raise the price of their ice cream a little without losing all their sales,

Table 3

Competition versus Monopoly: A Summary Comparison

	Competition	Monopoly
Similarities		
Goal of firms	Maximize profits	Maximize profits
Rule for maximizing	$MR = MC$	$MR = MC$
Can earn economic profits in the short run?	Yes	Yes
Differences		
Number of firms	Many	One
Marginal revenue	$MR = P$	$MR < P$
Price	$P = MC$	$P > MC$
Produces welfare-maximizing level of output?	Yes	No
Entry in the long run?	Yes	No
Can earn economic profits in the long run?	No	Yes
Price discrimination possible?	No	Yes

but if they raise it a lot, sales will fall substantially as their customers switch to other brands.

In the end, monopoly power is a matter of degree. It is true that many firms have some monopoly power. Still, for most firms, their monopoly power is limited. In many situations, we will not go far wrong assuming that firms operate in competitive markets, even if that is not precisely the case.

Chapter in a Nutshell

- A monopoly is the sole seller in its market. A monopoly arises when a single firm owns a key resource, when the government gives a firm the exclusive right to produce a good, or when a single firm can supply the entire market at a lower cost than many firms could.
- Because a monopoly is the sole producer in its market, it faces a downward-sloping demand curve for its product. When a monopoly increases production by 1 unit, it causes the price of its good to fall, which reduces the amount of revenue earned on all units produced. As a result, a monopoly's marginal revenue is always less than the price of its good.
- Like a competitive firm, a monopoly maximizes profit by producing the quantity at which marginal revenue equals marginal cost. The monopoly then sets the price at which consumers demand that quantity. Unlike a competitive firm, a monopoly's price exceeds its marginal revenue, so its price exceeds marginal cost.
- A monopolist's profit-maximizing output is below the level that maximizes the sum of consumer and producer surplus. That is, when the monopoly charges a price above marginal cost, some consumers who value the good more than its cost of production don't buy it. As a result, monopoly leads to deadweight losses similar to those that arise from taxes.

- A monopolist can often increase profits by charging different prices for the same good based on a buyer's willingness to pay. This practice of price discrimination can raise economic welfare by getting the good to some consumers who would otherwise not buy it. In the extreme case of perfect price discrimination, the deadweight loss of monopoly is eliminated, and the entire surplus in the market goes to the monopoly producer. More generally, when price discrimination is imperfect, it can either raise or lower welfare compared with the outcome from a single monopoly price.
- Policymakers can respond to the inefficiency of monopoly behavior in several ways. They can use the antitrust laws to try to make the industry more competitive. They can regulate the prices that the monopoly charges. They can turn the monopolist into a government-run enterprise. Or, if the market failure is deemed small compared with the inevitable imperfections of policies, they can do nothing at all.

Key Concepts

monopoly, p. 312

natural monopoly, p. 314

price discrimination, p. 326

Questions for Review

1. Give an example of a government-created monopoly. Is creating this monopoly necessarily bad public policy? Explain.
2. Define **natural monopoly**. What does the size of a market have to do with whether an industry is a natural monopoly?
3. Why is a monopolist's marginal revenue less than the price of its good? Can marginal revenue ever be negative? Explain.
4. Draw the demand, marginal-revenue, average-total-cost, and marginal-cost curves for a monopolist. Show the profit-maximizing output, the profit-maximizing price, and the amount of profit.
5. In your diagram from the previous question, show the output that maximizes total surplus. Show the deadweight loss from the monopoly. Explain your answer.
6. Give two examples of price discrimination. In each case, explain why the monopolist chooses to follow this business strategy.
7. What gives the government the power to regulate mergers between firms? From the perspective of society's welfare, give one reason that a merger might be good and one reason that a merger might be bad.
8. Describe the two problems that arise when regulators tell a natural monopoly that it must set a price equal to marginal cost.

Problems and Applications

1. A publisher faces the following demand schedule for the next novel from one of its popular authors:

Price	Quantity Demanded
$100	0 novels
90	100,000
80	200,000
70	300,000
60	400,000
50	500,000
40	600,000
30	700,000
20	800,000
10	900,000
0	1,000,000

The author is paid $2 million to write the book, and the marginal cost of publishing the book is a constant $10 per book.

a. Compute total revenue, total cost, and profit at each quantity. What quantity would a profit-maximizing publisher choose? What price would it charge?
b. Compute marginal revenue. (Recall that $MR = \Delta TR/\Delta Q$.) How does marginal revenue compare with the price? Explain.
c. Graph the marginal-revenue, marginal-cost, and demand curves. At what quantity do the marginal-revenue and marginal-cost curves cross? What does this signify?
d. In your graph, shade in the deadweight loss. Explain in words what this means.
e. If the author were paid $3 million instead of $2 million to write the book, how would this affect the publisher's decision regarding what price to charge? Explain.
f. Suppose the publisher was not profit-maximizing but was instead concerned with maximizing economic efficiency. What price would it charge for the book? How much profit would it make at this price?

2. A small town is served by many competing supermarkets, which have the same constant marginal costs.
   a. Using a diagram of the market for groceries, show the consumer surplus, producer surplus, and total surplus.
   b. Now suppose that the independent supermarkets combine into one chain. Using a new diagram, show the new consumer surplus, producer surplus, and total surplus. Relative to the competitive market, what is the transfer from consumers to producers? What is the deadweight loss?

3. Taylor Swift has just finished recording her latest album. Her recording company determines that the demand for the CD is as follows:

Price	Number of CDs
$24	10,000
22	20,000
20	30,000
18	40,000
16	50,000
14	60,000

The company can produce the CD with no fixed cost and a variable cost of $5 per CD.
   a. Find total revenue for quantity equal to 10,000, 20,000, and so on. What is the marginal revenue for each 10,000 increase in the quantity sold?
   b. What quantity of CDs would maximize profit? What would the price be? What would the profit be?
   c. If you were Swift's agent, what recording fee would you advise her to demand from the recording company? Why?

4. A company is considering building a bridge across a river. The bridge would cost $2 million to build and nothing to maintain. The following table shows the company's anticipated demand over the lifetime of the bridge:

Price per Crossing	Number of Crossings, in Thousands
$8	0
7	100
6	200
5	300
4	400
3	500
2	600
1	700
0	800

   a. If the company were to build the bridge, what would be its profit-maximizing price? Would that level of output be efficient? Why or why not?
   b. If the company is interested in maximizing profit, should it build the bridge? What would be its profit or loss?
   c. If the government were to build the bridge, what price should it charge?
   d. Should the government build the bridge? Explain.

5. Consider the relationship between monopoly pricing and price elasticity of demand.
   a. Explain why a monopolist will never produce a quantity at which the demand curve is inelastic. (Hint: If demand is inelastic and the firm raises its price, what happens to total revenue and total costs?)
   b. Draw a diagram for a monopolist, precisely labeling the portion of the demand curve that is inelastic. (Hint: The answer is related to the marginal-revenue curve.)
   c. On your diagram, show the quantity and price that maximize total revenue.

6. You live in a town with 300 adults and 200 children, and you are thinking about putting on a play to entertain your neighbors and make some money. A play has a fixed cost of $2,000, but selling an extra ticket has zero marginal cost. Here are the demand schedules for your two types of customer:

Price	Adults	Children
$10	0	0
9	100	0
8	200	0
7	300	0
6	300	0
5	300	100
4	300	200
3	300	200
2	300	200
1	300	200
0	300	200

a. To maximize profit, what price would you charge for an adult ticket? For a child's ticket? How much profit do you make?
b. The city council passes a law prohibiting you from charging different prices to different customers. What price do you set for a ticket now? How much profit do you make?
c. Who is worse off because of the law prohibiting price discrimination? Who is better off? (If you can, quantify the changes in welfare.)
d. If the fixed cost of the play were $2,500 rather than $2,000, how would your answers to parts (a), (b), and (c) change?

7. The residents of the town of Ectenia all love economics, and the mayor proposes building an economics museum. The museum has a fixed cost of $2,400,000 and no variable costs. There are 100,000 town residents, and each has the same demand for museum visits: $Q^D = 10 - P$, where P is the price of admission.
   a. Graph the museum's average-total-cost curve and its marginal-cost curve. What kind of market would describe the museum?
   b. The mayor proposes financing the museum with a lump-sum tax of $24 and then opening the museum to the public for free. How many times would each person visit? Calculate the benefit each person would get from the museum, measured as consumer surplus minus the new tax.
   c. The mayor's anti-tax opponent says the museum should finance itself by charging an admission fee. What is the lowest price the museum can charge without incurring losses? (Hint: Find the number of visits and museum profits for prices of $2, $3, $4, and $5.)
   d. For the break-even price you found in part (c), calculate each resident's consumer surplus. Compared with the mayor's plan, who is better off with this admission fee, and who is worse off? Explain.
   e. What real-world considerations absent in the problem above might justify an admission fee?

8. Henry Potter owns the only well in town that produces clean drinking water. He faces the following demand, marginal-revenue, and marginal-cost curves:

$$\text{Demand: } P = 70 - Q$$
$$\text{Marginal Revenue: } MR = 70 - 2Q$$
$$\text{Marginal Cost: } MC = 10 + Q$$

   a. Graph these three curves. If Mr. Potter maximizes profit, what quantity does he produce? What price does he charge? Show these results on your graph.
   b. Mayor George Bailey, concerned about water consumers, is considering a price ceiling 10 percent below the monopoly price derived in part (a). What quantity would be demanded at this new price? Would the profit-maximizing Mr. Potter produce that amount? Explain. (Hint: Think about marginal cost.)
   c. George's Uncle Billy says that a price ceiling is a bad idea because price ceilings cause shortages. Is he right in this case? What size shortage would the price ceiling create? Explain.
   d. George's friend Clarence, who is even more concerned about consumers, suggests a price ceiling 50 percent below the monopoly price. What quantity would be demanded at this price? How much would Mr. Potter produce? In this case, is Uncle Billy right? What size shortage would the price ceiling create?

9. Only one firm produces and sells soccer balls in the country of Wiknam, and as the story begins, international trade in soccer balls is prohibited. The following equations describe the monopolist's demand, marginal revenue, total cost, and marginal cost:

$$\text{Demand: } P = 10 - Q$$
$$\text{Marginal Revenue: } MR = 10 - 2Q$$
$$\text{Total Cost: } TC = 3 + Q + 0.5Q^2$$
$$\text{Marginal Cost: } MC = 1 + Q$$

   where Q is quantity and P is the price measured in Wiknamian dollars.
   a. How many soccer balls does the monopolist produce? At what price are they sold? What is the monopolist's profit?
   b. One day, the King of Wiknam decrees that henceforth there will be free trade—either imports or exports—of soccer balls at the world price of $6. The firm is now a price taker in a competitive market. What happens to domestic production of soccer balls? To domestic consumption? Does Wiknam export or import soccer balls?
   c. In our analysis of international trade in Chapter 9, a country becomes an exporter when the price without

trade is below the world price and an importer when the price without trade is above the world price. Does that conclusion hold in your answers to parts (a) and (b)? Explain.

d. Suppose that the world price was not $6 but, instead, happened to be exactly the same as the domestic price without trade as determined in part (a). Would allowing trade have changed anything in the Wiknamian economy? Explain. How does the result here compare with the analysis in Chapter 9?

10. Based on market research, a company obtains the following information about the demand and production costs of its new product:

$$\text{Demand: } P = 1{,}000 - 10Q$$
$$\text{Total Revenue: } TR = 1{,}000Q - 10Q^2$$
$$\text{Marginal Revenue: } MR = 1{,}000 - 20Q$$
$$\text{Marginal Cost: } MC = 100 + 10Q$$

where Q indicates the number of units sold and P is the price in dollars.

a. Find the price and quantity that maximize the company's profit.
b. Find the price and quantity that would maximize social welfare.
c. Calculate the deadweight loss from monopoly.
d. Suppose, in addition to the costs above, the product's inventor has to be paid. The company is considering four options:
   i. a flat fee of 2,000 dollars.
   ii. 50 percent of the profits.
   iii. 150 dollars per unit sold.
   iv. 50 percent of the revenue.
   For each option, calculate the profit-maximizing price and quantity. Which, if any, of these compensation schemes would alter the deadweight loss from monopoly? Explain.

11. Larry, Curly, and Moe run the only saloon in town. Larry wants to sell as many drinks as possible without losing money. Curly wants the saloon to bring in as much revenue as possible. Moe wants to make the largest possible profit. Using a single diagram of the saloon's demand curve and its cost curves, show the price and quantity combinations favored by each of the three partners. Explain. (Hint: Only one of these partners will want to set marginal revenue equal to marginal cost.)

12. Many schemes for price discrimination involve some cost. For example, discount coupons take up the time and resources of both the buyer and the seller. This question considers the implications of costly price discrimination. To keep things simple, let's assume that our monopolist's production costs are simply proportional to output so that average total cost and marginal cost are constant and equal to each other.

a. Draw the cost, demand, and marginal-revenue curves for the monopolist. Show the price the monopolist would charge without price discrimination.
b. In your diagram, mark the area equal to the monopolist's profit and call it X. Mark the area equal to consumer surplus and call it Y. Mark the area equal to the deadweight loss and call it Z.
c. Now suppose that the monopolist can perfectly price discriminate. What is the monopolist's profit? (Give your answer in terms of X, Y, and Z.)
d. What is the change in the monopolist's profit from price discrimination? What is the change in total surplus from price discrimination? Which change is larger? Explain. (Give your answer in terms of X, Y, and Z.)
e. Now suppose that there is some cost associated with price discrimination. To model this cost, let's assume that the monopolist has to pay a fixed cost, C, to price discriminate. How would a monopolist make the decision whether to pay this fixed cost? (Give your answer in terms of X, Y, Z, and C.)
f. How would a social planner who cares about total surplus decide whether the monopolist should price discriminate? (Give your answer in terms of X, Y, Z, and C.)
g. Compare your answers to parts (e) and (f). How does the monopolist's incentive to price discriminate differ from the social planner's? Is it possible that the monopolist will price discriminate even though doing so is not socially desirable?

QuickQuiz Answers

1. **c** 2. **d** 3. **b** 4. **d** 5. **a** 6. **b** 7. **a** 8. **c** 9. **c** 10. **c**

Chapter

17

Monopolistic Competition

You walk into a bookstore to buy a book to read. On the store's shelves, you find a James Patterson thriller, a Maya Angelou memoir, a Ron Chernow history, a Stephenie Meyer paranormal romance, and many other choices. When you pick out a book and buy it, what kind of market are you participating in?

On the one hand, the market for books seems competitive. As you browse, hundreds of authors and publishers vie for your attention. And because anyone can enter the industry by writing and publishing a book, the business is not very profitable. For every highly paid writer, there are dozens of struggling ones.

On the other hand, the market for books also seems monopolistic. Because each title is unique, publishers have some latitude in pricing. Sellers in this market are price makers rather than price takers, and the price of books greatly exceeds the marginal cost of producing them. The list price of a typical hardcover novel, for instance, is about $30, but the cost of printing one additional copy is less than $10. The price of e-books is often around $15, while the marginal cost of permitting one extra download is zero.

GAUDILAB/SHUTTERSTOCK.COM; LECHATNOIR/E+/GETTY IMAGES

The market for books fits neither the competitive nor the monopoly model. Instead, it is best described by the model of **monopolistic competition**, the subject of this chapter. The term "monopolistic competition" might at first seem to be an oxymoron, like "jumbo shrimp." But as we will see, monopolistically competitive industries are monopolistic in some ways and competitive in others. The model describes not only the publishing industry but also the market for many other goods and services.

17-1 Between Monopoly and Perfect Competition

The previous two chapters analyzed markets with many competitive firms and markets with a single monopoly firm. Chapter 15 showed that the price in a perfectly competitive market always equals the marginal cost of production. In addition, in the long run, entry and exit drive economic profit to zero, so the price also equals average total cost. Chapter 16 examined how a monopoly firm can use its market power to keep price above marginal cost, leading to a positive economic profit for the firm and a deadweight loss for society. Perfect competition and monopoly are two extreme forms of market structure. Perfect competition describes a market with many firms offering essentially identical products; monopoly describes a market with only one firm.

Although the cases of perfect competition and monopoly illustrate important ideas about how markets work, most markets in the real world include elements of both these cases and are not completely described by either of them. The typical firm faces competition, but the competition is not so rigorous that it makes the firm a price taker like the firms in Chapter 15. The typical firm also has some degree of market power, but not so much that the firm can be described exactly by the monopoly model in Chapter 16. In other words, many industries fall somewhere between the polar cases of perfect competition and monopoly. Economists call this situation **imperfect competition**.

oligopoly
a market structure in which only a few sellers offer similar or identical products

One type of imperfectly competitive market is an **oligopoly**, a market with only a few sellers, each offering a product similar or identical to those of other sellers in the market. Economists often measure a market's domination by a small number of firms with a statistic called the **concentration ratio**, which is the percentage of total output in the market supplied by the four largest firms. In the U.S. economy, most industries have a four-firm concentration ratio under 50 percent, but in some industries, the biggest firms are more dominant. Industries with four-firm concentration ratios of 90 percent or more include aircraft manufacturing, tobacco, passenger car rentals, and express delivery services. These industries are best described as oligopolies. As the next chapter discusses, the small number of firms in oligopolies makes strategic interactions among them crucial to how these markets work. When deciding how much to produce and what price to charge, each firm in an oligopoly is concerned not only with what its competitors are doing but also with how its competitors would react to what it might do.

monopolistic competition
a market structure in which many firms sell products that are similar but not identical

A second type of imperfectly competitive market is called **monopolistic competition**, a market structure in which many firms sell similar but not identical products. In such a market, each firm has a monopoly over its product, but many other firms make similar products that compete for the same customers.

To be more precise, monopolistic competition describes a market with the following attributes:

- **Many sellers:** Numerous firms are competing for the same group of customers.
- **Product differentiation:** Each firm offers a product that is at least slightly different from those of other firms. Rather than being a price taker, each firm faces a downward-sloping demand curve.
- **Free entry and exit:** Firms can enter or exit the market without restriction. The number of firms in the market adjusts until economic profits are driven to zero.

A moment's thought reveals a long list of markets with these attributes: books, video games, restaurants, piano lessons, cookies, clothing, and so on.

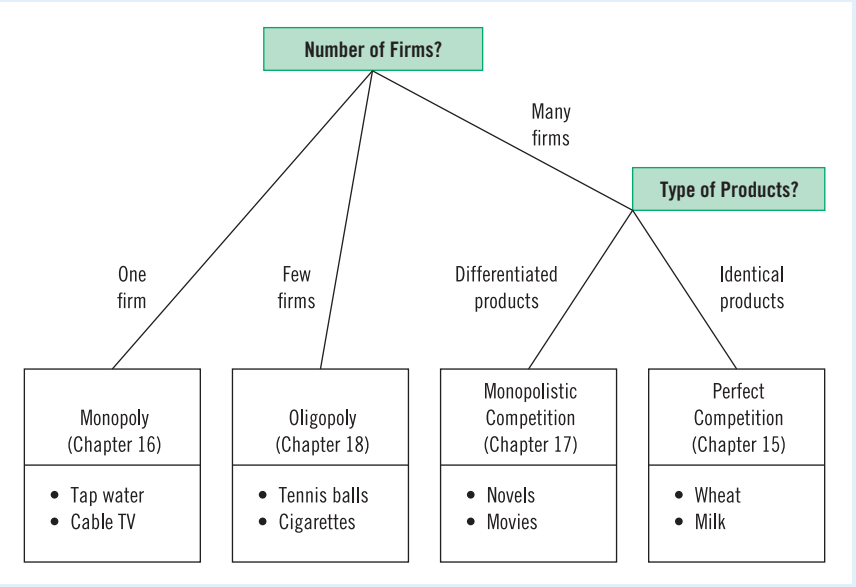
Monopolistic competition, like oligopoly, is a market structure that lies between the extreme cases of perfect competition and monopoly. But oligopoly and monopolistic competition are quite different. Oligopoly departs from the perfectly competitive ideal of Chapter 15 because the market has only a few sellers. The small number of sellers makes rigorous competition less likely and strategic interactions among them vitally important. By contrast, a monopolistically competitive market has many sellers, each of which is small compared with the market. It departs from the perfectly competitive ideal because each seller offers a somewhat different product.

Figure 1 summarizes the four types of market structure. The first question to ask about any market is how many firms there are. If there is only one firm, the market is a monopoly. If there are only a few, it is an oligopoly. If there are many firms, we need to ask another question: Do they sell identical or differentiated products? If their products are identical, the market is perfectly competitive. But if their products are differentiated, the market is monopolistically competitive.

Figure 1

The Four Types of Market Structure

Economists who study industrial organization divide markets into four types: monopoly, oligopoly, monopolistic competition, and perfect competition.



Because reality is never as clear-cut as theory, you may sometimes find it hard to decide what structure best describes a particular market. There is, for instance, no magic number that separates "few" from "many" when counting the number of firms. (Do the approximately dozen companies that sell cars in the United States make the market an oligopoly, or is it more competitive? The answer is open to debate.) Similarly, there is no sure way to determine when products are differentiated and when they are largely identical. (Are different brands of milk really the same? Again, the answer is debatable.) When analyzing actual markets, economists must keep in mind the lessons learned from studying all types of market structures and then apply each lesson as they deem appropriate.

Having defined the various market structures, let's continue our analysis of each of them. This chapter examines monopolistic competition, and the next examines oligopoly.

QuickQuiz

1. Which of the following conditions does NOT describe a firm in a monopolistically competitive market?
   a. It sells a product different from its competitors.
   b. It takes its price as given by market conditions.
   c. It maximizes profit both in the short run and in the long run.
   d. It has the freedom to enter or exit in the long run.
2. Which of the following markets best fits the definition of monopolistic competition?
   a. wheat
   b. tap water
   c. crude oil
   d. haircuts

Answers are at the end of the chapter.

17-2 Competition with Differentiated Products

To understand monopolistically competitive markets, we first consider the decisions facing an individual firm. We then examine what happens in the long run as firms enter and exit the industry. Next, we compare the equilibrium under monopolistic competition to the equilibrium under perfect competition that we examined in Chapter 15. Finally, we discuss whether the outcome in a monopolistically competitive market is desirable from the standpoint of society as a whole.

17-2a The Monopolistically Competitive Firm in the Short Run

Each firm in a monopolistically competitive market is, in many ways, like a monopoly. Because its product differs from those offered by other firms, its demand curve slopes down. (By contrast, a perfectly competitive firm faces a horizontal demand curve at the market price.) The monopolistically competitive firm follows a monopolist's rule for profit maximization: It produces the quantity at which marginal revenue equals marginal cost and then uses its demand curve to find the price at which it can sell that quantity.

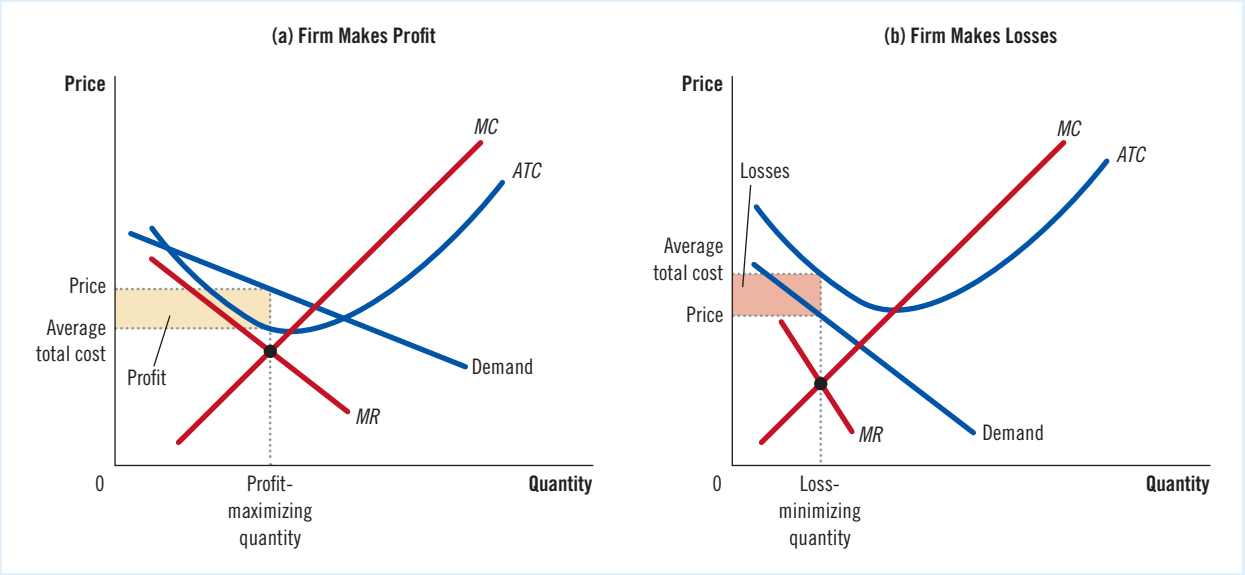
Figure 2 shows the cost, demand, and marginal-revenue curves for two typical firms, each in a different monopolistically competitive industry. In both panels, the profit-maximizing quantity is found where the marginal-revenue and marginal-cost curves intersect. The two panels show different outcomes for the firm's profit. In panel (a), price exceeds average total cost, so the firm makes a profit. In panel (b), price is below average total cost. In this case, the firm cannot make a positive profit, so the best it can do is to minimize its losses.

All this should seem familiar. A monopolistically competitive firm chooses its quantity and price just as a monopoly does. In the short run, these two market structures are similar.

Figure 2

Monopolistic Competitors in the Short Run

Monopolistic competitors, like monopolists, maximize profit by producing the quantity at which marginal revenue equals marginal cost. The firm in panel (a) makes a profit because, at this quantity, price is greater than average total cost. The firm in panel (b) makes losses because, at this quantity, price is less than average total cost.



SCIENCECARTOONSPLUS.COM

"GIVEN THE DOWNWARD SLOPE OF OUR DEMAND CURVE, AND THE EASE WITH WHICH OTHER FIRMS CAN ENTER THE INDUSTRY, WE CAN STRENGTHEN OUR PROFIT POSITION ONLY BY EQUATING MARGINAL COST AND MARGINAL REVENUE. ORDER MORE JELLY BEANS."

17-2b The Long-Run Equilibrium

The situations depicted in Figure 2 do not last long. When firms are making profits, as in panel (a), new firms have an incentive to enter the market. This entry increases the number of products from which customers can choose and, therefore, reduces the demand faced by each firm already in the market. In other words, profit encourages entry, and entry shifts the demand curves of the incumbent firms to the left. As the demand for incumbent firms' products falls, these firms experience declining profits.

Conversely, when firms are making losses, as in panel (b), firms in the market have an incentive to exit. As firms exit, customers have fewer products from which to choose. This decrease in the number of firms expands the demand faced by those that stay in the market. In other words, losses encourage exit, and exit shifts the demand curves of the remaining firms to the right. With increased demand, the remaining firms enjoy greater profits (that is, their losses decline).

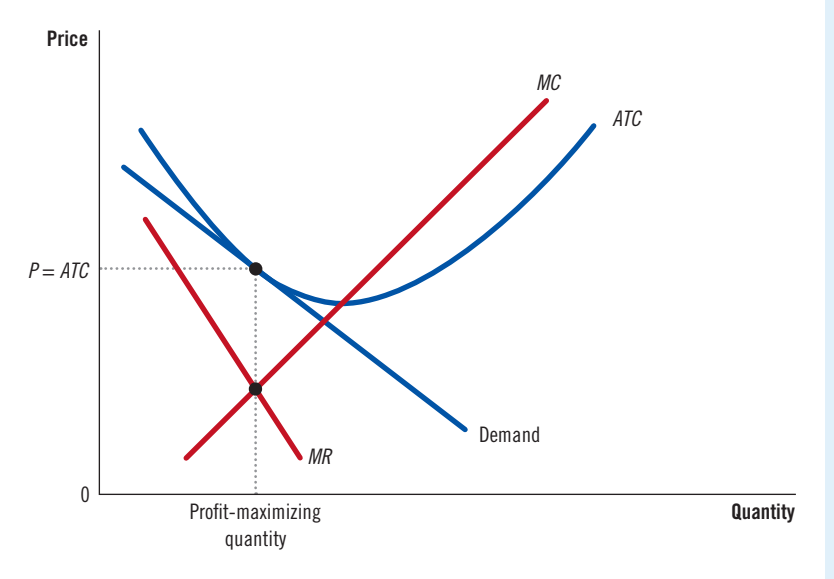
This process of entry and exit continues until the firms in the market make exactly zero economic profit. Figure 3 depicts the long-run equilibrium. Once the market reaches this equilibrium, new firms have no incentive to enter, and existing firms have no incentive to exit.

Notice that the demand curve in this figure just barely touches the average-total-cost curve. Mathematically, the two curves are said to be **tangent** to each other. These two curves must be tangent once entry and exit have driven profit to zero. Because profit per unit sold is the difference between price (found on the demand curve) and average total cost, the maximum profit is zero only if these two curves touch each other without crossing. Also, note that this point of tangency occurs at the same quantity where marginal revenue equals marginal cost. That these two points line up is not a coincidence: It is required because this quantity maximizes profit, which must be exactly zero in the long run.

Figure 3

A Monopolistic Competitor in the Long Run

In a monopolistically competitive market, if firms are making profits, new ones enter, causing the demand curves for the incumbent firms to shift to the left. Similarly, if firms are making losses, some of the firms in the market exit, causing the demand curves of the remaining firms to shift to the right. Because of these shifts in demand, monopolistically competitive firms eventually find themselves in the long-run equilibrium shown here. In this long-run equilibrium, price equals average total cost, and each firm earns zero profit.



To sum up, two characteristics describe the long-run equilibrium in a monopolistically competitive market:

- As in a monopoly market, price exceeds marginal cost ($P > MC$). This occurs because profit maximization requires marginal revenue to equal marginal cost ($MR = MC$) and because the downward-sloping demand curve makes marginal revenue less than the price ($MR < P$).
- As in a perfectly competitive market, price equals average total cost ($P = ATC$). This arises because free entry and exit drive economic profit to zero in the long run.

The second characteristic shows how monopolistic competition differs from monopoly. Because a monopoly is the sole seller of a product without close substitutes, it can earn positive economic profit, even in the long run. By contrast, because monopolistically competitive markets have free entry, the economic profit of a firm in this type of market is driven to zero in the long run.

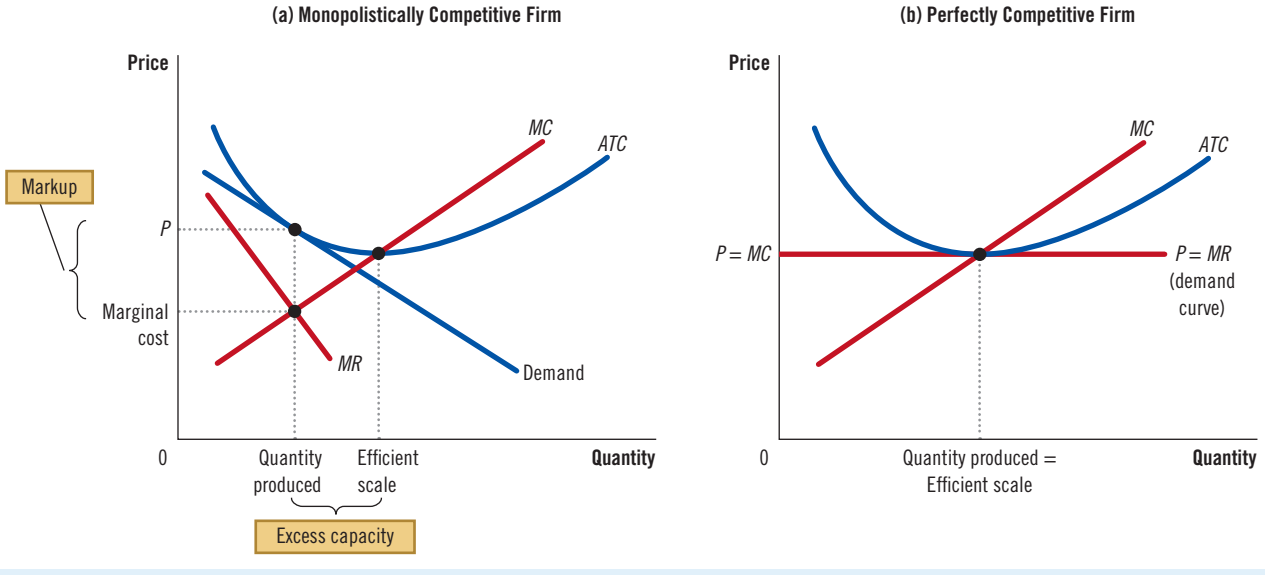
17-2c Monopolistic versus Perfect Competition

Figure 4 compares the long-run equilibria under monopolistic competition and perfect competition. (Chapter 15 discussed the equilibrium with perfect competition.) There are two noteworthy differences: excess capacity and the markup.

Figure 4

Monopolistic versus Perfect Competition

Panel (a) shows the long-run equilibrium in a monopolistically competitive market, and panel (b) shows the long-run equilibrium in a perfectly competitive market. Two differences are notable. (1) The perfectly competitive firm produces at the efficient scale, where average total cost is minimized. By contrast, the monopolistically competitive firm produces at less than the efficient scale. (2) Price equals marginal cost under perfect competition, but price is above marginal cost under monopolistic competition.



Excess Capacity As we have just seen, the process of entry and exit drives each firm in a monopolistically competitive market to a point of tangency between its demand and average-total-cost curves. Panel (a) of Figure 4 shows that the quantity of output at this point is smaller than the quantity that minimizes average total cost. Under monopolistic competition, firms produce on the downward-sloping portion of their average-total-cost curves. In this way, monopolistic competition contrasts starkly with perfect competition. As panel (b) of Figure 4 shows, free entry in competitive markets drives firms to produce at the minimum of average total cost.

The quantity that minimizes average total cost is called the **efficient scale** of the firm. In the long run, perfectly competitive firms produce at the efficient scale, while monopolistically competitive firms produce below this level. Firms are said to have **excess capacity** under monopolistic competition. In other words, a monopolistically competitive firm, unlike a perfectly competitive firm, could increase the quantity it produces and lower the average total cost of production. The firm forgoes this opportunity because, to sell the additional output, it would need to cut its price for all the units it produces. It is more profitable for a monopolistic competitor to continue operating with excess capacity.

Markup over Marginal Cost A second difference between perfect competition and monopolistic competition is the relationship between price and marginal cost. For a perfectly competitive firm, such as the one in panel (b) of Figure 4, price equals marginal cost. For a monopolistically competitive firm, such as the one in panel (a), price exceeds marginal cost because the firm always has some market power.

How is this markup over marginal cost consistent with free entry and zero profit? The zero-profit condition ensures only that price equals average total cost. It does **not** ensure that price equals marginal cost. Indeed, in the long-run equilibrium, monopolistically competitive firms operate on the declining portion of their average-total-cost curves, so marginal cost is below average total cost. For price to equal average total cost, price must be above marginal cost.

This relationship between price and marginal cost highlights a key difference between perfect competitors and monopolistic competitors. Imagine that you were to ask a firm the following question: "Would you like to see another customer come through your door ready to buy from you at your current price?" A perfectly competitive firm would answer that it didn't care. Because price exactly equals marginal cost, the profit from an extra unit sold is zero. By contrast, a monopolistically competitive firm is always eager to get another customer. Because its price exceeds marginal cost, an extra unit sold at the posted price means more profit.

An old economist's joke says that monopolistically competitive markets are those in which sellers send holiday cards to buyers. Currying favor with customers to attract more of them makes sense only if price exceeds marginal cost. And since the business practice of sending out holiday cards is widespread, monopolistically competitive markets must be commonplace.

17-2d Monopolistic Competition and the Welfare of Society

Is the outcome in a monopolistically competitive market desirable from the standpoint of society as a whole? Can government policymakers improve on the market outcome? Previous chapters evaluated markets from the standpoint of efficiency by asking whether the economy is getting the most it can out of its scarce resources. We learned that perfectly competitive markets achieve efficient outcomes (unless there are externalities), while monopoly markets entail deadweight losses.

Monopolistically competitive markets are more complex than either of these polar cases, so evaluating welfare in these markets is a more subtle exercise.

One source of inefficiency in monopolistically competitive markets is the markup of price over marginal cost. Because of the markup, some consumers who value the good at more than the marginal cost of production (but less than the price) will be deterred from buying it. A monopolistically competitive market has the normal deadweight loss of monopoly pricing.

This outcome is undesirable compared with the efficient quantity that arises when price equals marginal cost, but policymakers can't easily fix the problem. To enforce marginal-cost pricing, they would need to regulate all firms that produce differentiated products. Because such products are so common, the administrative burden of such regulation would be overwhelming.

Regulating monopolistic competitors would also entail all the problems of regulating natural monopolies. In particular, because monopolistic competitors are already making zero profits, requiring them to lower their prices to equal marginal cost would cause them to incur losses. To keep these firms in business, the government would need to help them cover these losses. Rather than financing these subsidies with higher taxes, which would entail their own deadweight losses, policymakers may decide it is better to live with the inefficiency of monopolistic pricing.

Another source of inefficiency under monopolistic competition is that the number of firms in the market may not be ideal. That is, there may be too much or too little entry. Think of the externalities associated with entry. Whenever a new firm considers entering the market with a new product, it takes into account only the profit it would make. Yet its entry would also have two effects that are external to the firm:

- **The product-variety externality:** Because consumers benefit from the introduction of a new product, the entry of a new firm confers a positive externality on consumers.
- **The business-stealing externality:** Because other firms lose customers and profits when faced with a new competitor, the entry of a new firm imposes a negative externality on existing firms.

Thus, in a monopolistically competitive market, the entry of new firms entails both positive and negative externalities. Depending on which is larger, a monopolistically competitive market could have too few or too many products.

Both externalities are closely related to the conditions for monopolistic competition. The product-variety externality arises because new firms under monopolistic competition offer products that differ from those of the existing firms. The greater range of choices increases consumer surplus. The business-stealing externality arises because monopolistically competitive firms post a price above marginal cost and, therefore, are eager to sell additional units. Conversely, because perfectly competitive firms produce identical goods and charge a price equal to marginal cost, neither of these externalities exists under perfect competition.

In the end, we can conclude only that monopolistically competitive markets do not have all the desirable welfare properties of perfectly competitive markets. That is, the invisible hand does not ensure that total surplus is maximized under monopolistic competition. Yet because the inefficiencies are subtle, hard to measure, and hard to fix, public policy cannot easily improve the market outcome.

QuickQuiz

3. A monopolistically competitive firm will increase its production if
   a. marginal revenue is greater than marginal cost.
   b. marginal revenue is greater than average total cost.
   c. price is greater than marginal cost.
   d. price is greater than average total cost.
4. New firms will enter a monopolistically competitive market if
   a. marginal revenue is greater than marginal cost.
   b. marginal revenue is greater than average total cost.
   c. price is greater than marginal cost.
   d. price is greater than average total cost.
5. What is true of a monopolistically competitive market in long-run equilibrium?
   a. Price is greater than marginal cost.
   b. Price is equal to marginal revenue.
   c. Firms make positive economic profits.
   d. Firms produce at the minimum of average total cost.

Answers are at the end of the chapter.

17-3 Advertising

In the modern world, it is nearly impossible to go through a typical day without being bombarded with advertising. Whether you are surfing the Internet, watching television, or driving down the highway, some firm will try to convince you to buy its product. Such behavior is a natural feature of monopolistic competition (as well as some oligopolistic industries). When firms sell differentiated products and charge prices above marginal cost, each firm has an incentive to advertise to attract more buyers to its particular product.

The amount of advertising varies substantially across products. Firms that sell differentiated consumer goods, such as over-the-counter drugs, perfumes, soft drinks, razor blades, breakfast cereals, and dog food, typically spend between 10 and 20 percent of their revenue on advertising. Firms that sell industrial products, such as drill presses and communications satellites, typically spend very little on advertising. And those that sell homogeneous products, such as wheat, salt, sugar, and crude oil, often spend nothing at all.

For the overall economy, about 2 percent of total firm revenue is spent on advertising. This spending takes many forms, including ads on websites, social media, television, radio, and billboards and in newspapers, magazines, and direct mail.

17-3a The Debate over Advertising

Is society wasting the resources it devotes to advertising? Or does advertising serve a valuable purpose? Assessing the social value of advertising is difficult and often generates heated arguments among economists. Let's consider both sides of the debate.

The Critique of Advertising Critics argue that firms advertise to manipulate people's tastes. Much advertising is psychological rather than informational. Consider, for example, the typical television commercial for some brand of soft drink. The commercial most likely does not tell the viewer about the product's price or quality. Instead, it might show a group of happy and beautiful people at a party on a beach on a sunny day. In their hands are cans of the soft drink. The goal of the commercial is to convey a subconscious (if not subtle) message: "You too can have many friends and be happy and beautiful, if you drink our product." Critics of advertising argue that such a commercial creates a desire that otherwise might not exist.

Critics also argue that advertising impedes competition. Advertising often tries to convince consumers that products are more different than they truly are. By increasing the perception of product differentiation and fostering brand loyalty, advertising makes buyers less concerned with price differences among similar goods, making the demand for the brand being advertised less elastic. When a firm faces a less elastic demand curve, it can increase its profits by charging a larger markup over marginal cost.

The Defense of Advertising Defenders of advertising argue that firms use advertising to inform customers. Advertising often conveys the prices of the goods offered for sale, the existence of new products, and the ways in which they can be purchased. This information allows customers to make better choices about what to buy, contributing to the efficient allocation of resources.

Defenders also argue that advertising fosters competition. Because advertising may make customers more aware of the available products, customers can more easily take advantage of price differences, thereby reducing the market power of each firm. In addition, advertising allows new firms to enter more easily because it gives them a way to inform and attract customers.

Over time, the view that advertising can make markets more competitive has gained adherents. One important example is the regulation of advertising for lawyers, doctors, and pharmacists. In the past, these groups succeeded in getting state governments to prohibit advertising in their fields on the grounds that it was "unprofessional." In recent years, however, the courts have concluded that the primary effect of these restrictions was to curtail competition. They have, therefore, overturned many of the laws that prohibit advertising in these fields.

Case Study

How Advertising Affects Prices

What effect does advertising have on prices? On the one hand, it might make consumers view products as being more different from each other than they otherwise would. If so, it would make markets less competitive and firms' demand curves less elastic, allowing firms to charge higher prices. On the other hand, advertising might make it easier for consumers to find the firms with the best prices. In this case, it would make markets more competitive and firms' demand curves more elastic, which would lead to lower prices.

In an article published in *The Journal of Law and Economics* in 1972, the economist Lee Benham tested these two hypotheses. In the United States during the 1960s, state governments had vastly different rules about advertising by optometrists. Some states allowed advertising for eyeglasses and eye examinations, but many prohibited it. For example, a Florida law justified the advertising ban as "in the interest of public health, safety, and welfare." Optometrists endorsed these restrictions.

Benham used the differences in state laws as a natural experiment to test the two views of advertising. The results were striking. In states that prohibited advertising, the average price paid for a pair of eyeglasses was $33, or $288 in 2021 dollars. In states that did not restrict advertising, the average price was $26, or $227 in 2021 dollars. Advertising reduced average prices by more than 20 percent.

A similar natural experiment occurred in 1996 when the U.S. Supreme Court struck down a Rhode Island law that banned advertising the prices of liquor products. A study by Jeffrey Milyo and Joel Waldfogel, published in the *American*

Economic Review in 1999, examined liquor prices in Rhode Island after the legal change, compared with liquor prices in the neighboring state of Massachusetts, where there was no change. According to this research, stores in Rhode Island that started advertising cut their prices substantially, often by more than 20 percent, but only on those products that they or their rivals advertised. In addition, after these stores began advertising, they attracted a larger share of customers.

The bottom line: In many markets, advertising fosters competition and leads to lower prices for consumers. ●

17-3b Advertising as a Signal of Quality

Advertising often contains little apparent information about the product being advertised. Consider a firm introducing a new breakfast cereal. It might saturate the airwaves with advertisements showing some actor eating the cereal and exclaiming how wonderful it tastes. How much information does that provide?

According to one theory, the answer is more than you might think. Even advertising that appears to contain little hard information may tell consumers something about product quality. The willingness of the firm to spend a large amount of money on advertising can itself be a **signal** to consumers about the quality of the product being offered.

To see how this works, let's examine the problem facing two firms—General Mills and Kellogg. Each company has just come up with a recipe for a new cereal, which it would sell for $5 a box. To keep things simple, assume that the marginal cost of making cereal is zero, so the $5 is all profit. Each company knows that if it spends $20 million on advertising, it will get 1 million consumers to try its cereal. And each knows that if consumers like the cereal, they will buy it many times.

First, consider General Mills' decision. Based on market research, General Mills knows that its cereal tastes like shredded newspaper with sugar on top. Advertising would sell one box to each of the 1 million consumers, but they would quickly learn that the cereal is not very good and stop buying it. General Mills decides it is not worth spending $20 million on advertising to get only $5 million in sales. So it does not bother to advertise. It sends its cooks back to the kitchen to come up with a better recipe.

Kellogg, on the other hand, knows that its cereal is great. Each person who tries it will buy a box a month for the next year, so the $20 million in advertising will bring in $60 million in sales. In this case, advertising is profitable because Kellogg has a good product that consumers will buy repeatedly. Thus, Kellogg chooses to advertise.

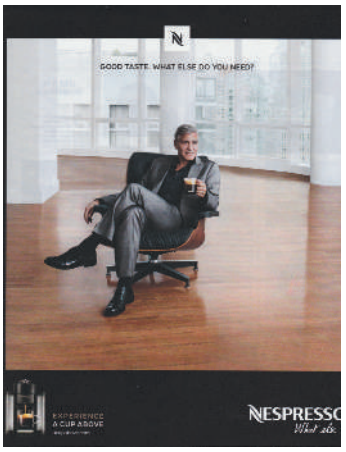
Now consider the behavior of consumers. We began by asserting that consumers are inclined to try a new cereal that they see advertised. But is this behavior rational? Should a consumer try a new cereal just because the seller has chosen to advertise it?

In fact, it may be completely rational for consumers to try new products that they see advertised. In this story, consumers decide to try Kellogg's new cereal because Kellogg advertises. Kellogg chooses to advertise because it knows that its cereal is quite good, while General Mills chooses not to advertise because it knows that its cereal is not good at all. By its willingness to spend on advertising, Kellogg signals to consumers the quality of its cereal. Each consumer thinks, quite sensibly, "If the Kellogg Company is willing to spend so much money advertising this new cereal, it must be really good."

What is striking about this theory of advertising is that the content of the advertisement is irrelevant. Kellogg signals the quality of its product by its willingness

to spend money on advertising. What the advertisements say is not as important as the fact that consumers know the ads are expensive. By contrast, cheap advertising cannot be effective at signaling quality to consumers. In this example, if an advertising campaign cost less than $5 million, both General Mills and Kellogg would use it to market their new cereals. Because both good and bad cereals would now be advertised, consumers could not infer the quality of a new cereal from the fact that it is advertised. Over time, consumers would learn to ignore such cheap advertising.

This theory can explain why firms pay famous actors large sums of money to make advertisements that, on the surface, appear to convey no information at all. The information is not in the ad's content but simply in its existence and expense.



UGPIX/MEG/NEWSCOM

Is it rational for consumers to be impressed that George Clooney endorses this product?

17-3c Brand Names

Many markets have two types of firms: those that sell products with widely recognized brand names and those that sell generic substitutes. For example, Bayer aspirin competes with generic aspirin. Coke and Pepsi compete with less familiar colas. Firms with brand names usually spend more on advertising and charge more for their products. Just as there is debate about the economics of advertising, there is debate about the economics of brand names.

Critics argue that brand names cause consumers to perceive differences that do not really exist. In many cases, the generic is almost indistinguishable from the brand-name good. Consumers' willingness to pay more for the brand-name good, these critics assert, is a form of irrationality fostered by advertising. The economist Edward Chamberlin, one of the early developers of the theory of monopolistic competition, concluded from this argument that brand names were bad for the economy. He proposed that the government discourage their use by refusing to enforce the trademarks that companies use to identify their products.

More recently, economists have defended brand names as a way for consumers to ensure that the goods they buy are of high quality. There are two related arguments. First, brand names provide consumers with **information** about quality when quality cannot be easily judged in advance of purchase. Second, brand names give firms an **incentive** to maintain high quality because firms have a financial stake in maintaining their brands' reputations.

To see how these arguments work in practice, consider a famous brand name: McDonald's. Imagine that you are driving through an unfamiliar town and want to stop for lunch. You see a McDonald's and a local restaurant next to it. Which do you choose? The local restaurant may offer better food at lower prices, but you don't really know. By contrast, McDonald's offers a consistent product across many cities and countries. Its brand name is useful to you as a way of judging the quality of what you are about to buy.

The McDonald's brand name also ensures that the company has an incentive to maintain quality. For example, if some customers were to become ill from spoiled food sold at a McDonald's, the news would be disastrous for the company. McDonald's would lose much of the valuable reputation that it has built up with years of expensive advertising. As a result, it would lose sales and profit not only in the outlet that sold the bad food but also in many other McDonald's outlets throughout the country. By contrast, if some customers were to become ill from bad food at a local restaurant, that restaurant might have to close, but the lost profits would be much smaller. McDonald's has a greater incentive to ensure that its food is safe.

The debate over brand names thus centers on the question of whether consumers are rational in preferring brand-name products. Critics argue that brand names are the result of an irrational consumer response to advertising. Defenders argue that consumers have good reason to pay more for brand-name products because they can be more confident in these products' quality.

QuickQuiz

6. If advertising makes consumers more loyal to particular brands, it could _____ the elasticity of demand and _____ the markup of price over marginal cost.
   a. increase; increase
   b. increase; decrease
   c. decrease; increase
   d. decrease; decrease
7. If advertising makes consumers more aware of alternative products, it could _____ the elasticity of demand and _____ the markup of price over marginal cost.
   a. increase; increase
   b. increase; decrease
   c. decrease; increase
   d. decrease; decrease
8. Advertising can be a signal of quality
   a. if advertising is freely available to all firms.
   b. if the benefit of attracting customers is greater for firms with better products.
   c. only if consumers are irrationally attracted to the products they see advertised.
   d. only if the content of the ads contains credible information about the products.

Answers are at the end of the chapter.

17-4 Conclusion

Monopolistic competition is true to its name: It is a hybrid of monopoly and competition. Like a monopoly, each monopolistic competitor faces a downward-sloping demand curve and charges a price above marginal cost. As in a perfectly competitive market, there are many firms, and entry and exit drive the profit of each monopolistic competitor toward zero in the long run. Table 1 summarizes these lessons.

Because monopolistically competitive firms produce differentiated products, each advertises to attract customers to its own brand. To some extent, advertising manipulates consumers' tastes, promotes irrational brand loyalty, and impedes

Table 1

Monopolistic Competition: Between Perfect Competition and Monopoly

	Market Structure		
	Perfect Competition	**Monopolistic Competition**	**Monopoly**
Features that all three market structures share			
Goal of firms	Maximize profits	Maximize profits	Maximize profits
Rule for maximizing	$MR = MC$	$MR = MC$	$MR = MC$
Can earn economic profits in the short run?	Yes	Yes	Yes
Features that monopolistic competition shares with monopoly			
Price taker?	Yes	No	No
Price	$P = MC$	$P > MC$	$P > MC$
Produces welfare-maximizing level of output?	Yes	No	No
Features that monopolistic competition shares with perfect competition			
Number of firms	Many	Many	One
Entry in the long run?	Yes	Yes	No
Can earn economic profits in the long run?	No	No	Yes

competition. Often, however, it informs consumers, establishes brand names of reliable quality, and fosters competition.

The theory of monopolistic competition describes many markets in the economy. It is somewhat disappointing, therefore, that the theory does not yield simple and compelling advice for public policy. From the standpoint of economic theorists, the allocation of resources in monopolistically competitive markets is not perfect. Yet from the standpoint of practical policymakers, there may be little that can be done to improve it.

Chapter in a Nutshell

- A monopolistically competitive market is characterized by three attributes: many firms, differentiated products, and free entry and exit.
- The long-run equilibrium in a monopolistically competitive market differs from that in a perfectly competitive market in two ways. First, in a monopolistically competitive market, each firm has excess capacity. That is, it chooses a quantity that puts it on the downward-sloping portion of the average-total-cost curve. Second, each firm charges a price above marginal cost.
- Monopolistic competition does not have all the desirable properties of perfect competition. There is the standard deadweight loss of monopoly caused by the markup of price over marginal cost. In addition, the number of firms (and thus the number of product varieties) can be too large or too small. In practice, the

ability of policymakers to correct these inefficiencies is limited.

- The product differentiation inherent in monopolistic competition leads to the use of advertising and brand names. Critics of advertising and brand names argue that firms use them to manipulate consumers' tastes and reduce competition. Defenders of advertising and brand names argue that firms use them to inform consumers and compete more vigorously on price and product quality.

Key Concepts

oligopoly, p. 342

monopolistic competition, p. 342

Questions for Review

1. Describe the three attributes of monopolistic competition. How is monopolistic competition like monopoly? How is it like perfect competition?
2. Draw a diagram depicting a firm that is making a profit in a monopolistically competitive market. Now show what happens to this firm as new firms enter the industry.
3. Draw a diagram of the long-run equilibrium in a monopolistically competitive market. How is price related to average total cost? How is price related to marginal cost?
4. Does a monopolistic competitor produce too much or too little output compared with the most efficient level? What practical considerations make it difficult for policymakers to solve this problem?
5. How might advertising reduce economic well-being? How might advertising increase economic well-being?
6. How might advertising with no apparent informational content still inform consumers?
7. Explain two benefits that might arise from the existence of brand names.

Problems and Applications

1. Among monopoly, oligopoly, monopolistic competition, and perfect competition, how would you classify the markets for each of the following drinks?
   a. tap water
   b. bottled water
   c. cola
   d. beer
2. Classify the following markets as perfectly competitive, monopolistic, or monopolistically competitive, and explain your answers.
   a. wooden no. 2 pencils
   b. copper
   c. local electricity service
   d. peanut butter
   e. lipstick
3. For each of the following characteristics, say whether it describes a perfectly competitive firm, a monopolistically competitive firm, both, or neither.
   a. sells a product differentiated from that of its competitors
   b. has marginal revenue less than price
   c. earns economic profit in the long run
   d. produces at the minimum of average total cost in the long run
   e. equates marginal revenue and marginal cost
   f. charges a price above marginal cost
4. For each of the following characteristics, say whether it describes a monopoly firm, a monopolistically competitive firm, both, or neither.
   a. faces a downward-sloping demand curve
   b. has marginal revenue less than price
   c. faces the entry of new firms selling similar products
   d. earns economic profit in the long run
   e. equates marginal revenue and marginal cost
   f. produces the socially efficient quantity of output
5. You are hired as a consultant to a monopolistically competitive firm, which reports the following information about its price, marginal cost, and average total cost. Can the firm possibly be maximizing profit? If not, what should it do to increase profit? If the firm is maximizing profit, is the market in a long-run equilibrium? If not, what will happen to restore long-run equilibrium?
   a. $P < MC$, $P > ATC$
   b. $P > MC$, $P < ATC$

c. $P = MC$, $P > ATC$
d. $P > MC$, $P = ATC$

6. Sparkle is one of the many firms in the market for toothpaste, which is in a long-run, monopolistically competitive equilibrium.
   a. Draw a diagram showing Sparkle's demand curve, marginal-revenue curve, average-total-cost curve, and marginal-cost curve. Label Sparkle's profit-maximizing output and price.
   b. What is Sparkle's profit? Explain.
   c. On your diagram, show the consumer surplus derived from the purchase of Sparkle toothpaste. Also, show the deadweight loss relative to the efficient outcome.
   d. If the government forced Sparkle to produce the efficient level of output, what would happen to the firm? What would happen to Sparkle's customers?

7. Consider a monopolistically competitive market with N firms. Each firm's business opportunities are described by the following equations:

$$\text{Demand: } Q = 100/N - P.$$

$$\text{Marginal Revenue: } MR = 100/N - 2Q.$$

$$\text{Total Cost: } TC = 50 + Q^2.$$

$$\text{Marginal Cost: } MC = 2Q.$$

   a. How does N, the number of firms in the market, affect each firm's demand curve? Why?
   b. How many units does each firm produce? (The answers to this and the next two questions depend on N.)
   c. What price does each firm charge?
   d. How much profit does each firm make?
   e. In the long run, how many firms will exist in this market?

8. The market for peanut butter in Nutville is monopolistically competitive and in long-run equilibrium. One day, consumer advocate Jif Skippy discovers that all brands of peanut butter in Nutville are identical. Thereafter, the market becomes perfectly competitive and again reaches its long-run equilibrium. Using an appropriate diagram, explain whether each of the following variables increases, decreases, or stays the same for a typical firm in the market.
   a. price
   b. quantity
   c. average total cost
   d. marginal cost
   e. profit

9. For each of the following pairs of firms, explain which one would be more likely to engage in advertising.
   a. a family-owned farm or a family-owned restaurant
   b. a manufacturer of forklifts or a manufacturer of cars
   c. a company that invented a very comfortable razor or a company that invented a less comfortable razor

10. Sleek Sneakers Co. is one of many firms in the market for shoes.
    a. Assume that Sleek is currently earning short-run economic profit. On a correctly labeled diagram, show Sleek's profit-maximizing output and price as well as the area representing profit.
    b. What happens to Sleek's price, output, and profit in the long run? Explain this change in words, and show it on a new diagram.
    c. Suppose that over time, consumers become more focused on stylistic differences among shoe brands. How would this change in attitudes affect each firm's price elasticity of demand? In the long run, how will this change in demand affect Sleek's price, output, and profit?
    d. At the profit-maximizing price you identified in part (c), is Sleek's demand curve elastic or inelastic? Explain.

QuickQuiz Answers

1. **b** 2. **d** 3. **a** 4. **d** 5. **a** 6. **c** 7. **b** 8. **b**

Chapter

18

Oligopoly

oligopoly
a market structure in which only a few sellers offer similar or identical products

game theory
the study of how people behave in strategic situations

If you play tennis, you have probably used balls from one of a handful of brands: Penn, Wilson, Dunlop, Prince, or Babolat. These few companies supply most of the tennis balls sold in the United States. Together, they determine the quantity of tennis balls produced and, given the market demand curve, the price at which tennis balls are sold.

The market for tennis balls is an example of an **oligopoly**. The essence of an oligopolistic market is that there are only a few sellers, so the actions of any one of them can have a large impact on the profits of all the others. This chapter examines how this interdependence shapes the firms' behavior and what problems it raises for public policy.

The analysis of oligopoly leads us to **game theory**, the study of how people behave in strategic situations. By "strategic," we mean a situation in which people, when choosing a course of action, must anticipate how others might respond to their choice. Strategic thinking is crucial not only in chess, checkers, and tic-tac-toe but also in many business decisions. Because oligopolistic markets have only a few firms, each firm must be strategic when making supply decisions. The firms are keenly aware that each firm's profit

depends not only on how much it produces but also on how much each of the others produce. When setting production, a firm in an oligopoly needs to consider how its choices might affect the choices of other firms in the market.

Game theory isn't necessary for analyzing competitive or monopoly markets. In a market that is either perfectly or monopolistically competitive, each firm is so small compared with the overall market that strategic interactions are insignificant. And for a monopoly, there are no other firms to worry about. But game theory is important for understanding oligopolies and may be applied whenever a small number of players interact with one another. It helps explain the strategies that people choose, whether they are playing tennis or selling tennis balls.

18-1 Markets with Only a Few Sellers

An oligopolistic market has only a small group of sellers and is characterized by the tension between cooperation and self-interest. Oligopolists can make the most profit if they cooperate and together act like one big monopolist—producing a small quantity of output and charging a price well above marginal cost. Yet because each oligopolist cares only about its own profit, powerful incentives pull them apart, making it hard to maintain the cooperative outcome.

18-1a A Duopoly Example

Consider the simplest type of oligopoly, one with only two members, called a **duopoly**. Oligopolies with three or more members face the same problems as duopolies, so little is lost by starting with the simpler case.

Imagine a town in which only two residents, Jack and Jill, own wells that produce water safe for drinking. Each Saturday, Jack and Jill decide how many gallons of water to pump, bring the water to town, and sell it for whatever price the market will bear. To keep things simple, suppose that they can pump as much as they want without cost. That is, the marginal cost of water is zero.

Table 1 shows the town's demand schedule for water. The first column shows the total quantity demanded, and the second shows the price. If the well owners sell a total of 10 gallons of water, water goes for $110 a gallon. If they sell a total of 20 gallons, the price falls to $100 a gallon. And so on. If you graphed these two columns of numbers, you would get a standard downward-sloping demand curve.

The last column in Table 1 shows total revenue from the sale of water. It equals the quantity sold times the price. Because there is no cost to pumping water, the total revenue of the two producers equals their total profit.

Now consider how the organization of the town's water industry affects the price of water and the quantity sold.

18-1b Competition, Monopolies, and Cartels

Before examining the price and quantity of water that results from the Jack and Jill duopoly, let's consider the outcomes that would result if the water market were either perfectly competitive or monopolistic. These polar cases are natural benchmarks.

If the market for water were perfectly competitive, the production decisions of each firm would drive price to equal marginal cost. Because here the marginal cost

Table 1

The Demand Schedule for Water

Quantity	Price	Total Revenue (and total profit)
0 gallons	$120	$0
10	110	1,100
20	100	2,000
30	90	2,700
40	80	3,200
50	70	3,500
60	60	3,600
70	50	3,500
80	40	3,200
90	30	2,700
100	20	2,000
110	10	1,100
120	0	0

of pumping additional water is zero, the equilibrium price of water under perfect competition would be zero as well. The equilibrium quantity would then be 120 gallons. The price of water would reflect the cost of producing it, and the efficient quantity of water would be produced and consumed.

Now consider how a monopoly would behave. Table 1 shows that total profit is maximized at a quantity of 60 gallons and a price of $60 a gallon. A profit-maximizing monopolist, therefore, would produce this quantity and charge this price. As is standard for monopolies, price would exceed marginal cost. The result would be inefficient because the quantity of water produced and consumed would fall short of the socially efficient level of 120 gallons.

What outcome would the duopolists achieve? One possibility is that Jack and Jill get together and agree on the quantity of water to produce and the price to charge for it. Such an agreement among firms over production and price is called **collusion**, and the group of firms acting in unison is called a **cartel**. Once a cartel is formed, the market is in effect served by a monopoly, and the analysis from Chapter 16 applies. That is, if Jack and Jill collude, they will agree on the monopoly outcome because it maximizes their total profit. Together, they produce a total of 60 gallons, which sell at a price of $60 a gallon. Price exceeds marginal cost, and the outcome is socially inefficient.

collusion
an agreement among firms in a market about quantities to produce or prices to charge

cartel
a group of firms acting in unison

A cartel must agree not only on total production but also on the amount produced by each member. Each wants a larger share of the market because that means more individual profit. In this case, Jack and Jill must agree on how to split the production of 60 gallons. If they agree to split the market equally, each produces 30 gallons, the price is $60 a gallon, and each earns a profit of $1,800.

18-1c The Equilibrium for an Oligopoly

Oligopolists would like to form cartels and earn monopoly profits, but that is often impossible. Squabbling among cartel members over how to divide the profit can make agreement among members difficult. In addition, antitrust laws prohibit explicit agreements among oligopolists. Even talking about pricing and production restrictions with competitors can be a criminal offense. Let's consider, therefore, what happens if Jack and Jill decide separately how much water to produce.

One might expect Jack and Jill to reach the monopoly outcome on their own because this outcome maximizes their joint profit. In the absence of a binding agreement, however, the monopoly outcome is unlikely. To see why, imagine that Jack expects Jill to produce only 30 gallons (half the monopoly quantity). Jack would reason as follows:

"I could produce 30 gallons as well. Together, we'd sell 60 gallons of water at $60 a gallon. My profit would be $1,800 (30 gallons × $60 a gallon). But why settle for that? I could produce 40 gallons. Then together, we'd sell 70 gallons of water at $50 a gallon. My profit would be $2,000 (40 gallons × $50 a gallon). Total profit in the market would fall, but who cares? My profit would rise because I'd have more of the market."

That's logical, as far as it goes. But there's a catch: Jill might well think the same way. Then she would bring 40 gallons to the market, too. Total sales would be 80 gallons, the price would fall to $40, and total profits would be $3,200. Jack and Jill would each earn only $1,600. By pursuing their individual self-interest when deciding how much to produce, the duopolists produce a total quantity greater than the monopoly quantity, charge a price lower than the monopoly price, and earn total profit less than the monopoly profit.

The logic of self-interest increases the duopoly's output above the monopoly level, but it does not push the duopolists all the way to the competitive allocation. Consider what happens when Jack and Jill each produce 40 gallons. The price is $40, and they each make a profit of $1,600. Jack ponders the situation further, but his self-interested logic leads to a different conclusion:

"Right now, my profit is $1,600. Suppose I increase production to 50 gallons. In this case, a total of 90 gallons would be sold, and the price would be $30 a gallon. Then, my profit would be only $1,500. Hmmm. Rather than increasing production and driving down the price, I am better off keeping my production at 40 gallons."

Jill, on her own, reaches the same conclusion.

The outcome in which Jack and Jill each produce 40 gallons looks like some sort of equilibrium. In fact, this outcome is called a Nash equilibrium. (It is named after the Nobel Prize–winning mathematician and economic theorist John Nash, whose life was portrayed in the book and movie *A Beautiful Mind*.) A **Nash equilibrium** is a situation in which economic actors interacting with one another each choose their best strategy given the strategies that the others have chosen. In this case, once Jill is producing 40 gallons, the best strategy for Jack is also to produce 40 gallons. Similarly, once Jack is producing 40 gallons, the best strategy for Jill is also to produce 40 gallons. At this Nash equilibrium, neither Jack nor Jill has an incentive to make a different decision.

Nash equilibrium
a situation in which economic actors interacting with one another each choose their best strategy given the strategies that all the other actors have chosen

This example shows the tension between cooperation and self-interest that is the essence of oligopolies. Oligopolists would be better off cooperating to attain the monopoly outcome. Yet because they each pursue their own self-interest, they fail to do so. Each oligopolist is tempted to raise production and capture a larger share of the market. As each tries to do this, total production rises, the price falls, and total profit falls.

Yet self-interest does not drive the market all the way to the competitive outcome. Like monopolists, oligopolists know that producing more reduces the price they will receive, which in turn affects profits. So they stop short of following the competitive firm's rule of producing up to the point where price equals marginal cost.

In summary, when firms in an oligopoly individually choose production to maximize profit, they produce a quantity greater than the level produced by a monopoly and less than the level produced under perfect competition. The oligopoly price is less than the monopoly price but greater than the competitive price (which equals marginal cost).

18-1d How the Size of an Oligopoly Affects the Market Outcome

We can use the insights from studying duopoly to discuss how the size of an oligopoly affects the market outcome. Suppose, for instance, that Jude and Jade suddenly discover water sources on their properties and join Jack and Jill in the water oligopoly. The demand schedule in Table 1 remains the same, but more producers are available to satisfy this demand. How does an increase from two to four sellers affect the price and quantity of water in the town?

If the sellers formed a cartel, they would again try to maximize total profit by producing the monopoly quantity and charging the monopoly price. Just as with only two sellers, the cartel members would need to agree on individual production levels and find a way to enforce the agreement. As the cartel grows larger, however, this outcome is less likely. If you have ever been a member of a team or a club, you might have noticed that working harmoniously becomes more difficult as the size of the group increases. In addition, evading the antitrust laws may be harder with a larger group of conspirators.

If the oligopolists do not form a cartel, they must each decide on their own how much water to produce. To see how the increase in the number of sellers affects the outcome, consider the decision facing each seller. At any time, each well owner has the option to raise production by one gallon. In making this decision, the well owner weighs two effects:

- **The output effect:** Because price exceeds marginal cost, selling one more gallon of water at the going price increases profit.
- **The price effect:** Because raising production increases the total quantity sold, the price of water declines, as does the profit on all the other gallons sold.

If the output effect outweighs the price effect, the well owner increases production. If the price effect outweighs the output effect, the owner does not raise production. (In this case, it is profitable to reduce production.) Each oligopolist increases production until these two marginal effects exactly balance, taking the other firms' production as given.

Now consider how the number of firms in the industry affects the marginal analysis of each oligopolist. The more firms there are, the smaller each firm's market share is. As a firm's market share shrinks, the less the firm is concerned about its own impact on the market price. That is, as the oligopoly grows, the magnitude of the price effect falls.

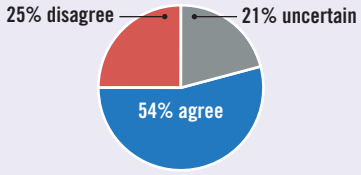
When the oligopoly grows very large, the price effect disappears altogether. In this extreme case, the production decision of an individual firm no longer affects the market price. Each firm takes the market price as given when deciding how much to produce and, therefore, increases production as long as price exceeds marginal

Ask the Experts

Market Share and Market Power

"If a small number of firms have a large combined market share in a properly defined market, it is strong evidence that those firms have substantial market power."

What do economists say?



Source: IGM Economic Experts Panel, September 25, 2018.

cost. In other words, a large oligopoly is essentially a group of competitive firms.

In summary, as the number of sellers in an oligopoly grows, an oligopolistic market increasingly resembles a competitive market. The price approaches marginal cost, and the quantity produced approaches the socially efficient level.

This analysis of oligopoly offers a new perspective on international trade. Imagine that Toyota and Honda are the only automakers in Japan, Volkswagen and BMW are the only ones in Germany, and Ford and General Motors are the only ones in the United States. If these countries prohibited international trade in autos, each would have an oligopoly with only two members, and the market price and quantity would likely depart substantially from the competitive ideal. With international trade, however, a world market arises, and the oligopoly in this example has six members. Allowing free trade increases the number of producers from which each consumer can choose, and the greater competition keeps prices closer to marginal cost. Thus, the theory of oligopoly provides another reason, in addition to the theory of comparative advantage discussed in Chapter 3, why countries can benefit from free trade.

QuickQuiz

1. The key feature of an oligopolistic market is that
   a. each firm sells a product different from other firms.
   b. a single firm chooses a point on the market demand curve.
   c. each firm takes the market price as given.
   d. a small number of firms are acting strategically.
2. If an oligopolistic industry organizes itself as a cooperative cartel, it will produce a quantity of output _____ the competitive level and _____ the monopoly level.
   a. less than; more than
   b. more than; less than
   c. less than; equal to
   d. equal to; more than
3. If an oligopoly does not cooperate and each firm chooses its own quantity, the industry will produce a quantity of output _____ the competitive level and _____ the monopoly level.
   a. less than; more than
   b. more than; less than
   c. less than; equal to
   d. equal to; more than
4. As the number of firms in an oligopoly grows, the industry approaches a level of output _____ the competitive level and _____ the monopoly level.
   a. less than; more than
   b. more than; less than
   c. less than; equal to
   d. equal to; more than

Answers are at the end of the chapter.

18-2 The Economics of Cooperation

Oligopolists would like to reach the monopoly outcome but may find it hard to cooperate. This situation is not unusual: Often in life, people fail to cooperate with one another even when cooperation would make them all better off. An oligopoly is just one example.

This section looks more closely at the problems that arise when cooperation is desirable but difficult. This requires an understanding of game theory. We focus on a "game" called the **prisoners' dilemma**, which teaches a general lesson that applies to any group trying to maintain cooperation among its members.

prisoners' dilemma a particular "game" between two captured prisoners that illustrates why cooperation is difficult to maintain even when it is mutually beneficial

18-2a The Prisoners' Dilemma

The prisoners' dilemma is a story about two criminals who have been captured by the police. Call them Bonnie and Clyde. The police have enough evidence to convict them of the minor crime of carrying an unregistered gun, so each would spend a year in jail. The police also suspect that these criminals committed a bank robbery but lack evidence to convict them for it. The police question Bonnie and Clyde in separate rooms and offer each of them the following deal:

"Right now, we can lock you up for 1 year. But if you confess to the bank robbery and implicate your partner, we'll give you immunity, and you can go free. Your partner will get 20 years in jail. If you both confess, we won't need your testimony, and we can avoid the cost of a trial, so you'll each get a sentence of 8 years."

If Bonnie and Clyde, heartless bank robbers that they are, care only about their own individual sentences, what would you expect them to do? Figure 1 shows the **payoff matrix** for their choices. Each prisoner has two strategies: confess or remain silent. The sentence each prisoner gets depends on the strategy he or she chooses and the strategy chosen by his or her partner in crime.

Consider first Bonnie's decision. She reasons as follows: "We had a fabulous time robbing banks together, but now I don't know what Clyde is going to do. If he remains silent, my best strategy is to confess because then I'll go free rather than spending a year in jail. If he confesses, I should still confess because then I'll spend just 8 years in jail rather than 20. So, whatever Clyde does, I'm better off confessing."

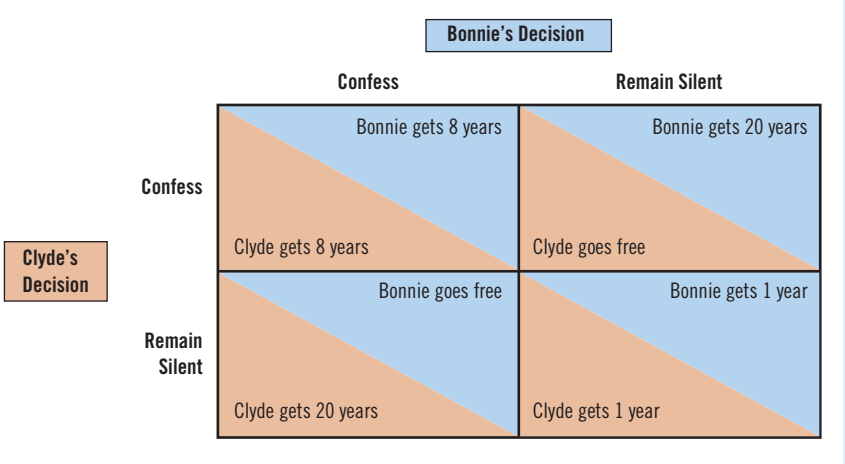
dominant strategy a strategy that is best for a player in a game regardless of the strategies chosen by the other players

In the language of game theory, a strategy is called a **dominant strategy** if it is the best one for a player to follow regardless of the strategies pursued by other players. In this case, confessing is a dominant strategy for Bonnie. She spends less time in jail if she confesses, regardless of whether Clyde confesses or remains silent.

Figure 1

The Prisoners' Dilemma

In this game between two criminals suspected of committing a major crime, the sentence that each receives depends both on his or her decision whether to confess or remain silent and on the decision made by the other.



Now consider Clyde's decision. He faces the same choices as Bonnie, and he reasons the same way. Regardless of what Bonnie does, Clyde can reduce his jail time by confessing. In other words, confessing is also a dominant strategy for Clyde.

In the end, both Bonnie and Clyde confess, and they each spend 8 years in jail. This outcome is a Nash equilibrium: Each chooses the best strategy available given the strategy the other is following. Yet, from their standpoint, the outcome is terrible. If they had **both** remained silent, both would have been better off, spending only 1 year in jail on the gun charge. Because they pursue their own interests, the two prisoners together reach an outcome that is worse for each of them.

You might have thought that Bonnie and Clyde would have planned for this situation. But even if they had, they would still run into problems. Imagine that, before the police captured them, the two criminals had sworn undying love and agreed never to confess. Clearly, this pact would make them both better off **if** they both lived up to it because each would spend only 1 year in jail. They could then ride off into a glorious sunset. But would Bonnie and Clyde remain silent simply because they had agreed they would? Once they are being questioned separately, the logic of self-interest takes over and leads them to confess. Cooperation between the prisoners is difficult to maintain because cooperation is individually irrational. So is love, but it's easier to maintain when you're not facing a prison sentence.

18-2b Oligopolies as a Prisoners' Dilemma

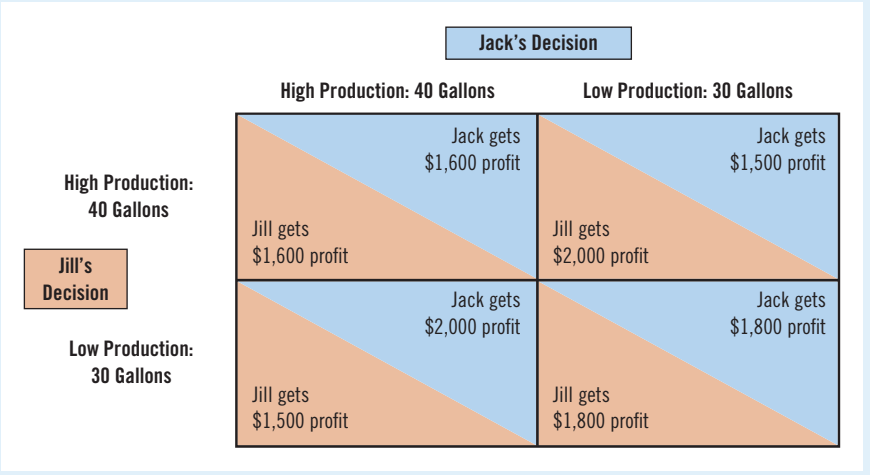
What does the prisoners' dilemma have to do with markets and imperfect competition? It turns out that the decisions oligopolists face in trying to reach the monopoly outcome are similar to those that Bonnie and Clyde faced in the prisoners' dilemma.

Consider again the choices facing Jack and Jill. After prolonged negotiation, the two water suppliers agree to keep production at 30 gallons. That way, the price will be high, and together, they will earn the maximum profit. After they agree on production levels, however, each of them must decide whether to honor this agreement or to ignore it and produce at a higher level. Figure 2 shows how the profits of the two producers depend on the strategies they choose.

Figure 2

Jack and Jill's Oligopoly Game

In this game between Jack and Jill, the profit that each earns from selling water depends on both the quantity he or she chooses to sell and the quantity the other chooses to sell.



Jack might reason as follows: "I could keep production at 30 gallons as we agreed, or I could raise my production and sell 40 gallons. If Jill lives up to the agreement and keeps her production at 30 gallons, then my profit is $2,000 if I sell 40 gallons and $1,800 if I sell 30 gallons. In this case, I am better off with the higher production level. If Jill fails to live up to the agreement and produces 40 gallons, then I earn $1,600 by selling 40 gallons and $1,500 by selling 30 gallons. Again, I am better off with higher production. So, whatever Jill chooses to do, I am better off reneging on our agreement and producing at the higher level."

Producing 40 gallons is a dominant strategy for Jack. Of course, Jill reasons in the same way, and so both produce at the higher level of 40 gallons. The result is the inferior outcome (from Jack and Jill's standpoint) with low profits for each of the two producers.

This example shows why oligopolies have trouble maintaining monopoly profits. The monopoly outcome is jointly rational, but each oligopolist has an incentive to cheat. Just as self-interest drives the suspects in the prisoners' dilemma to confess, self-interest makes it hard for the oligopolists to maintain the cooperative outcome with low production, high prices, and monopoly profits.

OPEC and the World Oil Market

The story about the town's market for water is fictional, but if we change water to crude oil and Jack and Jill to Saudi Arabia and Iraq, the story is close to reality. Much of the world's oil is produced by a few countries, mostly in the Middle East. These countries together make up an oligopoly. Their decisions about how much oil to pump are much the same as Jack and Jill's decisions about how much water to pump.

In 1960, the countries that produce much of the world's oil formed a cartel called the Organization of Petroleum Exporting Countries (OPEC). It includes Saudi Arabia, Iraq, Iran, United Arab Emirates, Kuwait, Venezuela, and several other nations. In 2016, ten other oil-producing nations, led by Russia, joined forces with OPEC, and the cartel is now known as OPEC Plus. Together, OPEC Plus countries control most of the world's oil reserves. The cartel tries to raise the price of its product through a coordinated reduction in quantity produced. To do so, it sets production targets for each of the member countries.

The problem that OPEC Plus faces is much the same as the problem that Jack and Jill face in our story. The countries in the cartel would like to maintain a high price for oil. But each member is tempted to increase its production to get a larger share of the total profit. OPEC Plus members frequently agree to reduce production but then cheat on their agreements.

OPEC was quite successful at maintaining cooperation and high prices in the period from 1973 to 1985. The price of crude oil rose from $3 a barrel in 1972 to $11 in 1974 and then to $35 in 1981. But in the mid-1980s, member countries began arguing about production levels, and OPEC became ineffective at maintaining cooperation. By 1986, the price of crude oil had fallen back to $13 a barrel.

In recent years, the members of OPEC have continued to meet regularly and to confer with allies in the larger oil bloc, but they have been less successful at reaching and enforcing agreements. Changes in technology, such as the development of fracking, have expanded oil supply around the world and reduced OPEC's market power. As a result, fluctuations in oil prices have been driven more by the natural forces of supply and demand than by the cartel's artificial restrictions on production. ●

18-2c Other Examples of the Prisoners' Dilemma

The logic of the prisoners' dilemma applies not only to oligopolies but also to many other situations. Here are two examples in which self-interest impedes cooperation, leading to inferior outcomes for the parties involved.

Arms Races In the decades after World War II, the world's two superpowers—the United States and the Soviet Union—were engaged in a prolonged competition over military power. This struggle motivated some of the early work on game theory. Theorists pointed out that an arms race is much like the prisoners' dilemma. Today, it applies to relations among the United States, Russia, and another great military power, China.

Consider the decisions of the United States and the Soviet Union about whether to build new weapons or to disarm. Each country wants to have more arms than the other because larger arsenals give it more influence in world affairs. But each country also worries about the other's weapons.

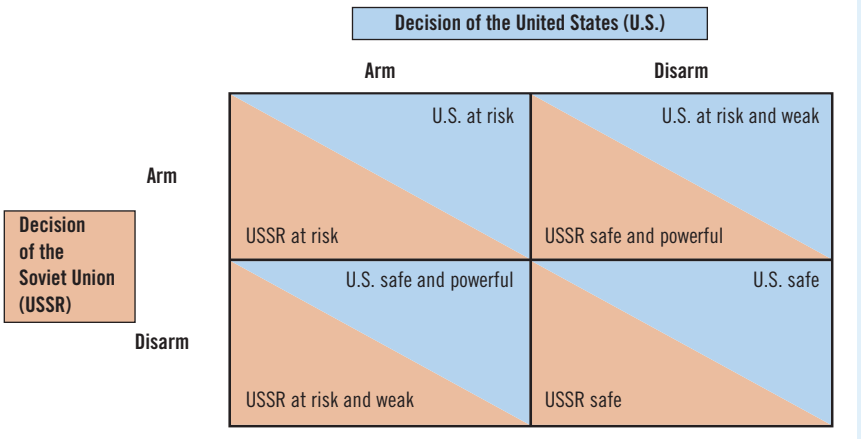
Figure 3 shows the payoff matrix for this deadly game. If the Soviet Union arms, the United States is better off doing the same to prevent the loss of power. If the Soviet Union disarms, the United States is better off arming because doing so would make it more powerful. For each country, arming is a dominant strategy. So each country chooses to continue the arms race, resulting in the inferior outcome with both countries at risk.

From about 1945 to 1991, the United States and the Soviet Union attempted to solve this problem through arms control negotiations and agreements. The difficulties the two countries faced were like those encountered by oligopolists in trying to maintain a cartel. Much as oligopolists argue over production levels, the United States and the Soviet Union argued over the amount and type of arms that each country would be allowed. And just as cartels have trouble enforcing production levels, the United States and the Soviet Union feared that the other country would find ways to cheat. In both arms races and oligopolies, the logic of self-interest can drive the participants toward the noncooperative outcome, which is worse for both parties. Yet with transparency and stringent methods for verifying that agreements are being honored, it is possible to break out of the boxes of the prisoners' dilemma. It helps, however, to understand the pressures that hamper cooperation on both sides.

Figure 3

An Arms-Race Game

In this game between two countries, the safety and power of each depends on what its adversary does, as well as on its own decision whether to arm.



Common Resources Chapter 11 noted that people tend to overuse common resources. One can view this problem as an example of the prisoners' dilemma.

Imagine that two oil companies—ExxonMobil and Chevron—own adjacent oil fields. Under the fields is a common pool of oil worth $120 million. Drilling a well to recover the oil costs $10 million. If each company drills one well, each will get half of the oil and earn a $50 million profit ($60 million in revenue minus $10 million in costs).

Because the pool of oil is a common resource, the companies will not use it efficiently. Suppose that either company could drill a second well. If one company has two of the three wells, that company gets two-thirds of the oil, which yields a profit of $60 million. The other company gets one-third of the oil for a profit of $30 million. Yet if each company drills a second well, the two companies again split the oil. In this case, each bears the cost of a second well and earns a profit of only $40 million.

Figure 4 shows the game. Drilling two wells is a dominant strategy for each company. Once again, the self-interest of the two players leads them to an inferior outcome.

18-2d The Prisoners' Dilemma and the Welfare of Society

The prisoners' dilemma shows that cooperation can be difficult to maintain, even when it would make both players in the game better off. This lack of cooperation is a problem for those directly involved. But is it a problem from the standpoint of society as a whole? The answer depends on the circumstances.

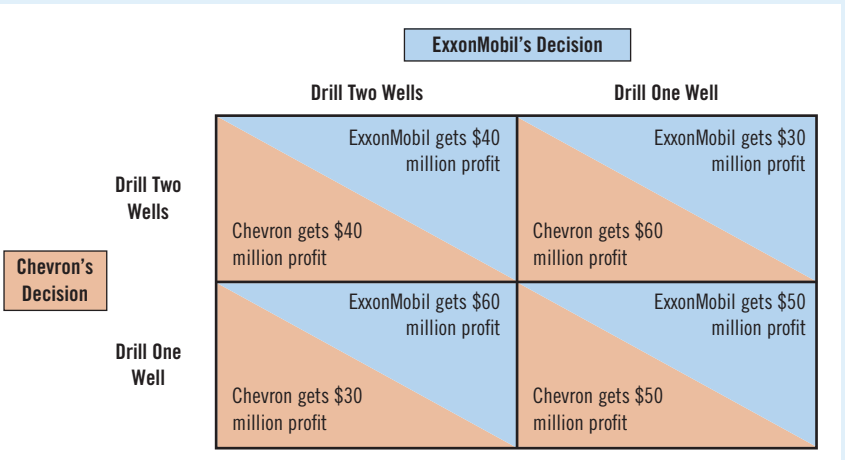
In some cases, the noncooperative equilibrium is bad for society as well as the players. In the arms race depicted in Figure 3, both the United States and the Soviet Union end up at risk (and so does everyone else on the planet). In the common-resources game in Figure 4, the extra wells dug by Chevron and ExxonMobil are pure waste. In these cases, society would be better off if the two players could reach the cooperative outcome.

By contrast, in the case of oligopolists trying to maintain monopoly profits, a lack of cooperation is desirable from the standpoint of society. The monopoly outcome is good for the oligopolists but bad for consumers. As Chapter 7 showed, the competitive outcome is best for society because it maximizes total surplus. When oligopolists fail to cooperate, the quantity they produce is closer to this optimal level.

Figure 4

A Common-Resources Game

When firms pump oil from a common pool, each firm's profit depends on both the number of wells it drills and the number of wells drilled by the other firm.



Put differently, the invisible hand guides markets to allocate resources efficiently only when markets are competitive, and markets are competitive only when firms in the market fail to cooperate with one another.

Similarly, consider the case of the police questioning two suspects. Lack of cooperation between the suspects is desirable for society because it allows the police to convict more criminals. The prisoners' dilemma is a dilemma for the prisoners, but it can be a boon to everyone else.

18-2e Why People Sometimes Cooperate

The prisoners' dilemma shows that cooperation is difficult. But is it impossible? Not all prisoners, when questioned by the police, turn in their partners. Cartels sometimes manage to maintain collusive arrangements, despite the incentive for members to defect. Very often, players can solve the prisoners' dilemma because they play the game not once but many times.

To see why cooperation is easier to enforce in repeated games, let's return to our duopolists, Jack and Jill, whose choices were given in Figure 2. They would like to achieve the monopoly outcome in which each produces 30 gallons. Yet if Jack and Jill are to play this game only once, neither has any incentive to live up to an agreement to do so. Self-interest drives each of them to renege and choose the dominant strategy of 40 gallons.

Now suppose that Jack and Jill know that they will play the same game every week. When making their initial agreement to keep production low, they can specify what happens if one party reneges. They might agree, for instance, that once one of them produces 40 gallons, both will produce 40 gallons forever after. This penalty is easy to enforce because if one party produces at the high level, the other has every reason to do the same.

The threat of this penalty may be all that is needed to maintain cooperation. Each person knows that defecting would raise his or her profit from $1,800 to $2,000. But this benefit would last for only one week. Thereafter, profit would fall to $1,600 and stay there. As long as the players care enough about future profits, they will forgo the one-time gain from defection. In a game of repeated prisoners' dilemma, like an ongoing arms race, the two players may well be able to reach the cooperative outcome.

The Prisoners' Dilemma Tournament

Imagine that you are playing the prisoners' dilemma with a person being questioned in a separate room and that you are to play with this other person many times. Your score at the end of the game is the total number of years you will spend in jail, a total you'd like to make as small as possible. What strategy would you play? Would you begin by confessing or remaining silent? How would the other player's actions in one round affect your choices in subsequent rounds?

This has now become a complicated game. To encourage cooperation, each player will want to impose some penalty when the player in the other room does not cooperate. Yet the strategy described earlier for Jack and Jill's water cartel—defect forever as soon as the other player defects—is not in the least forgiving. If the game is going to be repeated many times, a strategy that allows players to return to the cooperative outcome after a period of noncooperation may be preferable.

To see what strategies work best, the political scientist Robert Axelrod held a contest. People entered by submitting computer programs designed to play a game of repeated prisoners' dilemma. Each program was then paired with each of the others as in a round-robin tournament. The goal was to receive the fewest total years in jail.

The winning program turned out to be a simple strategy called **tit-for-tat**. According to tit-for-tat, a player should start by cooperating and then do whatever the other player did last time. A tit-for-tat player cooperates until the other player defects and then defects until the other player cooperates again. This strategy starts out friendly, penalizes unfriendly players, and forgives them if warranted. To Axelrod's surprise, this simple strategy did better than the more complex strategies that people had sent in.

The tit-for-tat strategy has a long history. It is essentially "an eye for an eye, a tooth for a tooth." The prisoners' dilemma tournament suggests that this classic strategy may be a good rule of thumb for playing some of the games of life. ●

QuickQuiz

5. The prisoners' dilemma is a two-person game illustrating that
   a. the cooperative outcome could be worse for both people than the Nash equilibrium.
   b. even if the cooperative outcome is better than the Nash equilibrium for one person, it might be worse for the other.
   c. even if cooperation is better than the Nash equilibrium, each person might have an incentive not to cooperate.
   d. rational, self-interested individuals will naturally avoid the Nash equilibrium because it is worse for both of them.

6. Two people facing the prisoners' dilemma may cooperate if they
   a. recognize that the Nash equilibrium is worse for both people than the cooperative equilibrium.
   b. play the game repeatedly and expect noncooperation to be met with future retaliation.
   c. each choose their dominant strategy.
   d. each realize that the strategy they choose is not known to the other until the outcome is realized.

Answers are at the end of the chapter.

18-3 Public Policy toward Oligopolies

One of the **Ten Principles of Economics** in Chapter 1 is that governments can sometimes improve market outcomes. This principle applies to oligopolistic markets, where cooperation leads to production that is too low and prices that are too high from the standpoint of society as a whole. The allocation of resources will be closer to the social optimum if firms in an oligopoly compete rather than cooperate. Let's consider how policymakers can foster competition.

18-3a Restraint of Trade and the Antitrust Laws

The common law can inhibit cooperation among oligopolists. Normally, freedom of contract is an essential part of a market economy. Businesses and households use contracts to arrange mutually advantageous trades, and they rely on the court system to enforce contracts. Yet, for many centuries, judges in England and the United States have deemed agreements among competitors to reduce quantities and raise prices to be contrary to the public good. They have, therefore, refused to enforce such agreements.

The Sherman Antitrust Act of 1890 codified and reinforced this policy in the United States:

> Every contract, combination in the form of trust or otherwise, or conspiracy, in restraint of trade or commerce among the several States, or with foreign nations, is declared to be illegal . . . 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 misdemeanor, and on conviction thereof, shall be punished by fine not exceeding fifty thousand dollars, or by imprisonment not exceeding one year, or by both said punishments, in the discretion of the court.

The Sherman Act elevated agreements among oligopolists from unenforceable contracts to criminal conspiracies.

The Clayton Act of 1914 further strengthened the antitrust laws. According to this statute, plaintiffs who could prove that they were damaged by an illegal arrangement to restrain trade could sue and recover three times the damages sustained. The purpose of this unusual rule of triple damages is to encourage private lawsuits against conspiring oligopolists.

Today, the U.S. Justice Department and private parties have the authority to bring legal suits to enforce the antitrust laws. As Chapter 16 discussed, these laws are used to prevent mergers that would give a firm excessive market power. These laws are also used to prevent oligopolists from acting together in ways that would make their markets less competitive.

An Illegal Phone Call

Firms in oligopolies have a strong incentive to collude to reduce production, raise prices, and increase profits. The great 18th-century economist Adam Smith was well aware of this potential market failure. In *The Wealth of Nations*, he wrote, "People of the same trade seldom meet together, but the conversation ends in a conspiracy against the public, or in some diversion to raise prices."

For a modern example of Smith's observation, consider this phone conversation between two airline executives in the early 1980s. The call was reported in the *New York Times* on February 24, 1983. Robert Crandall was president of American Airlines, and Howard Putnam was president of Braniff Airways, a major airline at the time. Here's an excerpt:

> **Crandall:** I think it's dumb as hell . . . to sit here and pound the @#$% out of each other and neither one of us making a #$%& dime.
> **Putnam:** Do you have a suggestion for me?
> **Crandall:** Yes, I have a suggestion for you. Raise your $%*& fares 20 percent. I'll raise mine the next morning.
> **Putnam:** Robert, we . . .
> **Crandall:** You'll make more money, and I will, too.
> **Putnam:** We can't talk about pricing!
> **Crandall:** Oh @#$%, Howard. We can talk about any &*#@ thing we want to talk about.

Putnam was right: The Sherman Antitrust Act prohibits competing executives from even talking about fixing prices. When Putnam gave a recording of this conversation to the Justice Department, the Justice Department filed suit against Crandall.

Two years later, Crandall and the Justice Department reached a settlement in which Crandall agreed to restrictions on his business activities, including his contacts with officials at other airlines. The Justice Department said that the terms of the settlement would "protect competition in the airline industry, by preventing American and Crandall from any further attempts to monopolize passenger airline service on any route through discussions with competitors about the prices of airline services." ●

18-3b Controversies over Antitrust Policy

What kinds of behavior the antitrust laws should prohibit is often controversial. Most commentators agree that price-fixing agreements among competing firms should be illegal. Yet the antitrust laws have been used to condemn some business practices whose effects are less obvious. Here are three examples.

Resale Price Maintenance One example of a controversial business practice is **resale price maintenance**. Imagine that Superduper Electronics sells smartphones to retail stores for $400. If Superduper requires the retailers to charge customers $500, it is said to engage in resale price maintenance. Any retailer that charged less than $500 would violate its contract with Superduper.

At first, resale price maintenance might seem anticompetitive. Like an agreement among cartel members, it prevents the retailers from competing on price. For this reason, the courts have sometimes viewed resale price maintenance as an antitrust violation.

Yet some economists defend the practice. First, they deny that it is aimed at reducing competition. If Superduper Electronics wanted to exert its market power, it would raise the wholesale price rather than control the resale price. What's more, Superduper has no reason to discourage competition among its retailers. Because a cartel of retailers sells less than a group of competitive retailers, Superduper would be worse off if its retailers were a cartel.

Second, resale price maintenance may have a legitimate goal. Superduper may want its retailers to provide customers with a pleasant showroom and a knowledgeable sales force. Yet, without resale price maintenance, some customers would take advantage of one store's service to learn about the smartphone's special features and then buy the item at a discount retailer that does not provide this service. Good customer service can be viewed as a public good among the retailers that sell Superduper products. As Chapter 11 discussed, when one person provides a public good, others can use it without paying for it. In this case, discounters would free ride on the service provided by other retailers, leading to less service than is desirable. Resale price maintenance is one way for Superduper to solve this free-rider problem.

The example of resale price maintenance illustrates an important principle: **Business practices that appear to reduce competition sometimes have legitimate purposes.** This principle makes the application of the antitrust laws all the more difficult. Those in charge of enforcing these laws must determine what kinds of behavior impede competition and reduce economic well-being. Often that job is not easy.

Predatory Pricing Firms with market power typically use it to raise prices above the competitive level. But should policymakers ever be concerned that firms with market power might charge prices that are too low? This question is at the heart of a second debate over antitrust policy.

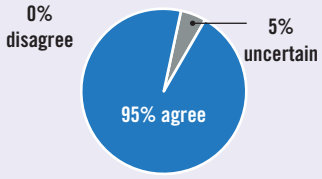
Imagine that a large airline, call it Coyote Air, has a monopoly on some route. Then, Roadrunner Express enters and takes 20 percent of the market, leaving Coyote with 80 percent. In response to this competition, Coyote starts slashing its fares. Some antitrust analysts argue that Coyote's move could be anticompetitive: The price cuts may be intended to drive Roadrunner out of the market so Coyote can recapture its monopoly and raise prices again. Such behavior is called **predatory pricing**.

Although predatory pricing is a common claim in antitrust suits, some economists say that predatory pricing is rarely, if ever, a profitable business strategy. Why? For a price war to drive out a rival, prices must be below cost. Yet if Coyote starts selling

Ask the Experts — Antitrust in the Digital Economy

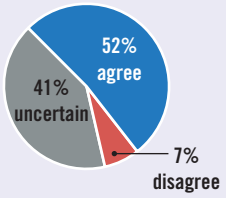
"Google's dominance of the market for internet search arose mainly from a combination of economies of scale and a quality algorithm."

What do economists say?



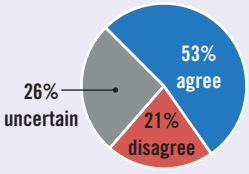
"In light of Google's dominance, its current operating practices could have a substantial negative effect on social welfare in the long run."

What do economists say?



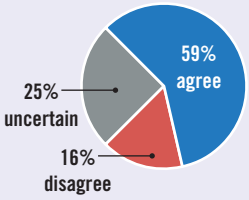
"The nature of the market dominance of technology giants in the digital economy warrants either the imposition of some kind of regulation or a fundamental change in antitrust policy."

What do economists say?



"Requiring Facebook to divest WhatsApp and Instagram is likely to make society better off."

What do economists say?



Source: IGM Economic Experts Panel, November 3, 2020, December 22, 2020.

cheap tickets at a loss, it had better be ready to fly more planes because low fares will attract more customers. Roadrunner, meanwhile, can respond to Coyote's predatory move by cutting back on flights. As a result, Coyote ends up bearing more than 80 percent of the losses, putting Roadrunner in a good position to survive the price war. As in the old Roadrunner–Coyote cartoons, the predator suffers more than the prey.

Economists debate whether predatory pricing should concern antitrust policymakers. When, if ever, is predatory pricing a profitable strategy? Are the courts capable of distinguishing between competitive and predatory price cuts? These are hard questions.

Bundling A third example of a controversial business practice is **bundling**. Suppose that Makemoney Movies produces two new films—*Superheroes* and *Hamlet*. If Makemoney offers theaters the two films together at a single price, rather than separately, the studio is said to be bundling its two products.

When the practice of bundling movies was challenged, the Supreme Court banned it. The court reasoned as follows: Imagine that *Superheroes* is a blockbuster and *Hamlet* is an unprofitable art film. The studio could use the high demand for *Superheroes* to force theaters to buy *Hamlet*. Bundling, the court concluded, could be a mechanism for a firm to expand its market power.

In 1963, the economist George Stigler offered a rebuttal to this argument. Imagine that theaters are willing to pay $200,000 for *Superheroes* and nothing for *Hamlet*. Then the most that a theater would pay for the two movies together is $200,000—the same as it would pay for *Superheroes* by itself. Forcing the theater to accept a worthless movie as part of the deal does not increase its willingness to pay. Makemoney cannot increase its market power simply by bundling the two movies together.

Stigler suggested another explanation for why bundling exists: It is a form of price discrimination. Suppose there are two theaters. City Theater is willing to pay $150,000 for *Superheroes* and $50,000 for *Hamlet*. Country Theater is just the opposite: It is willing to pay $50,000 for *Superheroes* and $150,000 for *Hamlet*. If Makemoney charges separate prices for the two films, its best strategy is to charge $150,000 for each film, and each theater chooses to show only one film. Yet if Makemoney offers the two movies as a bundle, it can charge each theater $200,000 for the movies. If different theaters value the films differently, bundling may allow the studio to increase profit by charging a combined price closer to the buyers' total willingness to pay.

Bundling remains a controversial business practice. The Supreme Court's argument that bundling allows a firm to extend its market power to other goods may not be well founded, at least in its simplest form. Yet economists have proposed more elaborate theories for how bundling can impede competition. Given the current state of economic knowledge, debate continues whether bundling is generally adverse for society.

Case Study

The Microsoft Case

A particularly important and controversial antitrust case was the U.S. government's suit against the Microsoft Corporation, filed in 1998. The case did not lack drama. It pitted one of the world's most powerful regulatory agencies (the U.S. Justice Department) against one of the world's richest men (Bill Gates). Testifying for the government was a prominent economist (MIT professor Franklin Fisher). Testifying for Microsoft was another prominent economist (MIT professor Richard Schmalensee, a former student of Franklin Fisher). At stake was the future of one of the world's most valuable companies (Microsoft) in one of the economy's fastest-growing industries (software).

Bundling was a central issue in the Microsoft case—in particular, whether Microsoft should be allowed to integrate its Internet Explorer browser into its Windows operating system. The government said that Microsoft was bundling these two products together to extend its power in the market for operating systems to the unrelated market of Internet browsers. Allowing Microsoft to incorporate such products into its operating system, the government argued, would deter other software companies from entering the market and offering new products.

Microsoft said that putting new features into old products is a natural part of technological progress. By the 1990s, cars included CD players and air conditioners, which were once sold separately, and cameras came with built-in flashes. The same was true with operating systems. Over time, Microsoft added many features to Windows that were previously stand-alone products. This has made computers more reliable and easier to use because consumers could be confident that the pieces worked together. The integration of Internet technology, Microsoft argued, was the natural next step.

One point of disagreement concerned the extent of Microsoft's market power. Noting that more than 80 percent of new personal computers used a Microsoft operating system, the government argued that the company was effectively a monopoly and was trying to expand its influence into new markets. Microsoft replied that the software market was always changing and that Microsoft's Windows was constantly being challenged by competitors, such as the Apple Mac and Linux operating systems. It also argued that the low price it charged for Windows—about $50 then, or only 3 percent of the price of a typical computer—was evidence that its market power was severely limited.

Like many large antitrust suits, the Microsoft case became a legal morass. In November 1999, after a long trial, Judge Penfield Jackson ruled that Microsoft had great monopoly power and that it had illegally abused that power. In June 2000, after hearings on possible remedies, he ordered that Microsoft be broken up into two companies—one that sold the operating system and one that sold applications software. A year later, an appeals court overturned Jackson's breakup order and handed the case to a new judge. In September 2001, the Justice Department announced that it no longer sought a breakup of the company and wanted to settle the case quickly.

The two sides reached a settlement in November 2002. Microsoft accepted some restrictions on its business practices, and the government accepted that a browser would remain part of the Windows operating system. But the settlement did not end Microsoft's antitrust troubles. In subsequent years, the company contended with several private antitrust suits, as well as suits brought by the European Union alleging a variety of anticompetitive behaviors.

Technological development has relegated the dispute over the once-mighty Explorer browser to the level of a historical footnote. In June 2021, Microsoft said it was retiring Explorer, which had lost most of its market share, replacing it with a new browser called Edge. Some analysts credited the Justice Department's settlement with Microsoft as a crucial step that allowed browsers like Google Chrome and Apple Safari to grow and ultimately supplant Explorer. ●

AP IMAGES/LAURA RAUCH

"Me? A monopolist? Now just wait a minute . . ."

In the News

Amazon in the Crosshairs

The Washington, D.C., attorney general takes aim at the giant online retailer.

A New Antitrust Case Cuts to the Core of Amazon's Identity

By Gilad Edelman

"I founded Amazon 26 years ago with the long-term mission of making it Earth's most customer-centric company," Jeff Bezos testified before the House Antitrust Subcommittee last summer. "Not every business takes this customer-first approach, but we do, and it's our greatest strength."

Bezos' obsession with customer satisfaction is at the center of Amazon's self-mythology. Every move the company makes, in this account, is designed with only one goal in mind: making the customer happy. If Amazon has become an economic juggernaut, the king of ecommerce, that's not because of any unfair practices or sharp elbows; it's simply because customers love it so much.

The antitrust lawsuit filed against Amazon on Tuesday directly challenges that narrative. The suit, brought by Karl Racine, the Washington, DC, attorney general, focuses on Amazon's use of a so-called most-favored-nation clause in its contracts with third-party sellers, who account for most of the sales volume on Amazon. A most-favored-nation clause requires sellers not to offer their products at a lower price on any other website, even their own. According to the lawsuit, this harms consumers by artificially inflating prices across the entire internet, while preventing other ecommerce sites from competing against Amazon on price. "I filed this antitrust lawsuit to put an end to Amazon's ability to control prices across the online retail market," Racine said in a press conference announcing the case.

For a long time, Amazon openly did what DC is alleging; its "price parity provision" explicitly restricted third-party sellers from offering lower prices on other sites. It stopped in Europe in 2013, after competition authorities in the UK and Germany began investigating it. In the US, however, the provision lasted longer, until Senator Richard Blumenthal wrote a letter to antitrust agencies in 2018 suggesting Amazon was violating antitrust law. A few months later, in early 2019, Amazon dropped price parity.

But that wasn't the end of the story. The DC lawsuit alleges that Amazon simply substituted a new policy that uses different language to accomplish the same result as the old rule. Amazon's Marketplace Fair Pricing Policy informs third-party sellers that they can be punished or suspended for a variety of offenses, including "setting a price on a product or service that is significantly higher than recent prices offered on or off Amazon." This rule can protect consumers when used to prevent price-gouging for scarce products, as happened with face masks in the early days of the pandemic. But it can also be used to **inflate** prices for items that sellers would prefer to offer more cheaply. The key phrase is "off Amazon." In other words, Amazon reserves the right to cut off sellers if they list their products more cheaply on another website—just as it did under the old price parity provision. According to the final report filed by the House Antitrust Subcommittee last year, based on testimony from third-party sellers, the new policy "has the same effect of blocking sellers from offering lower prices to consumers on other retail sites."

The main form that this price discipline takes, according to sellers who have spoken out against Amazon either publicly or in anonymous testimony, is through manipulating access to the Buy Box—those Add to Cart and Buy Now buttons at the top right of an Amazon product listing. When you go to buy something, there are often many sellers trying to make the sale. Only one can "win the Buy Box," meaning they're the one who gets the sale when you click one of those buttons. Because most customers don't scroll down to see what other sellers are offering the product, winning the Buy Box is crucial for anyone trying to make a living by selling on Amazon . . .

Jason Boyce, a longtime Amazon seller turned consultant, explained to me how this works. He and his partners were excited when the last third-party seller contract they signed with Amazon, to sell sporting goods on the site, didn't include the

QuickQuiz

7. The antitrust laws aim to
   a. facilitate cooperation among firms in oligopolistic industries.
   b. encourage mergers to take advantage of economies of scale.
   c. discourage firms from moving production facilities overseas.
   d. prevent firms from acting in ways that reduce competition.

8. Antitrust enforcement is controversial mainly because
   a. cooperative domestic firms are best equipped to deal with international competitors.
   b. some business practices that seem anticompetitive may have legitimate purposes.
   c. excessive competition can drive some firms out of business, causing job losses.
   d. vigorous enforcement can reduce business profitability, lowering shareholder value.

Answers are at the end of the chapter.

price parity provision. "We thought, 'This is great! We can offer discounts on Walmart, and Sears, and wherever else,'" he said. But then something odd happened. Boyce (who spoke with House investigators as part of the antitrust inquiry) noticed that once his company lowered prices on other sites, sales on Amazon started tanking. "We went to the listing, and the Add to Cart button was gone, the Buy Now button was gone. Instead, there was a gray box labeled 'See All Buying Options.' You could still buy the product, but it was an extra click. Now, an extra click on Amazon is an eternity—they're all about immediate gratification." Moreover, his company's ad spending plummeted, which he realized was because Amazon doesn't show users ads for products without a Buy Box. "So what did we do? We went back and raised our prices everywhere else, and within 24 hours everything came back. Traffic improved, clicks improved, and sales came back." . . .

Boyce's experience illustrates something important about most-favored-nation clauses: On their own, they aren't illegal. The problem comes when they're used by a company with a dominant share of the market. If a store wants to feature a certain brand on its shelves in exchange for an agreement not to sell more cheaply at a rival chain, the brand can decide whether the deal is worth it. But in the case of Amazon, according to sellers like Boyce, there is no real choice. The DC attorney general's lawsuit points out that Amazon accounts for somewhere between 50 and 70 percent of the US online retail market, and it notes that "a staggering 74 percent [of consumers] go directly to Amazon when they are ready to buy a specific product." It accuses Amazon of using its price policy to maintain that monopoly power by preventing rival platforms from using lower prices to eat into its market share.

RVLSOFT/SHUTTERSTOCK.COM

In a statement emailed to reporters, Amazon did not exactly deny that it punishes sellers who offer lower prices elsewhere. Rather, it suggested that this is ultimately good for the consumer. "The DC attorney general has it exactly backward—sellers set their own prices for the products they offer in our store," the company said. "Amazon takes pride in the fact that we offer low prices across the broadest selection, and like any store we reserve the right not to highlight offers to customers that are not priced competitively. The relief the AG seeks would force Amazon to feature higher prices to customers, oddly going against core objectives of antitrust law."

But this logic relies on a very idiosyncratic definition of "priced competitively." When someone goes to Amazon to buy something, they want the site to show them the best deal available **on Amazon**. If Jenny's Bike Supply has the best deal on Amazon for chain locks, then it's the best deal, regardless of whether Jenny is also selling the locks for an even better price on eBay. If Amazon makes it harder to buy the lock from Jenny in this scenario, the only thing it accomplishes is forcing customers to settle for the second-best deal. And, of course, it will probably succeed in forcing Jenny to raise prices on eBay. What it won't do is result in **lower** prices on Amazon.

All of which makes the DC lawsuit a narrower and potentially more winnable case than some of the other antitrust litigation that has been brought against tech companies. ■

Questions to Discuss

1. Have you ever bought anything from Amazon? If so, do you think you got a good price and good service?
2. Do you think Amazon should be barred from favoring sellers that offer Amazon customers their lowest prices? Why or why not?

Source: *Wired*, May 25, 2021.

18-4 Conclusion

Oligopolies would like to act like monopolies, but self-interest drives them toward competition. Where oligopolies end up on this spectrum depends on the number of firms in the oligopoly and the extent to which the firms cooperate. The story of the prisoners' dilemma shows why oligopolies can fail to maintain cooperation, even when cooperation is in their best interest.

Policymakers regulate the behavior of oligopolists through the antitrust laws. The proper scope of these laws is the subject of ongoing debate. There is little doubt that price fixing among competing firms reduces economic welfare and is an appropriate target for regulators, but some business practices that appear to reduce competition may have legitimate, if subtle, purposes. As a result, policymakers need to be careful when they use the substantial powers of the antitrust laws to place limits on firm behavior.

Chapter in a Nutshell

- Oligopolists maximize their total profits by forming a cartel and acting like a monopolist. Yet, if oligopolists make decisions about production levels individually, the result is a greater quantity and a lower price than under the monopoly outcome. The larger the number of firms in the oligopoly, the closer the quantity and price will be to the levels that would prevail under perfect competition.
- The prisoners' dilemma shows that self-interest can prevent people from maintaining cooperation, even when cooperation is in their mutual interest. The logic of the prisoners' dilemma applies to many situations, including arms races, common-resource problems, and oligopolies.
- Policymakers use the antitrust laws to prevent oligopolies from engaging in behavior that reduces competition. The application of these laws can be controversial because some behavior that can appear to reduce competition may have legitimate business purposes.

Key Concepts

oligopoly, p. 359
game theory, p. 359
collusion, p. 361
cartel, p. 361
Nash equilibrium, p. 362
prisoners' dilemma, p. 365
dominant strategy, p. 365

Questions for Review

1. If a group of sellers could form a cartel, what quantity and price would they try to set?
2. Compare the quantity and price of an oligopoly with that of a monopoly.
3. Compare the quantity and price of an oligopoly with that of a perfectly competitive market.
4. How does the number of firms in an oligopoly affect the outcome in the market?
5. What is the prisoners' dilemma, and what does it have to do with oligopoly?
6. Give two examples other than oligopoly that can be explained by the logic of the prisoners' dilemma.
7. What kinds of behavior do the antitrust laws prohibit?

Problems and Applications

1. A large share of the world supply of diamonds comes from Russia and South Africa. Suppose that the marginal cost of mining diamonds is constant at $1,000 per diamond and the demand for diamonds is described by the following schedule:

Price	Quantity
$8,000	5,000 diamonds
7,000	6,000
6,000	7,000
5,000	8,000
4,000	9,000
3,000	10,000
2,000	11,000
1,000	12,000

   a. If there were many suppliers of diamonds, what would the price and quantity be?
   b. If there were only one supplier of diamonds, what would the price and quantity be?
   c. If Russia and South Africa formed a cartel, what would the price and quantity be? If the countries split the market evenly, what would South Africa's production and profit be? What would happen to South Africa's profit if it increased its production by 1,000 while Russia stuck to the cartel agreement?
   d. Use your answers to part (c) to explain why cartel agreements are often not successful.

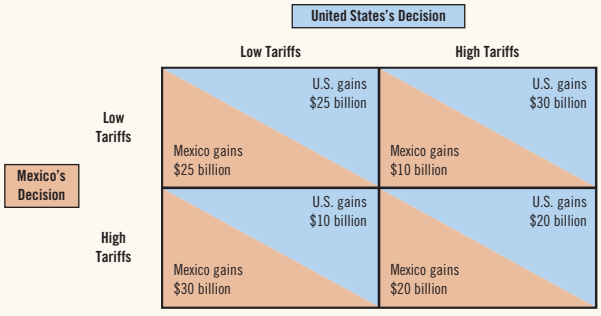
2. Some years ago, the *New York Times* reported that "the inability of OPEC to agree last week to cut production has sent the oil market into turmoil . . .

[leading to] the lowest price for domestic crude oil since June 1990."

a. Why were the members of OPEC trying to agree to cut production?
b. Why do you suppose OPEC was unable to agree on cutting production? Why did the oil market go into "turmoil" as a result?
c. The newspaper also noted OPEC's view "that producing nations outside the organization, like Norway and Britain, should do their share and cut production." What does the phrase "do their share" suggest about OPEC's desired relationship with Norway and Britain?

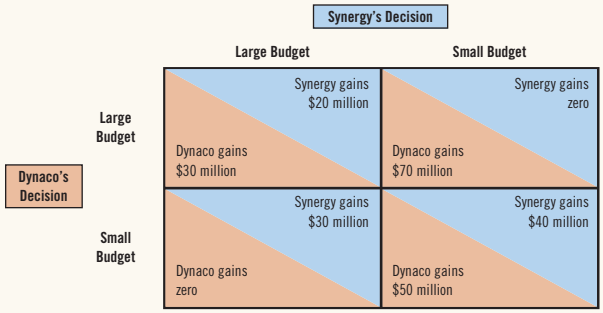
3. This chapter discusses companies that are oligopolists in the markets for the goods they sell. Many of the same ideas apply to companies that are oligopolists in the markets for the inputs they buy.
   a. If sellers who are oligopolists try to increase the price of goods they sell, what is the goal of buyers who are oligopolists?
   b. Major league baseball team owners have an oligopoly in the market for baseball players. What is the owners' goal regarding players' salaries? Why is this goal difficult to achieve?
   c. Baseball players went on strike in 1994 because they would not accept the salary cap that the owners wanted to impose. If the owners were already colluding over salaries, why did they feel the need for a salary cap?

4. Consider trade relations between the United States and Mexico. Assume that the leaders of the two countries believe the payoffs to alternative trade policies are as follows:



   a. What is the dominant strategy for the United States? For Mexico? Explain.
   b. Define **Nash equilibrium**. What is the Nash equilibrium for trade policy?
   c. In 1993, the U.S. Congress ratified the North American Free Trade Agreement, in which the United States and Mexico agreed to reduce trade barriers simultaneously. Do the perceived payoffs shown here justify this approach to trade policy? Explain.
   d. Based on your understanding of the gains from trade (discussed in Chapters 3 and 9), do you think that these payoffs actually reflect a nation's welfare under the four possible outcomes?

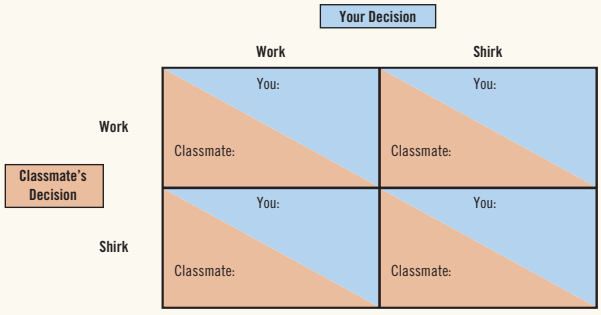
5. Synergy and Dynaco are the only two firms in a specific high-tech industry. They face the following payoff matrix as they determine the size of their research budget:



   a. Does Synergy have a dominant strategy? Explain.
   b. Does Dynaco have a dominant strategy? Explain.
   c. Is there a Nash equilibrium for this scenario? Explain. (Hint: Look closely at the definition of Nash equilibrium.)

6. You and a classmate are assigned a project on which you will receive one combined grade. You each want to receive a good grade, but you also want to avoid hard work. In particular, here is the situation:
   - If both of you work hard, you both get an A, which gives each of you 40 units of happiness.
   - If only one of you works hard, you both get a B, which gives each of you 30 units of happiness.
   - If neither of you works hard, you both get a D, which gives each of you 10 units of happiness.
   - Working hard costs 25 units of happiness.

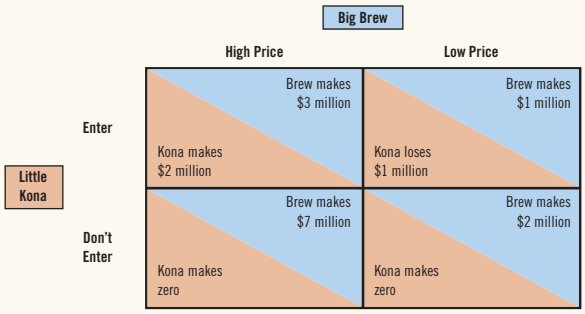
   a. Fill in the following payoff matrix:



   b. What is the likely outcome? Explain your answer.
   c. If you get this classmate as your partner on a series of projects throughout the year rather than only once, how might that change the outcome you predicted in part (b)?
   d. Another classmate cares more about good grades: She gets 50 units of happiness for a B and 80 units

of happiness for an A. If this classmate were your partner (but your preferences remained the same), how would your answers to parts (a) and (b) change? Which of the two classmates would you prefer as a partner? Would she also want you as a partner?

7. A case study in the chapter describes a phone conversation between the presidents of American Airlines and Braniff Airways. Let's use game theory to analyze the interaction between the two companies. Suppose that each company can charge either a high price for tickets or a low price. If one company charges $300, it earns low profit if the other company also charges $300 and high profit if the other company charges $600. On the other hand, if the company charges $600, it earns very low profit if the other company charges $300 and medium profit if the other company also charges $600.
   a. Draw the payoff matrix for this game.
   b. What is the Nash equilibrium in this game? Explain.
   c. Is there an outcome that would be better than the Nash equilibrium for both airlines? How could it be achieved? Who would lose if it were achieved?
8. Two athletes of equal ability are competing for a prize of $10,000. Each is deciding whether to take a dangerous performance-enhancing drug. If one athlete takes the drug and the other does not, the one who takes the drug wins the prize. If both or neither take the drug, they tie and split the prize. Taking the drug imposes health risks that are equivalent to a loss of X dollars.
   a. Draw a 2×2 payoff matrix describing the decisions the athletes face.
   b. For what X is taking the drug the Nash equilibrium?
   c. Does making the drug safer (that is, lowering X) make the athletes better or worse off? Explain.
9. Little Kona is a small coffee company that is considering entering a market dominated by Big Brew. Each company's profit depends on whether Little Kona enters and whether Big Brew sets a high price or a low price:



   a. Does either player in this game have a dominant strategy?
   b. Does your answer to part (a) help you figure out what the other player should do? What is the Nash equilibrium? Is there only one?
   c. Big Brew threatens Little Kona by saying, "If you enter, we're going to set a low price, so you had better stay out." Do you think Little Kona should believe the threat? Why or why not?
   d. If the two firms could collude and agree on how to split the total profits, what outcome would they pick?

QuickQuiz Answers

1. **d** 2. **c** 3. **a** 4. **d** 5. **c** 6. **b** 7. **d** 8. **b**

Chapter

19

The Markets for the Factors of Production

When you finish school, your income will be determined largely by what kind of job you take. If you work as a computer programmer, you are likely to earn more than if you work as a gas station attendant. This fact is puzzling. No law requires that computer programmers be paid more than gas station attendants. No ethical principle says that programmers are more deserving. What, then, determines which job pays the higher wage?

Your income is a small piece of a larger economic picture. In 2021, the total income of all U.S. residents, a statistic called **national income**, was about $20 trillion. People earned this income in various ways. Workers earned about two-thirds of it in the form of wages and fringe benefits, such as health insurance and pension contributions. The rest went to landowners and to the owners of **capital**—the economy's stock of equipment and structures—in the form of rent, profit, and interest. What determines how much goes to workers? To landowners? To the owners of capital? Why do some workers earn higher wages than others, some landowners higher rental income than others, and some capital owners greater profit than others? Why, in particular, do computer programmers earn more than gas station attendants?

SHIFT DRIVE/SHUTTERSTOCK.COM; ERDARKVE+/GETTY IMAGES

The answers to these questions, like most in economics, hinge on supply and demand. The supply and demand for labor, land, and capital determine the prices paid to workers, landowners, and capital owners. To understand why some people earn more than others, we need to look more deeply at the markets for the services they provide. We take up that task in this and the next two chapters.

factors of production the inputs used to produce goods and services

This chapter presents the basic theory of factor markets. As you may recall from Chapter 2, the **factors of production** are the inputs used to produce goods and services. Labor, land, and capital are the three most important factors of production. When a computer firm produces software, it uses programmers' time (labor), the physical space where its offices are located (land), and an office building and computer equipment (capital). When a gas station sells gas, it uses attendants' time (labor), the station's physical space (land), and gas tanks and pumps (capital).

Factor markets differ from the markets for goods and services analyzed in previous chapters in one important way: The demand for a factor of production is a **derived demand**. That is, a firm's demand for a factor of production is derived from its decision to supply a good in another market. For example, the demand for programmers is linked to the supply of software, and the demand for gas station attendants is linked to the supply of gasoline.

This chapter analyzes factor demand by considering how competitive, profit-maximizing firms decide how much of a factor to buy. The analysis begins by examining the demand for labor. Labor is the most important factor of production, as reflected by the fact that workers receive most of national income. Later in the chapter, the analysis is extended to the other factors of production.

The theory of factor markets developed in this chapter takes a large step toward explaining how the income of the U.S. economy is distributed among workers, landowners, and owners of capital. Chapter 20 builds on this analysis to examine in more detail why some workers earn more than others. Chapter 21 examines how much income inequality results from the functioning of factor markets and then considers the government's role in altering the income distribution.

19-1 The Demand for Labor

The forces of supply and demand govern most markets in the economy, and the labor market is no exception. This is illustrated in Figure 1. In panel (a), the supply and demand for apples determine the price of apples. In panel (b), the supply and demand for apple pickers determine the price, or wage, of apple pickers.

As we have noted, labor demand is a derived demand. Rather than being final goods ready to be consumed, most labor services are inputs into the production of other goods. To understand labor demand, we focus on the firms that hire the labor and use it to produce goods for sale. The link between the supply of goods and the demand for labor to produce them is crucial in determining equilibrium wages.

19-1a The Competitive, Profit-Maximizing Firm

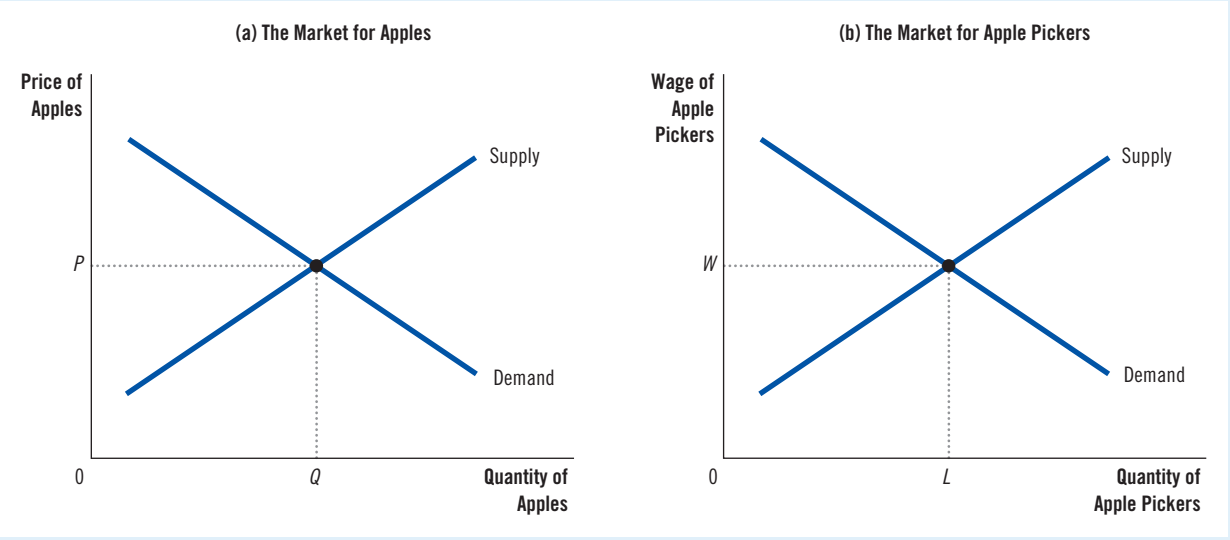
Consider how an apple producer decides what quantity of labor to demand. The firm owns an orchard and each week decides how many apple pickers to hire to harvest its crop. After the firm hires its workers, they pick the apples. The firm sells the apples, pays the workers, and keeps what's left as profit.

The theory developed in this chapter is based on two assumptions about the firm. First, the firm is **competitive** both in the market for apples (where it is a seller) and in the market for apple pickers (where it is a buyer). A competitive firm is a price taker. Because many firms sell apples and hire apple pickers, a single firm has little

Figure 1

The Versatility of Supply and Demand

The basic tools of supply and demand apply to goods and to labor services. Panel (a) shows how the supply and demand for apples determine the price of apples. Panel (b) shows how the supply and demand for apple pickers determine the wage of apple pickers.



influence over the price it gets for apples or the wage it pays apple pickers. The firm takes the price and wage as given by market conditions. It only has to decide how many apples to sell and how many workers to hire.

Second, the firm is **profit-maximizing**. It does not care about the number of workers it employs or the number of apples it produces, except insofar as they affect profit, which equals the total revenue from the sale of apples minus the total cost of producing them. The firm's supply of apples and its demand for workers are derived from its primary goal of maximizing profit.

19-1b The Production Function and the Marginal Product of Labor

To make its hiring decision, a firm considers how the size of its workforce affects its output. For the apple producer, the question is how the number of apple pickers affects the quantity of apples it can harvest and sell. Table 1 gives a numerical example. Column (1) shows the number of workers. Column (2) shows the quantity of apples the workers harvest each week.

These two columns of numbers describe the firm's ability to produce apples. Recall that economists use the term **production function** to describe the relationship between the quantity of the inputs used in production and the quantity of output from production. Here the "input" is the apple pickers, and the "output" is the apples. The other inputs—the trees themselves, the land, the firm's trucks and tractors, and so on—are held fixed for now. This firm's production function shows that if the firm hires 1 worker, that worker will pick 100 bushels of apples per week. If the firm hires 2 workers, the 2 workers together will pick 180 bushels per week. And so on.

production function the relationship between the quantity of inputs used to make a good and the quantity of output of that good

Table 1

How the Competitive Firm Decides How Much Labor to Hire

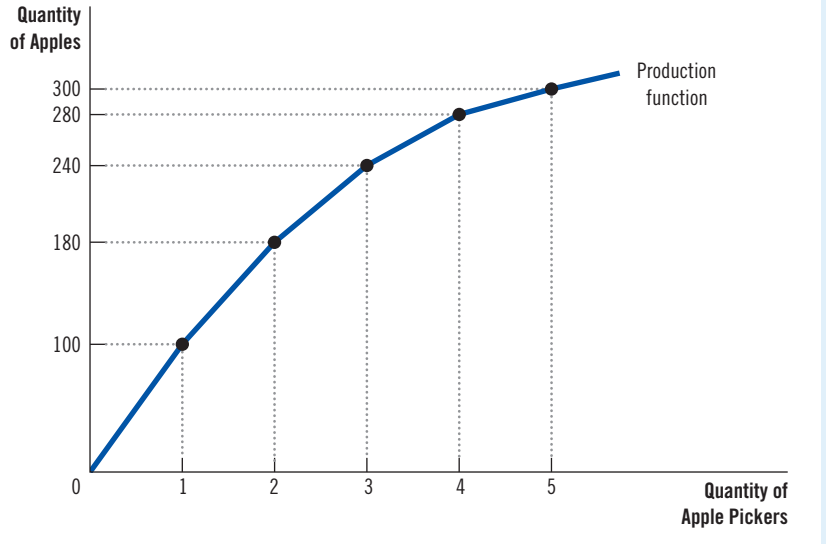
(1) Labor *L*	(2) Output *Q*	(3) Marginal Product of Labor *MPL* = $\Delta Q/\Delta L$	(4) Value of the Marginal Product of Labor *VMPL* = *P* × *MPL*	(5) Wage *W*	(6) Marginal Profit Δ*Profit* = *VMPL* − *W*
0 workers	0 bushels				
		100 bushels	$1,000	$500	$500
1	100				
		80	800	500	300
2	180				
		60	600	500	100
3	240				
		40	400	500	–100
4	280				
		20	200	500	–300
5	300				

Figure 2 graphs the data on labor and output presented in Table 1. The number of workers is on the horizontal axis, and the amount of output is on the vertical axis. This figure illustrates the production function.

Figure 2

The Production Function

The production function shows how an input into production (apple pickers) influences the output from production (apples). As the quantity of the input increases, the production function gets flatter, reflecting the property of diminishing marginal product.



One of the **Ten Principles of Economics** in Chapter 1 is that rational people think at the margin. This idea is the key to understanding how firms decide how much labor to hire. To take a step toward this decision, column (3) in Table 1 shows the **marginal product of labor**, the additional output produced by an additional unit of labor. When the firm increases the number of workers from 1 to 2, for example, the quantity of apples produced rises from 100 to 180 bushels. Therefore, the marginal product of the second worker is 80 bushels.

marginal product of labor
the increase in the amount of output from an additional unit of labor

Notice that as the number of workers increases, the marginal product of labor declines. That is, the production process exhibits **diminishing marginal product**. At first, when only a few workers are hired, they can pick the low-hanging fruit. As the number of workers increases, additional workers must climb higher up the ladders to find apples to pick. So as more workers are hired, each additional one contributes less to production. For this reason, the production function in Figure 2 becomes flatter as the number of workers rises.

diminishing marginal product
the property whereby the marginal product of an input declines as the quantity of the input increases

19-1c The Value of the Marginal Product and the Demand for Labor

Our profit-maximizing firm is concerned not about the apples themselves but rather about the money it can make by producing and selling them. As a result, when deciding how many workers to hire to pick apples, the firm considers how much profit each worker will bring in. Because profit is total revenue minus total cost, the profit from an additional worker is the worker's contribution to revenue minus the worker's wage.

To find the worker's contribution to revenue, we must convert the marginal product of labor (which is measured in bushels of apples) into the **value** of the marginal product (which is measured in dollars). We do this using the price of apples. If a bushel of apples sells for $10 and an additional worker produces 80 bushels of apples, then the worker produces $800 of revenue.

The **value of the marginal product** of any input is the marginal product of that input multiplied by the market price of the output. Column (4) in Table 1 shows the value of the marginal product of labor, assuming the price of apples is $10 per bushel. Because the market price is constant for a competitive firm while the marginal product declines with more workers, the value of the marginal product diminishes as the number of workers rises. Economists sometimes call this column of numbers the firm's **marginal revenue product**: It is the extra revenue the firm gets from hiring an additional unit of a factor of production.

value of the marginal product
the marginal product of an input times the price of the output

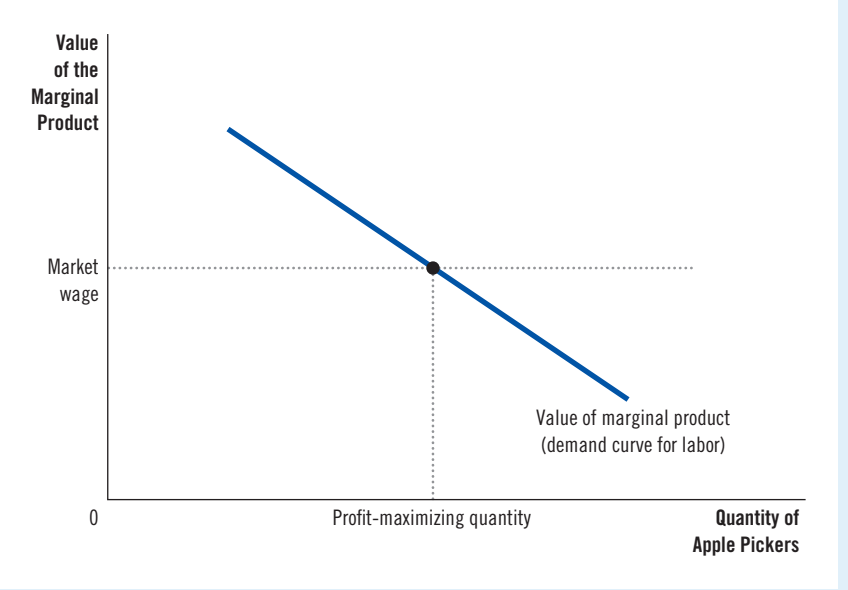
Now consider how many workers the firm will hire. Suppose that the market wage for apple pickers is $500 per week. In this case, as Table 1 shows, hiring the first worker is profitable: The first worker yields $1,000 in revenue and $500 in profit. Similarly, the second worker yields $800 in additional revenue and $300 in profit. The third yields $600 in additional revenue and $100 in profit. After the third worker, however, hiring workers is unprofitable. The fourth worker would generate only $400 of additional revenue. Because the worker's wage is $500, hiring the fourth worker would mean a $100 reduction in profit. The rational decision is clear: The firm hires 3 workers.

Figure 3 graphs the value of the marginal product. This curve slopes downward because the marginal product of labor diminishes as the number of workers rises. The figure also includes a horizontal line at the market wage. To maximize profit, the firm hires workers up to the point where these two curves cross. Below this level of employment, the value of the marginal product exceeds the wage, so hiring

Figure 3

The Value of the Marginal Product of Labor

This figure shows how the value of the marginal product (the marginal product times the price of the output) depends on the number of workers. The curve slopes downward because of diminishing marginal product. For a competitive, profit-maximizing firm, this value-of-marginal-product curve is also the firm's labor-demand curve.



another worker increases profit. Above this level of employment, the value of the marginal product is less than the wage, so the marginal worker is unprofitable. **A competitive, profit-maximizing firm hires workers up to the point at which the value of the marginal product of labor equals the wage.**

Based on the profit-maximizing hiring strategy for a competitive firm, we can offer a theory of labor demand. Recall that a firm's labor-demand curve tells us the quantity of labor that a firm decides to hire at any given wage. Figure 3 shows that the firm makes that decision by choosing the quantity of labor at which the value of the marginal product equals the wage. **As a result, the value-of-marginal-product curve is the labor-demand curve for a competitive, profit-maximizing firm.**

19-1d What Causes the Labor-Demand Curve to Shift?

The labor-demand curve reflects the value of the marginal product of labor. With this insight in mind, consider a few of the things that might cause the labor-demand curve to shift.

The Output Price The value of the marginal product is marginal product times the price of the firm's output. When the output price changes, the value of the marginal product changes, and the labor-demand curve shifts. An increase in the price of apples, for instance, raises the value of the marginal product of each worker who picks apples and, therefore, increases labor demand from the firms that supply apples. Conversely, a decrease in the price of apples reduces the value of the marginal product and decreases labor demand.

Technological Change Between 1960 and 2020, the output a typical U.S. worker produced in an hour rose by 236 percent. Why? The most important reason is

FYI

Input Demand and Output Supply: Two Sides of the Same Coin

Chapter 15 discussed how a competitive, profit-maximizing firm decides how much of its output to sell: It chooses the quantity of output at which the price of the good equals the marginal cost of production. We have just seen how such a firm decides how much labor to hire: It chooses the quantity of labor at which the wage equals the value of the marginal product. Because the production function links the quantity of inputs to the quantity of output, the firm's decision about input demand and its decision about output supply are two sides of the same coin.

To see this relationship more fully, consider how the marginal product of labor (*MPL*) and marginal cost (*MC*) are related. Suppose an additional worker costs $500 and has a marginal product of 50 bushels of apples. In this case, producing 50 more bushels costs the firm $500, so the marginal cost of a bushel is $500/50, or $10. More generally, if *W* is the wage, and an extra unit of labor produces *MPL* units of output, then the marginal cost of a unit of output is $MC = W/MPL$.

This analysis shows that diminishing marginal product is closely related to increasing marginal cost. When the apple orchard grows crowded with workers, each additional worker adds less to the production of apples (*MPL* falls). Similarly, when the apple firm is producing a large quantity of apples, the orchard is already crowded with workers, so it is more costly to produce an additional bushel of apples (*MC* rises).

Now consider our criterion for profit maximization. We determined that a profit-maximizing firm chooses the quantity of labor at which the value of the marginal product ($P \times MPL$) equals the wage (*W*). We can write this mathematically as:

$$P \times MPL = W.$$

Dividing both sides of this equation by *MPL* yields:

$$P = W/MPL.$$

We just noted that *W*/*MPL* equals marginal cost, *MC*. Therefore, we can substitute to obtain:

$$P = MC.$$

This equation states that the price of the firm's output equals the marginal cost of producing a unit of output. **When a competitive firm hires labor up to the point at which the value of the marginal product equals the wage, it also produces up to the point at which the price equals marginal cost.** The analysis of labor demand in this chapter is just another way of looking at the production decision discussed in Chapter 15. ■

technological progress: Scientists and engineers figured out new and better ways of doing things. This has profound implications for the labor market. Advances in technology usually raise the marginal product of labor, increasing the demand for labor and shifting the labor-demand curve to the right.

But technological change can also reduce labor demand. The invention of a cheap industrial robot, for instance, could reduce the marginal product of labor, shifting the labor-demand curve to the left. Economists call this a **labor-saving** technological change. History suggests, however, that most technological progress is instead **labor-augmenting**. For example, a carpenter with a nail gun is more productive than a carpenter with only a hammer. Labor-augmenting technological advances explain persistently rising employment in the face of rising wages: Even though wages (adjusted for inflation) increased by 201 percent from 1960 to 2020, firms roughly doubled the amount of labor they employed.

The Supply of Other Factors The quantity of one factor of production that is available can affect the marginal product of other factors. The productivity of apple pickers depends, for instance, on the availability of ladders. If the supply of ladders declines, the marginal product of apple pickers will decline as well, reducing the demand for apple pickers. We consider the linkage among the factors of production more fully later in the chapter.

QuickQuiz

1. Approximately what percentage of U.S. national income is paid to workers rather than to owners of capital and land?
   a. 25 percent
   b. 45 percent
   c. 65 percent
   d. 85 percent
2. If firms are competitive and profit-maximizing, the demand curve for labor is determined by
   a. the opportunity cost of workers' time.
   b. the value of the marginal product of labor.
   c. the value of the marginal product of capital.
   d. the ratio of the marginal product of labor to the marginal product of capital.
3. A bakery operating in competitive markets sells its output for $20 per cake and pays workers $10 per hour. To maximize profit, it should hire workers until the marginal product of labor is
   a. 1/2 cake per hour.
   b. 2 cakes per hour.
   c. 10 cakes per hour.
   d. 15 cakes per hour.

Answers are at the end of the chapter.

19-2 The Supply of Labor

Having analyzed labor demand, let's turn to the other side of the market and consider labor supply. A formal model of labor supply is included in Chapter 22, which develops the theory of household decision making. Here we informally discuss the decisions that lie behind the labor-supply curve.

19-2a The Trade-Off between Work and Leisure

One of the **Ten Principles of Economics** in Chapter 1 is that people face trade-offs. Probably no trade-off in a person's life is more important than the trade-off between work and leisure. The more hours you spend working, the fewer hours you have to enjoy dinner with friends, browse social media, or pursue your favorite hobby. The trade-off between labor and leisure lies behind the labor-supply curve.

PETER C. VEY/THE NEW YORKER COLLECTION/THE CARTOON BANK

"I really didn't enjoy working five days a week, fifty weeks a year for forty years, but I needed the money."

Another of the **Ten Principles of Economics** is that the cost of something is what you give up to get it. What do you give up to get an hour of leisure? You give up an hour of work, which in turn means an hour of wages. If your wage is $20 per hour, the opportunity cost of an hour of leisure is $20. And when you get a raise to $25 per hour, the opportunity cost of leisure increases.

The labor-supply curve reflects how workers' decisions about the labor-leisure trade-off respond to a change in that opportunity cost. An upward-sloping labor-supply curve means that an increase in the wage induces workers to increase the quantity of labor they supply. Because time is limited, more work means less leisure. That is, workers respond to the increase in the opportunity cost of leisure by taking less of it.

It is worth noting that the labor-supply curve need not be upward-sloping. Imagine you got that raise from $20 to $25 per hour. The opportunity cost of leisure is now greater, but you are also richer than you were before. You might decide that, with your extra wealth, you can now afford to enjoy more leisure. That is, at the higher wage, you might choose to work fewer hours. If so, your labor-supply curve would slope backward. Chapter 22 discusses this possibility

in terms of conflicting effects on your labor-supply decision, called the **income effect** and **substitution effect**. The income effect reflects the response of hours worked due to a change in a person's level of economic well-being, while the substitution effect reflects the response of hours worked due to a change in the opportunity cost of leisure. For now, let's put aside the possibility of backward-sloping labor supply. That is, we assume that the substitution effect dominates, so the labor-supply curve slopes upward.

19-2b What Causes the Labor-Supply Curve to Shift?

The labor-supply curve shifts whenever people change the amount they want to work at a given wage. Consider some of the events that might cause such a shift.

Changes in Preferences In 1950, 34 percent of women were employed at paid jobs or looking for work. By 2020, that number had risen to 56 percent. One of the many explanations for this development is changing preferences or attitudes toward work. In 1950, women routinely stayed at home and raised their children. Today, the typical family size is smaller, and more mothers choose to work. One result is an increase in the supply of labor.

Changes in Alternative Opportunities The supply of labor in any one labor market depends on the opportunities available in other labor markets. If the wage earned by pear pickers suddenly rises, some apple pickers may choose to switch occupations, causing the supply of labor in the market for apple pickers to fall.

Immigration The movement of workers from region to region or country to country is an important source of shifts in labor supply. When migrant workers come north for the autumn harvest, the supply of labor increases in apple orchards, but it declines in orange-processing plants in the south. When immigrants come to the United States, the supply of labor increases in the United States and falls in the immigrants' home countries. Much of the policy debate about immigration centers on its effect on labor supply and equilibrium wages.

QuickQuiz

4. Who has a greater opportunity cost of leisure—janitors or surgeons?
   a. janitors because their wages are lower
   b. surgeons because their wages are higher
   c. whoever has the greater income effect
   d. whoever has the greater substitution effect
5. A person works more hours at a higher wage if the substitution effect
   a. equals zero.
   b. equals the income effect.
   c. is smaller than the income effect.
   d. is larger than the income effect.
6. Which of the following events will shift the labor supply curve to the right?
   a. More dads leave the workforce to spend time raising children.
   b. Great new video games are introduced, enhancing the value of leisure.
   c. Relaxed immigration laws allow more workers to come in from abroad.
   d. Government benefits for the retired are increased.

Answers are at the end of the chapter.

19-3 Equilibrium in the Labor Market

So far, we have established two facts about how wages are determined in competitive labor markets:

- The wage adjusts to balance the supply and demand for labor.
- The wage equals the value of the marginal product of labor.

At first, it might seem surprising that the wage can do both things at once. In fact, there is no real puzzle here, but understanding why is an important step toward understanding wage determination.

Figure 4 shows the labor market in equilibrium. The wage and the quantity of labor have adjusted to balance supply and demand. When the market is in this equilibrium, each firm has bought as much labor as it finds profitable at the equilibrium wage. That is, each firm has followed the rule for profit maximization: It has hired workers until the value of the marginal product equals the wage. Hence, the wage must equal the value of the marginal product of labor once it has brought supply and demand into equilibrium.

This brings us to an important lesson: **Any event that changes the supply or demand for labor must change the equilibrium wage and the value of the marginal product by the same amount because these must always be equal.** To see how this works, let's consider some events that shift these curves.

19-3a Shifts in Labor Supply

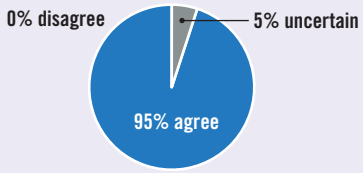
Suppose that immigration increases the number of workers willing to pick apples. As Figure 5 shows, the supply of labor shifts to the right from S_1 to S_2. At the initial wage W_1, the quantity of labor supplied now exceeds the quantity demanded. This surplus of labor puts downward pressure on the wage of apple pickers, and the fall in the wage from W_1 to W_2 makes it profitable for firms to hire more workers. As the number of workers employed in each apple orchard rises, the marginal product of a worker falls, and so does the value of the marginal product. In the new equilibrium, both the wage and the value of the marginal product of labor are lower than they were before the influx of new workers.

An episode from Israel, studied by the economist Joshua Angrist, illustrates how a shift in labor supply can alter the equilibrium in a labor market. During most of the 1980s, many thousands of Palestinians regularly commuted from their homes in the Israeli-occupied West Bank and Gaza Strip to jobs in Israel, primarily in the construction and agriculture industries. In 1988, however, political unrest in these occupied areas induced the Israeli government to take steps that, as a by-product, reduced this supply of workers. Curfews were imposed, work permits were checked more thoroughly, and a ban on overnight stays of Palestinians in Israel was enforced more rigorously. The economic impact of these steps was exactly as theory predicts: The number of Palestinians with jobs in Israel fell by half, while those who

Ask the Experts — Immigration

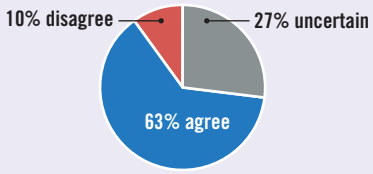
"The average U.S. citizen would be better off if a larger number of highly educated foreign workers were legally allowed to immigrate to the U.S. each year."

What do economists say?



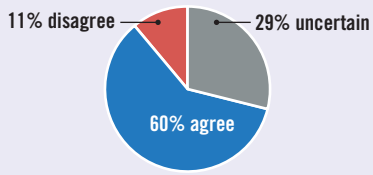
"The average U.S. citizen would be better off if a larger number of low-skilled foreign workers were legally allowed to enter the U.S. each year."

What do economists say?



"Unless they were compensated by others, many low-skilled American workers would be substantially worse off if a larger number of low-skilled foreign workers were legally allowed to enter the U.S. each year."

What do economists say?



Source: IGM Economic Experts Panel, February 12, 2013, December 10, 2013.

Figure 4

Equilibrium in a Labor Market

Like all prices, the price of labor (the wage) depends on supply and demand. Because the demand curve reflects the value of the marginal product of labor, in equilibrium, workers receive the value of their marginal contribution to the production of goods and services.

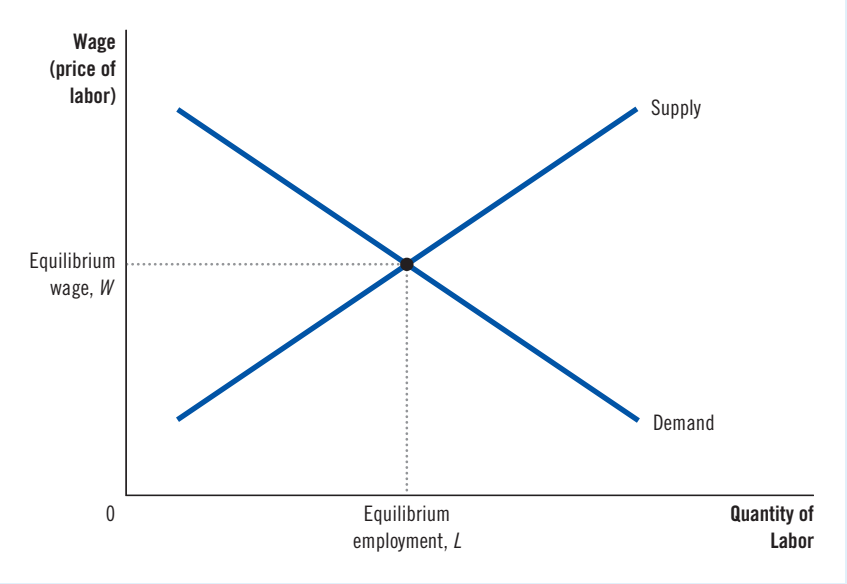
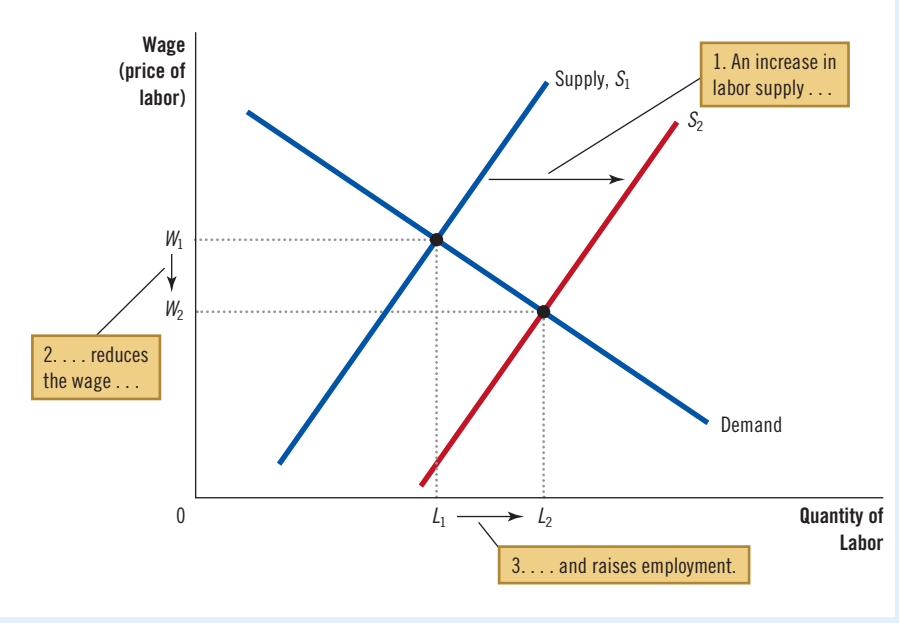


Figure 5

A Shift in Labor Supply

When labor supply increases from S_1 to S_2, perhaps because of an immigration wave of new workers, the equilibrium wage falls from W_1 to W_2. At this lower wage, firms hire more labor, so employment rises from L_1 to L_2. The change in the wage reflects a change in the value of the marginal product of labor: With more workers, the added output from an extra worker is smaller.



continued to work in Israel enjoyed wage increases of about 50 percent. With a reduced number of Palestinian workers in Israel, the value of the marginal product of the remaining workers was much higher.

When considering the economics of immigration, keep in mind that the economy consists not of a single labor market but of a variety of labor markets for different

kinds of workers. A wave of immigration may lower wages in those labor markets in which the new immigrants seek work, but it could have the opposite effect in other labor markets. For example, if the new immigrants look for jobs as apple pickers, the supply of apple pickers increases, and the wage of apple pickers declines. But suppose the new immigrants are physicians who use some of their income to buy apples. In this case, the wave of immigration increases the **supply** of physicians but also increases the **demand** for apples and thus apple pickers. As a result, the wages of physicians decline, and the wages of apple pickers rise. The linkages among various markets—sometimes called **general equilibrium effects**—make analyzing the full effect of immigration more complex than it first appears.

The Immigration Debate

Pat Paulsen was a comedian from the 1960s to the 1990s who, every four years, conducted a faux campaign for president. "All the problems we face in the United States today," Mr. Paulsen would say, "can be traced to an unenlightened immigration policy on the part of the American Indian."

That quip contains a deep truth. Most Americans today are beneficiaries of a policy that welcomed their ancestors when they arrived at the border. But that does not stop immigration from being a divisive political issue. One reason for this divisiveness is immigration's economic impact.

The welfare effects of immigration can be seen through the lens of international trade. Recall from Chapter 9 that when a nation allows a good to be imported, the price falls. Consumers of the good are better off, and domestic producers are worse off. But the increases in consumer surplus exceed the losses in producer surplus, so total surplus rises. In other words, imports expand the economic pie but leave some with a smaller slice.

Immigration entails an import of labor services. The consumers of these services are the firms that hire the labor and their customers, both of which benefit when immigration increases labor supply. The domestic producers in this case are the native workers who are now competing with new workers from abroad and, as a result, experience reduced earnings. The net benefit to the economy is positive, but that fact may not offer much comfort to those with depressed incomes.

How large are the labor market effects of immigration? The economist George Borjas estimates that the increased total surplus from immigration into the United States is about 0.25 percent of U.S. national income annually. In addition, about 2.5 percent of national income is redistributed from the native losers (the workers who compete with the immigrants) to the native winners (those who consume these labor services). Not included in these numbers are the benefits to the immigrants themselves, whose earnings in the United States far exceed what they would have been had they stayed in their home countries.

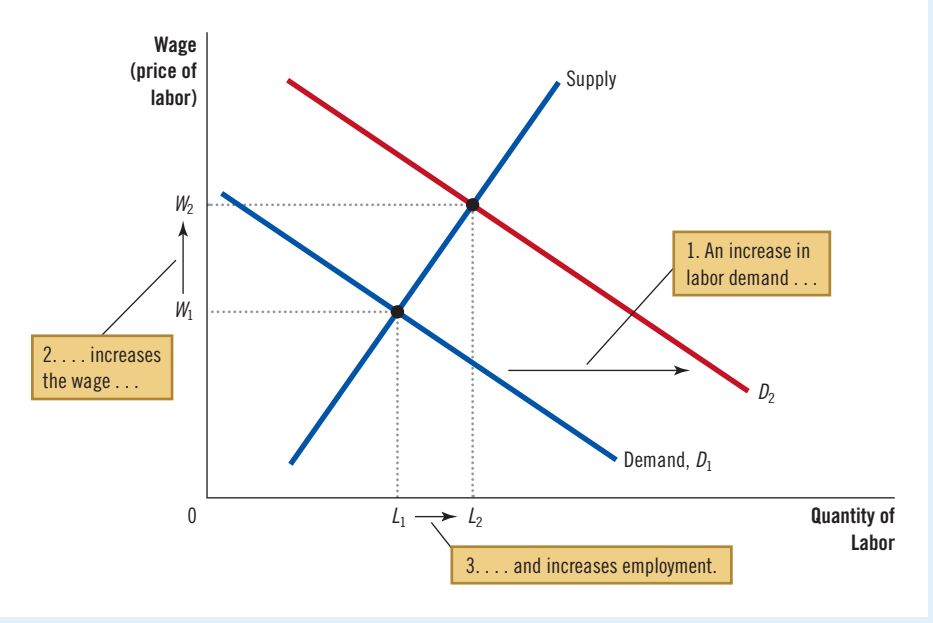
Some economists have proposed ways to distribute the gains from immigration more equitably. For example, immigrants could be subject to a special tax levied on either them or their employers. The revenue could be used to reduce the tax burden on native workers. If native workers shared more of the benefits from immigration, they might be more likely to welcome it.

The debate over immigration is not just about economics, however. It also has a powerful, emotional element that concerns cultural and national identity. But most Americans would do well to remember how lucky they are that the American Indian did not pursue the enlightened immigration policy suggested by Mr. Paulsen. ●

Figure 6

A Shift in Labor Demand

When labor demand increases from D_1 to D_2, perhaps because of an increase in the price of the firm's output, the equilibrium wage rises from W_1 to W_2, and employment rises from L_1 to L_2. The change in the wage reflects a change in the value of the marginal product of labor: With a higher output price, the added output from an extra worker is more valuable.



19-3b Shifts in Labor Demand

Now suppose that an increase in the popularity of apples causes their price to rise. This price increase does not change the marginal product of labor for any given number of workers, but it does raise the **value** of the marginal product. With a higher price for apples, hiring more apple pickers is now profitable. As Figure 6 shows, when the demand for labor shifts to the right from D_1 to D_2, the equilibrium wage rises from W_1 to W_2, and equilibrium employment rises from L_1 to L_2. Once again, the wage and the value of the marginal product of labor move together.

This analysis shows that prosperity for firms in an industry is often linked to prosperity for workers in that industry. When the price of apples rises, apple producers make a greater profit, and apple pickers earn higher wages. When the price of apples falls, apple producers earn a smaller profit, and apple pickers earn lower wages. This lesson is well known to workers in industries with highly volatile prices. Workers in oil fields, for instance, know from experience that their earnings are closely linked to the world price of crude oil.

From these examples, you should now have a good understanding of how wages are set in competitive labor markets. Labor supply and labor demand together determine the equilibrium wage, and shifts in the supply or demand curve for labor cause the equilibrium wage to change. At the same time, profit maximization by the firms that demand labor ensures that the equilibrium wage always equals the value of the marginal product of labor.

Case Study

Productivity and Wages

One of the **Ten Principles of Economics** in Chapter 1 is that a country's standard of living depends on its ability to produce goods and services. This principle is evident in the market for labor. Our analysis of labor demand shows that wages equal productivity as measured by

Figure 7

Growth in Productivity and Real Wages

When productivity grows rapidly, so do real wages. And when productivity growth is more modest, real wage growth is as well.

Source: Bureau of Labor Statistics. Growth in productivity is measured here as the annualized rate of change in output per hour in the nonfarm business sector. Growth in real wages is measured as the annualized change in compensation per hour in the nonfarm business sector divided by the price deflator for that sector. These productivity data measure average productivity—the quantity of output divided by the quantity of labor—rather than marginal productivity, but average and marginal productivity are thought to move closely together.

the value of the marginal product of labor. Put simply, highly productive workers tend to be highly paid, and less productive workers typically earn less.

This lesson is key to understanding why workers today are generally better off than workers in previous generations. From 1960 to 2020, economy-wide productivity as measured by output per hour of work grew about 2.0 percent per year. Real wages (that is, wages adjusted for inflation) grew at 1.9 percent per year—almost the same rate. This change in wages may be too small to notice year to year, but it compounds over many years. With a growth rate of 2 percent per year, productivity and real wages double about every 35 years.

The link between productivity and real wages appears again when we examine various historical periods with different productivity experiences, as shown in Figure 7. When productivity grows rapidly, real wages rise quickly. When productivity grows slowly, the increase in real wages is more modest. The most recent period, 2010 to 2020, exhibited low growth in both productivity and real wages.

The bottom line: Both theory and history confirm the close connection between productivity and real wages. ●

QuickQuiz

7. A technological advance that increases the marginal product of labor shifts the labor- _____ curve to the _____.
   a. demand; left
   b. demand; right
   c. supply; left
   d. supply; right

8. Around 1973, the U.S. economy experienced a significant _____ in productivity growth, coupled with a _____ in the growth of real wages.
   a. pickup; pickup
   b. pickup; slowdown
   c. slowdown; pickup
   d. slowdown; slowdown

Answers are at the end of the chapter.

19-4 The Other Factors of Production: Land and Capital

We have seen how firms decide how much labor to hire and how these decisions affect workers' wages. As firms hire workers, they also decide about other inputs to production. For example, our apple-producing firm might have to choose the size of its orchard and the number of ladders for apple pickers. Think of the firm's factors of production as falling into three categories: labor, land, and capital.

The meanings of the terms **labor** and **land** are clear, but the definition of **capital** is tricky. Economists use the term **capital** to refer to the stock of equipment and structures used for production. That is, capital represents the accumulation of goods produced in the past that are being used in the present to produce new goods and services. For our apple firm, the capital stock includes the ladders used to climb the trees, the trucks used to transport the apples, the buildings used to store the apples, and even the apple trees themselves.

capital
the equipment and structures used to produce goods and services

19-4a Equilibrium in the Markets for Land and Capital

What determines how much the owners of land and capital earn for their contribution to production? Before answering this question, we need to distinguish between two prices: the purchase price and the rental price. The **purchase price** of land or capital is the price paid to own it indefinitely. The **rental price** is the price paid to use that factor for a limited period. Keep this distinction in mind because, as we will see, these prices are determined by somewhat different economic forces.

Here's the key insight: The theory of factor demand that we developed for the labor market also applies to the markets for land and capital. Because the wage is the rental price of labor, much of what we have learned about wage determination is relevant to the rental prices of land and capital. As Figure 8 illustrates, the rental

Figure 8
The Markets for Land and Capital

Supply and demand determine the compensation paid to the owners of land, as shown in panel (a), and the compensation paid to the owners of capital, as shown in panel (b). The demand for each factor, in turn, depends on the value of its marginal product.

FYI

What Is Capital Income?

Labor income is easy to understand: It's the paycheck that workers get from their employers. The income earned by capital is less obvious.

Our analysis implicitly assumed that households own the economy's stock of capital—ladders, drill presses, warehouses, and so on—and rent it to the firms that use it. Capital income, in this case, is the rent that households receive for the use of their capital. This assumption simplified our analysis of how capital owners are compensated, but it is not entirely realistic. In fact, firms usually own the capital they use and receive the earnings from this capital.

These earnings from capital are eventually paid to households in various ways. Some of the earnings are paid as interest to households that have lent money to firms. Bondholders and bank depositors are two examples. When you receive interest on your bank account, that income is part of the economy's capital income.

In addition, some of the earnings from capital are paid to households as dividends. Dividends are payments by a firm to its stockholders. A stockholder (or shareholder) is a person who has bought a share in the firm's ownership and, therefore, is entitled to a portion of its profits.

A firm does not have to pay out all its earnings to households in the form of interest and dividends. Instead, it can keep some of its earnings within the firm. Retained earnings can be used to add to the firm's cash position or to buy additional capital. Unlike dividends, retained earnings do not yield a direct cash payment to the firm's stockholders, but the stockholders benefit from them nonetheless. Because retained earnings increase the firm's assets, they tend to increase the value of the firm's stock.

These institutional details are important, but they do not alter our conclusion about the income earned by the owners of capital. Capital is paid according to the value of its marginal product, regardless of whether this income is transmitted to households in the form of interest or dividends or whether it is kept within firms as retained earnings. ■

price of land, shown in panel (a), and the rental price of capital, shown in panel (b), are determined by supply and demand. What's more, the demands for land and capital are determined just like the demand for labor. That is, when the apple-producing firm decides how much land and how many ladders to rent, it follows the same logic as it does when deciding how many workers to hire. For both land and capital, the firm increases the quantity rented until the value of the factor's marginal product equals the factor's price. The demand curve for each factor reflects that factor's marginal productivity.

This theory can now explain how much income goes to labor, landowners, and the owners of capital. As long as the firms using the factors of production are competitive and profit-maximizing, each factor's rental price must equal the value of its marginal product. **Labor, land, and capital all earn the value of their marginal contributions to the production process.**

Now consider the purchase price of land and capital. The rental price and purchase price are related: Buyers are willing to pay more for a piece of land or capital if it produces a valuable stream of rental income. And the equilibrium rental income at any point in time equals the value of that factor's marginal product. As a result, the equilibrium purchase price of a piece of land or capital depends not only on the current value of the marginal product but also on the value of the marginal product expected to prevail in the future.

19-4b Linkages among the Factors of Production

In competitive factor markets, the price paid for any factor of production—labor, land, or capital—equals the value of its marginal product. The marginal product of any factor, in turn, depends on the quantity of that factor that is available. Because of diminishing marginal product, a factor in abundant supply has a low marginal

product and a low price, and a factor in scarce supply has a high marginal product and a high price. When the supply of a factor falls, its equilibrium price rises.

But when the supply of any factor changes, the effects are not limited to the market for that factor. In most situations, factors of production are used together in a way that makes the productivity of each factor depend on the quantities of the other factors available for use in the production process. As a result, when some event changes the supply of any one factor of production, it will typically affect not only the earnings of that factor but also the earnings of all the other factors as well.

For example, suppose a hurricane destroys many of the ladders that workers use to pick apples. (To keep things simple, imagine that the storm miraculously leaves the orchards intact.) What happens to the earnings of the various factors of production? Most obviously, when the supply of ladders falls, the equilibrium rental price of ladders rises. Those owners lucky enough to have avoided damage to their ladders now earn a higher return when they rent them out to the firms that produce apples.

Yet the effects of this event do not stop at the ladder market. Because there are fewer ladders, the workers who pick apples can't perform their jobs as efficiently. In other words, the marginal product of labor declines. The reduction in the supply of ladders reduces the demand for the labor of apple pickers, and this shift in demand causes the equilibrium wage to fall.

This story shows a general lesson: **An event that changes the supply of any factor of production can alter the earnings of all the factors.** The change in earnings of any factor can be found by analyzing the impact of the event on the value of the marginal product of that factor.

BETTMANN/GETTY IMAGES

Workers who survived the plague were lucky in more ways than one.

Case Study

The Economics of the Black Death

In 14th-century Europe, the bubonic plague wiped out about one-third of the population within a few years—a vastly more calamitous event than even the tragic Covid pandemic of 2020 and 2021, which killed less than 1 percent of the population. This event, called the **Black Death**, provides a grisly natural experiment to test the theory of factor markets that we have just developed. Consider the effects of the Black Death on those who were lucky enough to survive. What do you think happened to the wages earned by workers and the rents earned by landowners?

To answer this question, let's examine the effects of a reduced population on the marginal product of labor and the marginal product of land. With a smaller supply of workers, the marginal product of labor rises. This is diminishing marginal product working in reverse. We would, therefore, expect the Black Death to raise wages.

Because land and labor are used together in production, a smaller supply of workers also affects the market for land, the other major factor of production in medieval Europe. With fewer workers available to farm the land, an additional unit of land produced less additional output. This decline in the marginal product of land would be expected to reduce rents.

Both theoretical predictions conform with the historical evidence. Wages approximately doubled during this period, and rents declined 50 percent or more. For survivors, the Black Death led to economic prosperity for the peasant classes and reduced incomes for the landed classes. ●

QuickQuiz

9. A bakery operating in competitive markets sells its output for $20 per cake and rents ovens at $30 per hour. To maximize profit, it should rent ovens until the marginal product of an oven is
   a. 2/3 cake per hour.
   b. 3/2 cakes per hour.
   c. 10 cakes per hour.
   d. 25 cakes per hour.

10. A storm destroys several factories, reducing the stock of capital. What effect does this event have on factor markets?
    a. Wages and the rental price of capital both rise.
    b. Wages and the rental price of capital both fall.
    c. Wages rise, and the rental price of capital falls.
    d. Wages fall, and the rental price of capital rises.

Answers are at the end of the chapter.

19-5 Conclusion

This chapter has explained how labor, land, and capital are compensated for the roles they play in the production process. The theory developed here is called the **neoclassical theory of distribution**. According to the neoclassical theory, the amount paid to each factor of production depends on the supply and demand for that factor. The demand, in turn, depends on that factor's marginal productivity. In equilibrium, each factor of production earns the value of its marginal contribution to the production of goods and services.

The neoclassical theory of distribution is widely accepted. Most economists begin with it when trying to explain how the U.S. economy's $20 trillion of income is distributed among the economy's various members. The next two chapters consider the distribution of income in more detail. The neoclassical theory provides the framework for that discussion.

Even at this point, you can use the theory to answer the question that began this chapter: Why are computer programmers paid more than gas station attendants? It is because programmers can produce a good of greater market value than can gas station attendants. People are willing to pay dearly for a good video game, but they are willing to pay little to have their gas pumped and windshield washed. The wages of these workers reflect the market prices of the goods they produce. If people suddenly got tired of using computers and decided to spend more time driving, the prices of these goods would change and so would the equilibrium wages of these workers.

Chapter in a Nutshell

- The economy's income is distributed in the markets for the factors of production. The three most important factors are labor, land, and capital.
- The demand for factors, such as labor, is a derived demand that comes from firms that use the factors to produce goods and services. Competitive, profit-maximizing firms hire each factor up to the point at which the value of the factor's marginal product equals its price.
- The supply of labor arises from individuals' trade-off between work and leisure. An upward-sloping labor-supply curve means that people respond to an increase in the wage by working more hours and enjoying less leisure.
- In competitive factor markets, the price paid to each factor adjusts to balance supply and demand. Because factor demand reflects the value of the factor's marginal product, in equilibrium, each factor is compensated according to its marginal contribution to the production of goods and services.
- Because factors of production are used together, the marginal product of any one factor depends on the available quantities of all factors. A change in the supply of one factor alters the equilibrium earnings of all of them.

Key Concepts

factors of production, p. 382
production function, p. 383
marginal product of labor, p. 385
diminishing marginal product, p. 385
value of the marginal product, p. 385
capital, p. 395

Questions for Review

1. Explain how a firm's production function is related to its marginal product of labor, how a firm's marginal product of labor is related to the value of its marginal product, and how a firm's value of marginal product is related to its demand for labor.
2. Give two examples of events that could shift the demand for labor, and explain why they do so.
3. Give two examples of events that could shift the supply of labor, and explain why they do so.
4. Explain how the wage can adjust to balance the supply and demand for labor while simultaneously equaling the value of the marginal product of labor.
5. If the population of the United States suddenly grew because of a large wave of immigration, what would happen to wages? What would happen to the rents earned by the owners of land and capital?

Problems and Applications

1. Suppose that the president proposes a new law aimed at reducing healthcare costs: All Americans are required to eat one apple daily.
   a. How would this apple-a-day law affect the demand and equilibrium price of apples?
   b. How would the law affect the marginal product and the value of the marginal product of apple pickers?
   c. How would the law affect the demand and equilibrium wage for apple pickers?
2. Show the effect of each of the following events on the market for labor in the computer manufacturing industry.
   a. Congress buys personal computers for all U.S. college students.
   b. More college students major in engineering and computer science.
   c. Computer firms build new manufacturing plants.
3. Suppose that labor is the only input used by a perfectly competitive firm. The firm's production function is as follows:

Days of Labor	Units of Output
0 days	0 units
1	7
2	13
3	19
4	25
5	28
6	29
7	29

   a. Calculate the marginal product of each additional worker.
   b. Each unit of output sells for $10. Calculate the value of the marginal product of each worker.
   c. Compute the demand schedule showing the number of workers hired for all wages from zero to $100 a day.
   d. Graph the firm's labor-demand curve.
   e. What happens to this demand curve if the price of output rises from $10 to $12 per unit?
4. Smiling Cow Dairy can sell all the milk it wants for $4 a gallon, and it can rent all the robots it wants to milk the cows at a capital rental price of $100 a day. It faces the following production schedule:

Number of Robots	Total Product
0	0 gallons
1	50
2	85
3	115
4	140
5	150
6	155

a. In what kind of market structure does the firm sell its output? How can you tell?
b. In what kind of market structure does the firm rent robots? How can you tell?
c. Calculate the marginal product and the value of the marginal product of each additional robot.
d. How many robots should the firm rent? Explain.

5. The nation of Ectenia has 20 competitive apple orchards, all of which sell apples at the world price of $2 per apple. The following equations describe the production function and the marginal product of labor in each orchard:

$$Q = 100L - L^2$$
$$MPL = 100 - 2L$$

where Q is the number of apples produced in a day, L is the number of workers, and MPL is the marginal product of labor.
a. What is each orchard's labor demand as a function of the daily wage W? What is the market's labor demand?
b. Ectenia has 200 workers who supply their labor inelastically. Solve for the wage W. How many workers does each orchard hire? How much profit does each orchard owner make?
c. Calculate what happens to the income of workers and orchard owners if the world price doubles to $4 per apple.
d. Now suppose that the price is back at $2 per apple, but a hurricane destroys half the orchards. Calculate how the hurricane affects the income of each worker and of each remaining orchard owner. What happens to the income of Ectenia as a whole?

6. Your enterprising uncle opens a sandwich shop that employs 7 people. The employees are paid $12 per hour, and a sandwich sells for $6. If your uncle is maximizing his profit, what is the value of the marginal product of the last worker he hired? What is that worker's marginal product?

7. Leadbelly Co. sells pencils in a perfectly competitive product market and hires workers in a perfectly competitive labor market. Assume that the market wage rate for workers is $150 per day.
a. What rule should Leadbelly follow to hire the profit-maximizing amount of labor?
b. At the profit-maximizing level of output, the marginal product of the last worker hired is 30 boxes of pencils per day. Calculate the price of a box of pencils.
c. Draw a diagram of the labor market for pencil workers (as in Figure 4 of this chapter) next to a diagram of the labor supply and demand for Leadbelly Co. (as in Figure 3). Label the equilibrium wage and quantity of labor for both the market and the firm. How are these diagrams related?
d. Suppose some pencil workers switch to jobs in the growing computer industry. On the side-by-side diagrams from part (c), show how this change affects the equilibrium wage and quantity of labor both for the pencil market and for Leadbelly. How does this change affect the marginal product of labor at Leadbelly?

8. Sometimes, laws require firms to give workers certain fringe benefits, such as health insurance or paid parental leave. Let's consider the effects of such a policy on the labor market.
a. Suppose that a law requires firms to give each worker $3 of fringe benefits for every hour that the worker is employed by the firm. How does this law affect the marginal profit that a firm earns from each worker at a given cash wage? How does the law affect the demand curve for labor? Draw your answer on a graph with the cash wage on the vertical axis.
b. If there is no change in labor supply, how would this law affect employment and wages?
c. Why might the labor-supply curve shift in response to this law? Would this shift in labor supply raise or lower the impact of the law on wages and employment?
d. As discussed in Chapter 6, minimum-wage laws keep the wages of some workers, particularly the unskilled and inexperienced, above the equilibrium level. What effect would a fringe-benefit mandate have for these workers?

9. Some economists believe that the U.S. economy as a whole can be modeled with the following production function, called the **Cobb–Douglas production function**:

$$Y = AK^{1/3}L^{2/3},$$

where Y is the amount of output, K is the amount of capital, L is the amount of labor, and A is a parameter that measures the state of technology. For this production function, the marginal product of labor is

$$MPL = (2/3)\, A(K/L)^{1/3}.$$

Suppose that the price of output P is 2, A is 3, K is 1,000,000, and L is 1,000. The labor market is competitive, so labor is paid the value of its marginal product.

a. Calculate the amount of output produced Y and the dollar value of output PY.
b. Calculate the wage W and the real wage W/P. (Note: The wage is labor compensation measured in dollars, whereas the real wage is labor compensation measured in units of output.)
c. Calculate the labor share (the fraction of the value of output that is paid to labor), which is $(WL)/(PY)$.
d. Calculate what happens to output Y, the wage W, the real wage W/P, and the labor share $(WL)/(PY)$ in each of the following scenarios:
   i. Inflation increases P from 2 to 3.
   ii. Technological progress increases A from 3 to 9.
   iii. Capital accumulation increases K from 1,000,000 to 8,000,000.
   iv. A plague decreases L from 1,000 to 125.
e. Despite many changes in the U.S. economy over time, the labor share has been relatively stable. Is this observation consistent with the Cobb–Douglas production function? Explain.

QuickQuiz Answers

1. **c** 2. **b** 3. **a** 4. **b** 5. **d** 6. **c** 7. **b** 8. **d** 9. **b** 10. **d**

Chapter

20

Earnings and Discrimination

In the United States in 2020, fast-food cooks earned about $24,000 a year, high school teachers about $67,000, family physicians about $214,000, and the chief executives of the largest companies about $12 million. Meanwhile, according to *Billboard* magazine, the superstar Taylor Swift earned $24 million, making her the year's best-paid musician. These vast differences in earnings have enormous implications. They explain why some people live in mansions, ride in private jets, and vacation on their own islands, while others live in small apartments, ride the bus, and don't take much of a vacation at all.

Why do earnings vary so much? Chapter 19, which developed the basic neoclassical theory of the labor market, offered an answer. It said that wages are governed by supply and demand, like so much else in the economy. Labor demand reflects the marginal productivity of labor, and in equilibrium, workers are paid the value of their marginal contribution to the production of goods and services.

This theory of the labor market, though widely accepted by economists, is only the beginning of the story. To explain disparities in earnings, we must go beyond this general framework. This chapter examines more precisely what determines the supply and demand for different types of labor and why, in some cases, wages depart from their equilibrium levels. It also considers how discrimination can influence labor-market outcomes.

AZMANJAKA/E+/GETTY IMAGES; CARLOS FLORES/EYEEM/GETTY IMAGES

20-1 What Determines Wages?

Let's first consider how the characteristics of jobs and workers affect labor supply, labor demand, and equilibrium wages.

20-1a Compensating Differentials

DANA FRADON/CARTOON COLLECTIONS

"On the one hand, I know I could make more money if I left public service for the private sector, but, on the other hand, I couldn't chop off heads."

When a worker is deciding whether to take a job, the wage is only one of many factors to consider. Some jobs are easy, fun, and safe. Others are hard, boring, and even dangerous. The more a job appeals to people, the more people are willing to do it at a given wage. In other words, the supply of labor is greater for easy, fun, and safe jobs than for hard, boring, and dangerous ones. As a result, other things being equal, appealing jobs will tend to have lower equilibrium wages than less appealing ones.

Imagine you are looking for a summer job in a beach community. Two jobs are available: beach-badge checker and garbage collector. The beach-badge checkers take leisurely strolls near the water during the day and ensure that the tourists have bought the required permits. The garbage collectors wake up before dawn and drive dirty, noisy trucks around town to pick up garbage. Which job would you want? If the wages were the same, most people would prefer the job on the beach. To induce people to become garbage collectors, the town must offer higher wages to garbage collectors than to beach-badge checkers.

Economists use the term **compensating differential** to refer to a wage difference that arises from nonmonetary characteristics of different jobs. Compensating differentials are common. Here are some examples:

compensating differential
a difference in wages that arises to offset the nonmonetary characteristics of different jobs

- Roofers are paid more than other workers with similar levels of education. Their higher wage compensates them for the nasty smell of tar and the constant risk of accidents.
- Workers on night shifts are paid more than similar workers on day shifts. The higher wage compensates them for having to work at night and sleep during the day, a lifestyle that most people find undesirable.
- Professors are paid less than lawyers and doctors, who have similar amounts of education. The higher wages of lawyers and doctors compensate them for missing the intellectual and personal satisfaction that professors' jobs offer. (Indeed, teaching economics is so much fun that it is surprising economics professors are paid anything at all!)

20-1b Human Capital

As the previous chapter discussed, an economy's stock of equipment and structures is called **capital**. The capital stock includes the farmer's tractor, the manufacturer's factory, and the teacher's chalkboard. The essence of capital is that it is a factor of production that itself has been produced.

human capital
the accumulation of investments in people, such as education and on-the-job training

Another type of capital, though less tangible than physical capital, is just as important to the economy's production: It is **human capital**, the accumulation of investments in people. The most important type of human capital is education. Like all forms of capital, education represents an expenditure of resources to raise future productivity. But this investment is tied to a specific person, and this linkage makes it human capital.

Workers with more human capital earn more, on average, than those with less. College graduates in the United States, for example, earn almost twice as much as

those with only a high school diploma. This large difference has been documented around the world. The gap tends to be even larger in less developed countries, where educated workers are in scarce supply.

From the perspective of supply and demand, it is easy to see why education raises wages. Firms—the demanders of labor—pay more for highly educated workers because these workers have higher marginal products. Workers—the suppliers of labor—bear the cost of education because they expect a reward for doing so. The difference in wages between highly educated workers and less educated workers may be considered a compensating differential for the cost of acquiring human capital.

Case Study

The Increasing Value of Skills

"The rich get richer, and the poor get poorer." Like many adages, this one is not always true, but it has been recently in the United States and many other nations. Numerous studies have documented that the earnings gap between workers with high skills and workers with low skills has increased substantially over the past several decades.

Table 1 presents data on the average earnings of college graduates and of high school graduates without any additional education. These data show the increase in the financial reward from education. In 1974, a man with a college degree earned 42 percent more on average than a man without one; by 2019, this figure had risen to 85 percent. Among women, the earnings gap between those with and without college degrees rose from 35 percent in 1974 to 78 percent in 2019. The incentive to stay in school today is large by historical standards.

Why has the gap in earnings between skilled and unskilled workers widened? Economists have proposed two hypotheses, both of which suggest that the demand for skilled labor has risen relative to the demand for unskilled labor. The shift in demand has led to a corresponding change in the wages of both groups, increasing inequality.

The first hypothesis focuses on international trade. Over the past half century, the amount of trade with other countries has markedly increased. As a percentage

Table 1

Average Annual Earnings by Educational Attainment

College graduates have always earned more than workers who did not attend college, but the gap has grown larger over the past few decades.

	1974	2019
Men		
High school, no college	$56,855	$52,677
College graduates	$80,973	$97,554
Percent extra for college grads	+42%	+85%
Women		
High school, no college	$32,675	$39,669
College graduates	$44,200	$70,657
Percent extra for college grads	+35%	+78%

Note: Earnings data are adjusted for inflation and are expressed in 2019 dollars. Data apply to full-time, year-round workers age 18 and over. Data for college graduates exclude workers with additional schooling beyond college, such as a master's degree or Ph.D.

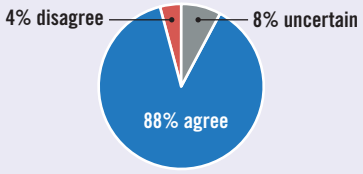
Source: U.S. Census Bureau, Tables P-32 and P-35, and author's calculations.

Ask the Experts

Inequality and Skills

"One of the leading reasons for rising U.S. income inequality over the past three decades is that technological change has affected workers with some skill sets differently than others."

What do economists say?



Source: IGM Economic Experts Panel, January 24, 2012.

of total U.S. production of goods and services, imports have risen from 5 percent in 1970 to 13 percent in 2020, and exports have risen from 6 percent in 1970 to 10 percent in 2020. Because unskilled labor is plentiful and cheap in many countries, the United States tends to import goods produced with unskilled labor and export those produced with skilled labor. This means that when international trade expands, domestic demand rises for skilled labor and falls for unskilled labor.

The second hypothesis emphasizes technological change. Consider computers. For keeping business records, many companies have replaced filing cabinets with computer databases, reducing the demand for filing clerks and increasing the demand for programmers and data analysts. Similarly, industrial robots have replaced the unskilled factory workers whose tasks can be automated, but they require skilled engineers to produce and maintain them. Economists call this shift in demand **skill-biased technological change**.

Economists debate the importance of trade, technology, and other forces on the distribution of wages. There is likely no single reason for the growing earnings gap between skilled and unskilled workers. The next chapter discusses income inequality in more detail. ●

20-1c Ability, Effort, and Chance

Why do major league baseball players earn more than those in the minor leagues? Certainly, the higher wage is not a compensating differential: Playing in the majors is hardly less pleasant than playing in the minors. Nor is human capital the key: The major leagues do not require more schooling or years of experience, though some training certainly helps. For the most part, players in the major leagues earn more simply because they have greater ability.

Ability is important in all occupations. Because of heredity and upbringing, people differ in personal attributes. Some people are strong, others weak. Some people are smart, others dull. Some people are affable, others awkward. These and many other characteristics affect workers' productivity and, therefore, their wages.

Closely related to ability is effort. Some people work harder than others, and this extra effort makes them more productive, which usually leads to higher earnings. Sometimes, firms reward hard work directly by paying people based on what they produce. Salespeople, for instance, are often paid a percentage of the sales they make. At other times, hard work is rewarded less directly in the form of a higher annual salary or a bonus.

But chance also plays a role in determining wages. Consider those who attended trade school to learn how to repair televisions with vacuum tubes and then found this skill made obsolete by the invention of solid-state electronics. They ended up earning a low wage compared with others with similar years of training. Or imagine what might happen to the future earnings of truck drivers if self-driving trucks are perfected. The decline in income experienced by these workers is due to random technological change.

Chance also arises from accidents of birth. If you are born to a family with a high income and deep education, that good fortune gives you an advantage in life. If you are born into extreme poverty, with a severe disability, or in a neighborhood without good role models, that's also a matter of chance.

How important are ability, effort, and chance in determining wages? These factors are hard to measure, but indirect evidence suggests that they are very important. When labor economists study wages, they relate a worker's wage to those variables that can be measured, such as years of schooling, years of experience, age, and job characteristics. All these measured variables affect a worker's wage as theory predicts, but they account for less than half of the variation in wages in the U.S. economy. Because so much of the variation in wages is left unexplained, omitted variables—including ability, effort, and chance—must be important.

Case Study

The Benefits of Beauty

People differ in many ways, one of which is physical attractiveness. The actor Gal Gadot, for instance, is a beautiful woman—indeed, she was once a beauty pageant winner. Her good looks help attract large audiences to her movies, and the large audiences mean a large income for Ms. Gadot. In 2020, she reportedly earned more than $30 million.

Physical attractiveness is obviously useful for an actor, but how widespread are the economic benefits of beauty? The labor economists Daniel Hamermesh and Jeff Biddle addressed this question in a study published in the December 1994 issue of the *American Economic Review*. Hamermesh and Biddle examined data from surveys of people in the United States and Canada. The interviewers who conducted the surveys were asked to rate each respondent's physical appearance. Hamermesh and Biddle then examined how much the wages of the respondents could be explained by the standard determinants—education, experience, and so on—and how much they depended on physical appearance.

They found that beauty pays. People deemed more attractive than average earn 5 percent more than people of average looks, and people of average looks earn 5 to 10 percent more than people considered less attractive than average. Similar results were found for men and women.

What explains these wage differences? There are several possible interpretations of the beauty premium.

One is that good looks are a type of innate ability determining productivity and wages. Some people are born with the physical attributes of a movie star; others are not. Good looks are useful in any public-facing job—not just acting, but also modeling, sales, and waiting on tables. In this case, attractive workers are more valuable to the firm, and its willingness to pay a beauty premium reflects its customers' preferences.

A second interpretation is that reported beauty is an indirect measure of other types of ability. How attractive a person appears depends on more than just heredity. It also depends on dress, hairstyle, personal demeanor, and other attributes that a person can control. Perhaps a person who successfully projects an attractive image in a survey interview is more likely to be a talented person who succeeds at other tasks as well.

A third interpretation is that the beauty premium is a type of discrimination, a topic to which we will return. ●

DFREE/SHUTTERSTOCK.COM

Good looks pay.

20-1d An Alternative View of Education: Signaling

Recall that the human-capital view of education holds that schooling makes workers more productive. Some economists have proposed another theory, which says that firms use educational attainment as a way of sorting between high-ability and low-ability workers. According to this view, when people earn a college degree,

they don't become more productive, but they do **signal** their high productivity to prospective employers. Because it is easier for high-ability people to earn a college degree than it is for low-ability people, more high-ability people get degrees. That makes it rational for firms to interpret a college degree as a signal of ability.

The signaling theory of education is similar to the signaling theory of advertising discussed in Chapter 17. In the case of advertising, the advertisements themselves contain no real information, but firms signal the quality of their products to consumers by their willingness to spend money on advertising. In the signaling theory of education, schooling has no real productivity benefit, but workers signal their innate productivity to employers by their willingness to spend years at school. In both cases, an action is being taken not for its intrinsic benefit but because the willingness to take that action conveys private information to someone observing it.

Both the human-capital theory and the signaling theory can explain why more educated workers tend to earn more than less educated ones. According to the human-capital view, education makes workers more productive; according to the signaling view, education is correlated with ability. But the two views have radically different predictions for policies that aim to increase educational attainment. According to the human-capital view, increasing educational levels for all workers would raise all workers' productivity and wages. According to the signaling view, education does not enhance productivity, so raising all workers' educational levels would not affect wages.

Most likely, the truth lies somewhere between these two extremes. The benefits of education are probably a combination of the productivity-enhancing effects of human capital and the productivity-revealing effects of signaling. The relative size of these two effects is an open question.

20-1e The Superstar Phenomenon

Although most actors earn little and often take jobs as waiters to support themselves, Scarlett Johansson earns millions of dollars for each film she makes. Similarly, while most people who play tennis do it as a hobby, Daniil Medvedev earns millions on the pro tour. Johansson and Medvedev are superstars in their fields, and their great public appeal is reflected in astronomical incomes.

Why do Johansson and Medvedev earn so much? It is not surprising that incomes differ within occupations. Good carpenters earn more than mediocre carpenters, and good plumbers earn more than mediocre plumbers. People vary in talent and effort, and these differences lead to differences in income. Yet the best carpenters and plumbers do not earn the fortunes that are common among the best actors and athletes. What explains the difference?

To understand the tremendous incomes of Johansson and Medvedev, consider the special features of the markets in which they sell their services. Superstars arise in markets with two characteristics:

- Every customer in the market wants to enjoy the services supplied by the best producers.
- The services are produced with a technology that makes it possible for the best producers to supply every customer at low cost.

If Scarlett Johansson is one of the best actors around, then everyone will want to see her next movie; seeing twice as many movies by an actor half as appealing is not a good substitute. Moreover, it is **possible** for everyone to enjoy a performance by Scarlett Johansson. Because it is easy to make multiple copies of a film, Johansson

can provide her acting services to millions of people simultaneously. Similarly, because tennis matches are broadcast on television, millions of fans can enjoy the extraordinary skills of Daniil Medvedev.

This logic shows why there are no superstar carpenters and plumbers. Other things being equal, everyone prefers to hire the best plumber, but a plumber, unlike a movie actor, can only work for a limited number of customers. Although the best plumber can command a somewhat higher wage than the average plumber, the average plumber can still earn a good living.

20-1f Below-Equilibrium Wages: Monopsony

Most often, economists analyze labor markets using the tools of supply and demand. The market is assumed to be competitive with many buyers and many sellers, each of whom has a negligible effect on the wage. Yet that assumption doesn't always apply.

Imagine that the labor market in a small town is dominated by a single, large employer. That employer can exert a large influence on the going wage, substantially changing the outcome in the labor market. A market in which there is a single buyer is called a **monopsony**.

monopsony
a market that has only one buyer

A monopsony (a market with one buyer) is like a monopoly (a market with one seller). Recall from Chapter 16 that a monopoly produces less of the good than a competitive firm would; by reducing the quantity offered for sale, the monopoly moves along the product's demand curve, increasing the price and its profit. Similarly, a monopsony in a labor market hires fewer workers than a competitive firm would; by reducing the number of jobs available, the monopsony moves along the labor supply curve, reducing the wage it pays and increasing its profit. Both monopolists and monopsonists reduce economic activity in a market below the socially optimal level. In both cases, the existence of market power distorts the outcome and causes deadweight losses. Workers employed by monopsonies earn less than they would under competition.

This book does not present the formal model of monopsony because true monopsonies are rare. In most labor markets, workers have many possible employers, and firms compete with one another to attract workers. In such cases, the model of supply and demand is the best one to use.

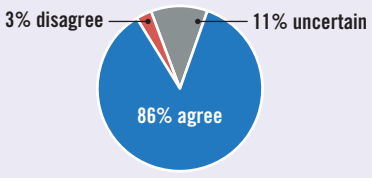
Yet the concept of monopsony, developed by the economist Joan Robinson in the 1930s, is important in some cases. In 2021, the Supreme Court declared that the NCAA has functioned as a monopsony, depriving college athletes of the compensation they would earn in a more competitive market. Around the same time, the Biden administration placed a spotlight on monopsonies in an executive order, seeking to reduce the power of big tech companies and increase workers' bargaining power. A growing problem, according to many economists, is the frequent use of employment contracts with non-compete clauses, which bar employees from leaving to work for a competitor. While these agreements protect employers' trade secrets, they also curb competition in the labor market, keeping wages below their equilibrium level.

Monopsonies in a pure form are not common, but the tendency toward monopsony still bedevils some parts of the modern economy. The effects of monopsony power may help explain some workers' wages.

Ask the Experts: Competition in Labor Markets

"The use of non-compete clauses in U.S. employment contracts reduces workers' mobility and wages by more than is justified by the protection of employers' intellectual property and trade secrets."

What do economists say?



Source: IGM Economic Experts Panel, August 3, 2021.

In the News

The Aftereffects of the Covid Pandemic

Some economists worry that students who fell behind because of Covid restrictions may never catch up in their skills, job prospects, and income.

The Long-Term Economic Costs of Lost Schooling

By Jon Hilsenrath

Imagine for a moment two objects in your hands. One is a piece of paper and the other a rubber band. If you squeeze your hands together hard and let go, the paper will remain crumpled, but the rubber band will return to its original shape.

Economists tend to think of the economy as the rubber band. After a shock, they expect it to go back to normal. When it doesn't, like the crumpled paper, they call the effect "hysteresis"—lasting changes caused by some large perturbation. The Covid-19 pandemic is a classic example. What permanent damage to the economy will it leave behind?

The first place to look is in classrooms, say Eric Hanushek and Margaret Raymond, economists and education researchers at Stanford University. Lost study time for children during the pandemic has the potential to do lasting harm not just to their own long-term prospects but to American prosperity in general, say the married couple.

Ms. Raymond studied 18 states and Washington, D.C. and concluded that, on average, children lost 116 days of reading time during the early stages of the pandemic last year and 215 days of math work—instruction that will be hard to regain and could leave a whole generation of children struggling to keep up in their studies and testing. If your child misses out on learning fractions now, how will she perform in algebra later?

And the shock has been distributed unevenly. Children in rural areas and areas with large Black and Hispanic populations were hit the hardest. Among the states suffering the most are South Carolina and Illinois, according to Ms. Raymond's study.

Economic output is a function of innovation, the skills that workers bring to their jobs and the machines that they use to create goods and services. Innovation and skills are shaped by education. Over the next century, the skill shock of 2020 will produce $25 trillion to $30 trillion of lost economic output in today's dollars, Mr. Hanushek estimates, and the lifetime household incomes of the affected students will be 6% to 9% lower.

He came to this conclusion in part by examining the experience of German students. In 1966 and 1967, the German government temporarily shortened the school year in a rejiggering of the school calendar. Longitudinal studies, he says, show that this lost class time reduced the incomes of that cohort of students by 5% over their lifetimes. Today's students "are going to feel the long-term effects of Covid even when they are back in school," Mr. Hanushek says.

Economists borrowed the term hysteresis from the physics of magnetism. If you apply a large enough magnetic force to a metal object, the polarity of the object can be permanently transformed. It's the mechanism used, for

20-1g Above-Equilibrium Wages: Minimum-Wage Laws, Unions, and Efficiency Wages

While workers employed by a monopsony have wages below the level that would prevail in a competitive market equilibrium, other workers are paid above that level. Above-equilibrium wages can arise for three reasons.

One reason is minimum-wage laws, as Chapter 6 discussed. Most workers in the economy are not affected by these laws because their equilibrium wages are well above the legal minimum. But for some workers, especially the least skilled and experienced, minimum-wage laws raise wages above the level they would earn in an unregulated labor market.

union
a worker association that bargains with employers over wages and working conditions

strike
a collective refusal to work organized as a form of protest

A second reason that wages might rise above their equilibrium level is the market power of labor unions. A **union** is a worker association that bargains with employers over wages and working conditions. Unions often raise wages above the level that would prevail in their absence, perhaps because they can organize workers to withhold their labor by calling a **strike**. Sometimes, unions act as a countervailing power to offset the monopsonistic behavior of employers; other

instance, to create memory in a hard drive. In economics, hysteresis is typically associated with damage after shocks, though there can be positive transformations too, such as the development of vaccine technologies and work-from-home options.

For many years, economists have looked for evidence of hysteresis in labor markets. In Europe in the 1970s and 1980s, economists Olivier Blanchard and Lawrence Summers noticed that unemployment tended to increase during economic downturns, as expected, but didn't fully return to previous levels when the economy revived. The rubber band, in other words, didn't regain its shape.

In a 1986 paper, "Hysteresis in Unemployment," the professors surmised that this was because of structural problems in these markets. Unions tended to fight hard to keep their workers from losing jobs but did little to help them after they were released, making it hard for them to be re-employed. Labor protections encoded in the law had the same effect: Firms were reluctant to rehire after downturns because it was so hard to let go of people in a recession. "Shocks that we thought should be temporary had long-lasting effects," Mr. Blanchard said in a recent interview.

Three decades later, Mr. Blanchard went back and looked at the problem of hysteresis in labor markets after the 2007–09 financial crisis. Millions of Americans were experiencing long spells of unemployment. In 2010, nearly half of unemployed workers were without a job for at least six months, an astonishingly high number. In the half-century before the crisis, just one of every eight unemployed workers, on average, went without work for that long.

Mr. Blanchard and other economists worried about the lasting damage to people who sat on the sidelines of the job market and saw their skills deteriorate. Some stopped looking for work; others found income on federal disability rolls. As the expansion marched on, however, some were drawn back into the labor market. Mr. Blanchard, to his surprise, found evidence of hysteresis less compelling than he had expected.

In the current economic crisis, policy makers in Washington are eager to drive the jobless rate down as fast as possible to reverse a new upsurge in long-term unemployment. It's one reason why Treasury Secretary Janet Yellen—a labor economist—has called for Congress to "go big" on a relief package.

Hysteresis may be acting on whole industries this time, Mr. Blanchard said. Air travel, commercial real estate and bricks-and-mortar retailing, for example, might never be the same.

Like Mr. Hanushek and Ms. Raymond, Mr. Blanchard is most worried about the long-run effects of the Covid crisis on children and their future as workers. "I would do everything I can to allow children to go safely back to school in person," he said. Ms. Raymond said that it might be time to start thinking about sending children to summer school to make up for lost time. At the very least, she said, it is time for educators to start thinking about how to fix schooling once the pandemic ends. ■

Source: *The Wall Street Journal,* February 27, 2021.

Questions to Discuss

1. Did you lose educational opportunities during the Covid pandemic? If so, do you think you subsequently caught up to where you otherwise would have been?
2. How do you think educational systems can best compensate for lost opportunities due to the pandemic?

times, they act more like monopolists, pricing their labor above the competitive level. Studies suggest that union workers earn about 10 to 20 percent more than similar, nonunion workers.

A third reason for above-equilibrium wages is based on the theory of **efficiency wages**. This theory holds that high wages increase worker productivity by reducing turnover, motivating greater effort, and enticing superior candidates to apply for jobs. If these effects are strong enough, firms may find it profitable to pay wages above the equilibrium level.

efficiency wages above-equilibrium wages paid by firms to increase worker productivity

Above-equilibrium wages, whether caused by minimum-wage laws, unions, or efficiency wages, have similar effects on the labor market. In each case, setting a wage above the equilibrium level increases the quantity of labor supplied and reduces the quantity of labor demanded. That creates a surplus of labor, or unemployment. The study of unemployment is usually considered a topic within macroeconomics, which is beyond the scope of this chapter. But these issues can be important when analyzing earnings. Most earnings differences can be understood while assuming that wages balance supply and demand, but above-equilibrium wages play a role in some cases.

QuickQuiz

1. Ted leaves his job as a high school math teacher and returns to school to study the latest developments in computer programming, after which he takes a higher-paying job at a software firm. This is an example of
   a. a compensating differential.
   b. human capital.
   c. monopsony.
   d. efficiency wages.
2. Marshall and Lily work at a local department store. Marshall, who greets customers as they arrive, is paid less than Lily, who cleans the bathrooms. This is an example of
   a. a compensating differential.
   b. monopsony.
   c. signaling.
   d. efficiency wages.
3. Barney runs a small manufacturing company. He pays his employees about twice as much as other firms, even though he could pay less and still recruit all the workers he needs. He believes that higher wages make his workers more loyal and hard-working. This is an example of
   a. monopsony.
   b. human capital.
   c. signaling.
   d. efficiency wages.
4. A business consulting firm hires Robin because she was a math major in college. Her new job does not require any of the mathematics she learned, but the firm's managers believe that anyone who can graduate with a math degree must be very smart. This is an example of
   a. a compensating differential.
   b. human capital.
   c. signaling.
   d. monopsony.

Answers are at the end of the chapter.

20-2 The Economics of Discrimination

discrimination
the offering of different opportunities to similar individuals who differ only by race, ethnicity, gender, age, religion, sexual orientation, or other personal characteristics

Another source of differences in wages is discrimination. **Discrimination** occurs when people are offered or denied opportunities based on race, ethnicity, gender, age, religion, sexual orientation, or other personal characteristics. Discrimination in the workplace reflects broader prejudices in society. Economists study the problem to attain a better understanding of its magnitude and sources.

20-2a Measuring Labor-Market Discrimination

How much does discrimination in labor markets affect the earnings of different groups of workers? This question is important, but answering it precisely is not easy.

There is no doubt that different groups of workers earn substantially different wages, as Table 2 shows. In 2019, the median Black man in the United States was paid 24 percent less than the median White man, and the median Black woman was paid

Table 2
Median Annual Earnings by Race and Sex

	White	Black	Percent by Which Earnings Are Lower for Black Workers
Men	$60,017	$45,644	24%
Women	$48,845	$41,098	16%
Percent by Which Earnings Are Lower for Women Workers	19%	10%	

Note: Earnings data are for the year 2019 and apply to full-time, year-round workers aged 14 and over. Individuals who report more than one race are excluded from these data.

Source: U.S. Census Bureau, Table P-38, and author's calculations.

16 percent less than the median White woman. The median White woman was paid 19 percent less than the median White man, and the median Black woman was paid 10 percent less than the median Black man. Taken at face value, these differentials look like incontrovertible evidence that employers discriminate against Black people and women.

Yet there is a potential problem with this inference. Even in a labor market free of discrimination, different people have different wages. People differ in the amount of human capital they have and in the kinds of work they are able and willing to do. The wage differences in an economy are, to some extent, attributable to the determinants of equilibrium wages discussed in the preceding section. Simply observing differences in wages among broad groups—White and Black people, men and women—does not prove that employers discriminate.

Consider the role of human capital. In 2019, among those aged 25 and older, 36 percent of White Americans had a bachelor's degree, compared with 26 percent of Black Americans. These educational differences explain some of the wage gap. And public schools in predominantly Black areas have historically been of lower quality—as measured by expenditure, class size, and so on—than public schools in predominantly White areas. If we could measure the quality as well as the quantity of education, the differences in human capital would likely appear even larger.

Human capital acquired in the form of job experience also helps explain wage differences. Women are more likely to interrupt their careers to raise children. Among the population aged 25 to 44 (when many people have children at home), 24 percent of women are out of the labor force, compared with 10 percent of men. As a result, female workers, especially at older ages, tend to have less job experience than male workers.

Compensating differentials are another source of wage differences. Men and women do not always choose the same type of work. For example, women are more likely to be administrative assistants, and men are more likely to be truck drivers. The relative wages of administrative assistants and truck drivers depend in part on the working conditions of each job. Because these nonmonetary aspects are hard to measure, it is difficult to gauge the practical importance of compensating differentials in explaining observed wage differences.

In the end, the study of wage differences among groups does not establish any clear conclusion about the prevalence of discrimination in U.S. labor markets. Most economists believe that some degree of discrimination is at work, but it's hard to say precisely how much. The only certain conclusion is a negative one: Because the differences in average wages among groups in part reflect differences in human capital and job characteristics, they do not by themselves measure the extent of labor-market discrimination.

Differences in human capital among groups of workers may, however, reflect a kind of discrimination. The less rigorous curriculums historically offered to female students, for instance, can be considered a discriminatory practice. Similarly, the inferior schools historically available to Black students may be traced to prejudice on the part of city councils and school boards. But this kind of discrimination occurs before workers enter the labor market. In this case, the disease is political, even if the symptom is economic.

Is Emily More Employable Than Lakisha?

Although gauging discrimination from labor-market outcomes is hard, compelling evidence for the existence of such discrimination comes from a creative field experiment. The economists Marianne Bertrand and Sendhil Mullainathan answered more than 1,300 help-wanted ads run in Boston and Chicago newspapers by sending in nearly 5,000 fake résumés. Half of the résumés had names that were common in the African American

community, such as Lakisha Washington or Jamal Jones. The other half had names that were more common among the White population, such as Emily Walsh and Greg Baker. Otherwise, the résumés were similar. The results of this experiment appeared in the *American Economic Review* in September 2004.

The researchers found large differences in how employers responded to the two groups of résumés. Job applicants with White names received about 50 percent more calls from employers than those with African American names. The study found that this discrimination occurred for all types of employers, including those who claimed to be an "Equal Opportunity Employer" in their help-wanted ads. The researchers concluded that "racial discrimination is still a prominent feature of the labor market."

More recently, the economist Philip Oreopoulos has examined Canadian labor markets by sending out some fake résumés with English names and others with Indian, Pakistani, Chinese, and Greek names. Published in the *American Economic Journal: Economic Policy* in November 2011, the study again found significant evidence of discrimination. English-sounding names received 39 percent more callbacks from employers. The differences were similar across the four ethnic groups. And the results were much the same if the fictional applicant had an English-sounding first name and a Chinese last name (such as James Liu or Amy Wang). When company recruiters were later asked about these findings, they tried to justify their behavior by saying that it is based on concern about language skills. Yet the discrimination occurred even when the applicant had a Canadian education and Canadian job experience, and there was no relationship between the advantage given to English names and the degree of language skills necessary for the type of job.

"What's in a name?" Shakespeare wrote in *Romeo and Juliet*. Like the Montagues and Capulets, many employers fail to look beyond the names of the people they are evaluating. ●

20-2b Discrimination by Employers

Let's now turn from measurement to the economic forces that lie behind discrimination in labor markets. If one group in society receives a lower wage than another, even after controlling for human capital and job characteristics, who is to blame?

The answer may seem obvious: employers. They make the hiring decisions that determine labor demand and wages. If some groups of workers earn less than they should, then employers seem to be the natural culprits. Yet many economists are skeptical of this answer. They believe that competitive market economies provide a natural antidote to employer discrimination: the profit motive.

Imagine an economy in which workers are differentiated only by hair color. In this simplified world, people are either blondes or brunettes. The two groups have the same skills, experience, and work ethic. Yet because of discrimination, employers prefer to hire brunettes, and this preference reduces the demand for blondes. This causes blondes to earn a lower wage than brunettes.

This wage differential, however, won't last for long. Entrepreneurs will soon notice an easy way to beat the competition: hiring blonde workers. Doing so means lower labor costs and higher profits. Over time, more and more "blonde" firms enter the market to exploit this cost advantage. Because the "brunette" firms have higher costs, they start losing money when faced with the new competitors and eventually go out of business. The entry of blonde firms and the exit of brunette firms increases the demand for blonde workers and reduces the demand for brunette workers. These shifts in demand pull their wages closer together. The process continues until the economy reaches a new equilibrium without the discriminatory wage differential.

Put simply, business owners who care only about making money are at an advantage when competing against those who also care about discriminating. As a result,

firms that do not discriminate tend to replace those that do. In this way, competitive markets have a natural remedy for employer discrimination.

Case Study

Segregated Streetcars and the Profit Motive

In the early 20th century, streetcars in many southern cities were segregated by race. White passengers sat in the front of the streetcars, and Black passengers sat in the back. The firms that ran the streetcars enforced this practice, but historical research shows that they didn't start it.

In a 1986 article in the *Journal of Economic History*, the economic historian Jennifer Roback found that the segregation of races on streetcars resulted from laws that required segregation. Before these laws existed, racial discrimination in seating was rare. It was far more common to segregate smokers and nonsmokers.

In fact, the firms that ran the streetcars often opposed the laws requiring racial segregation. Providing separate seating for different races raised the firms' costs and reduced their profits. One railroad company manager complained to the city council that, under the segregation laws, "the company has to haul around a good deal of empty space."

Here is how Roback describes the situation in one southern city:

> The railroad company did not initiate the segregation policy and was not at all eager to abide by it. State legislation, public agitation, and a threat to arrest the president of the railroad were all required to induce them to separate the races on their cars. . . . There is no indication that the management was motivated by belief in civil rights or racial equality. The evidence indicates their primary motives were economic; separation was costly. . . . Officials of the company may or may not have disliked blacks, but they were not willing to forgo the profits necessary to indulge such prejudice.

The story of southern streetcars illustrates a general lesson: Usually, business owners are most interested in making profits, not in discriminating against a particular group. When firms engage in discriminatory practices, the source of the discrimination often lies not with the firms themselves but elsewhere. In this case, the streetcar companies segregated White and Black people because discriminatory laws, which the companies opposed, required them to do so. ●

20-2c Discrimination by Customers and Governments

The profit motive acts to eliminate discriminatory wage differentials, but its corrective abilities only go so far. Two limiting factors are customer preferences and government policies.

To see how customer preferences for discrimination can affect wages, consider again the imaginary economy with blondes and brunettes. Suppose that restaurant owners discriminate against blondes when hiring waiters, so blonde waiters earn less than brunette waiters. In this case, an enterprising restaurant can open with blonde waiters and charge lower prices. If customers care only about the price and quality of meals, the discriminatory firms will be driven out of business, and the wage differential will disappear.

But imagine that customers prefer being served by brunettes. If this discriminatory preference is strong, the entry of blonde restaurants will not eliminate the wage differential. That is, if customers have discriminatory preferences, a competitive market may maintain a discriminatory wage differential. An economy with such discrimination would contain two types of restaurants. Blonde restaurants would have lower costs and charge lower prices. Brunette restaurants would have higher

costs and charge higher prices. Customers who did not care about the hair color of their waiters would be attracted to the lower prices at the blonde restaurants. Bigoted customers would go to the brunette restaurants and would pay for their discriminatory preference in the form of higher prices.

Another way for discrimination to persist in competitive markets is for the government to require it. If, for instance, the government passed a law stating that blondes could wash dishes but not work as waiters, a wage differential could persist in a competitive market. The segregated streetcars in the previous case study are one example of government-mandated discrimination. Similarly, before South Africa abandoned the formal policy of racial segregation called apartheid in 1990, Black people were prohibited from working in some jobs. When discriminatory governments pass such laws, they suppress the equalizing force of competitive markets.

To sum up: **Competitive markets contain a natural remedy for employer discrimination. The entry of firms that care only about profit tends to eliminate discriminatory wage differentials. These differentials persist in competitive markets when customers are willing to pay to maintain the discriminatory practice or when the government mandates it.**

Discrimination in Sports

Measuring discrimination is often difficult. To determine whether a group of workers is discriminated against, a researcher must correct for differences in the productivity between that group and others in the economy. Yet in most firms, it is difficult to measure a worker's contribution to the production of goods and services.

Sports are something of an exception. Professional sports teams revel in objective measures of productivity. In basketball, for instance, statistics on players' averages for scoring, assists, and rebounds are compiled instantaneously and pored over by voracious fans. For economists, this extensive documentation amounts to a bonanza.

Studies of sports teams suggest that racial discrimination has, in fact, been common and that much of the blame lies with customers. One study, published in the *Journal of Labor Economics* in 1988, examined the salaries of basketball players and found that Black players earned 20 percent less than White players of comparable ability. The study also found that attendance at basketball games was larger for teams with a greater proportion of White players. One interpretation of these facts is that, at least at the time of the study, customer discrimination made Black players less profitable than White players for team owners. In the presence of such customer discrimination, a discriminatory wage gap can persist, even if team owners care only about profit.

A similar situation once existed for baseball players. A study using data from the late 1960s showed that Black players earned less than comparable White players. Moreover, fewer fans attended games pitched by Black pitchers than games pitched by White pitchers, even though Black pitchers had better records than White pitchers. Studies of more recent salaries in baseball, however, have found no evidence of discriminatory wage differentials.

Another study, published in the *Quarterly Journal of Economics* in 1990, examined the market prices of old baseball cards. This study found similar evidence of discrimination. The cards of Black hitters sold for 10 percent less than the cards of comparable White hitters, and the cards of Black pitchers sold for 13 percent less than the cards of comparable White pitchers. These results suggest customer discrimination among baseball fans. ●

20-2d Statistical Discrimination

statistical discrimination discrimination that arises because an irrelevant but observable personal characteristic is correlated with a relevant but unobservable attribute

Beyond animosity toward particular groups, there is another possible cause of discrimination, called **statistical discrimination**. This theory assumes that employers have imperfect information about possible employees. If some relevant but unobservable employee characteristic happens to be correlated with an otherwise irrelevant but observable characteristic, employers may rely on the observable characteristic when making hiring decisions.

Suppose that employers care about punctuality but don't know whether a job applicant is likely to be punctual once hired. And suppose that employers have found that 10 percent of workers with blue eyes are chronically late, compared with only 5 percent of workers with brown eyes. Because of this correlation, employers might prefer hiring brown-eyed workers, even if they do not otherwise care about eye color. Blue-eyed people as a group would suffer from discrimination, even though 90 percent of them are punctual. The discrimination is "statistical" in the sense that each blue-eyed person is being stereotyped by the average behavior of the group.

This example is silly (punctuality is not really related to eye color). But the same phenomenon arises in real cases.

Some employers, for instance, prefer not to hire workers with criminal records. The simplest way to avoid doing so is to ask job applicants whether they have criminal records, and many employers do. Some states, however, have passed "ban the box" laws that prohibit employers from asking. (The "box" refers to the place on the job application that a person would check to signal a clean record.) The goal of these laws is to help ex-offenders find jobs and reenter society.

Despite the noble intent of these laws, one unintended consequence is that they foster statistical discrimination. Statistics show that Black men are more likely to have served time in prison than White men. Some employers who are aware of this but are prohibited from asking about criminal records may avoid hiring Black men. As a result, Black men without a criminal past would suffer from discrimination because of their group's average characteristics. Some studies have compared states with and without "ban the box" policies and have found that these laws significantly reduce employment for young Black men without college degrees. These results suggest that policymakers should look for ways to help ex-offenders that do not inadvertently increase statistical discrimination.

QuickQuiz

5. Among full-time U.S. workers, White women earn about _____ percent less than White men, and Black men earn about _____ percent less than White men.
   a. 5; 20
   b. 5; 40
   c. 20; 20
   d. 20; 40
6. It is difficult to measure to what extent discrimination affects labor market outcomes because
   a. data on wages are crucial but not readily available.
   b. firms misreport the wages they pay to hide discriminatory practices.
   c. workers differ in their attributes and the types of jobs they have.
   d. the same minimum-wage law applies to workers in all groups.
7. The forces of competition in markets with free entry and exit tend to eliminate wage differentials that arise from discrimination by
   a. employers.
   b. customers.
   c. government.
   d. all of the above.

Answers are at the end of the chapter.

20-3 Conclusion

In competitive markets, workers earn a wage equal to the value of their marginal contribution to the production of goods and services. But many things affect the value of the marginal product. Firms tend to pay more for workers who are talented, diligent, experienced, and educated because these workers are more productive. Firms are likely to pay less to those workers against whom customers discriminate because these workers contribute less to revenue.

The theory of the labor market developed in the last two chapters explains why some workers earn higher wages than other workers. The theory does not say that the resulting distribution of income is necessarily equal, fair, or desirable in any way. The next chapter takes up that topic.

Chapter in a Nutshell

- Workers earn different wages for many reasons. One is that wage differentials play a role in compensating workers for job attributes. Other things being equal, workers in hard, unpleasant jobs are paid more than workers in easy, pleasant jobs.
- Workers with more human capital are paid more than workers with less. The return to accumulating human capital is high and has increased over the past several decades.
- Years of education, experience, and job characteristics affect earnings in the way that theory predicts, but many differences in earnings cannot be explained by things that economists can easily measure. The unexplained variation in earnings is largely attributable to ability, effort, and chance.
- Some economists have suggested that more educated workers earn higher wages not because education raises productivity but because it signals to employers that these workers have high levels of ability. If this signaling theory is correct, then increasing the educational attainment of all workers would not raise the overall level of wages.
- Wages are sometimes kept away from the level that balances supply and demand. An explanation for below-equilibrium wages is the monopsony power of some employers. Three explanations for above-equilibrium wages are minimum-wage laws, unions, and efficiency wages.
- Some differences in earnings are attributable to discrimination based on race, gender, or other factors. Measuring the amount of discrimination is difficult because one must correct for differences in human capital and job characteristics.
- Competitive markets tend to limit discriminatory wage differences. If one group of workers earns less than another for reasons unrelated to productivity, then nondiscriminatory firms will be more profitable than discriminatory ones. Profit-seeking behavior can, therefore, reduce discriminatory wage gaps. Still, discrimination persists in competitive markets if customers are willing to pay more to discriminatory firms or if the government enacts laws that require firms to discriminate.
- Discrimination can also occur for statistical reasons. If employers have imperfect information about employees, they may discriminate against all members of a group whose average characteristics the employers find undesirable.

Key Concepts

compensating differential, p. 404
human capital, p. 404
monopsony, p. 409
union, p. 410
strike, p. 410
efficiency wages, p. 411
discrimination, p. 412
statistical discrimination, p. 417

Questions for Review

1. Why are roofers paid more than other workers with similar amounts of education?
2. In what sense is education a type of capital?
3. How might education raise a worker's wage without raising the worker's productivity?
4. What conditions lead to highly compensated superstars? Would you expect to see superstars in dentistry? In music? Explain.
5. Give three reasons a worker's wage might be above the level that balances supply and demand.
6. What difficulties arise in deciding whether a group of workers has a lower wage because of discrimination?
7. Explain how the forces of economic competition affect racial discrimination.
8. Give an example of how discrimination might persist in a competitive market.

Problems and Applications

1. College students sometimes work as summer interns for private firms or the government. Some of these positions pay little or nothing.
   a. What is the opportunity cost of taking such a job?
   b. Explain why students are willing to take these jobs.
   c. If you were to compare the earnings later in life of workers who had worked as interns and those who had taken summer jobs that paid more, what would you expect to find?
2. As Chapter 6 explained, a minimum-wage law distorts the market for low-wage labor. To reduce this distortion, some economists advocate a two-tiered minimum-wage system, with a regular minimum wage for adult workers and a lower, "subminimum" wage for teenage workers. Give two reasons a single minimum wage might distort the labor market for teenage workers more than it would the market for adult workers.
3. A basic finding of labor economics is that workers who have more experience in the labor force are paid more than workers who have less experience (holding constant the amount of formal education). Why might this be so? Some studies have also found that experience at the same job has an extra positive influence on wages. Explain why this might occur.
4. At some colleges and universities, economics professors receive higher salaries than professors in some other fields.
   a. Why might this be true?
   b. Some other colleges and universities have a policy of paying equal salaries to professors in all fields. At some of these schools, economics professors have lighter teaching loads than professors in some other fields. What role do the differences in teaching loads play?
5. Imagine that someone offered you a choice: You could spend four years studying at the world's best university, but you would have to forever keep your attendance there a secret. Or you could be awarded an official degree from the world's best university, but you couldn't actually attend. Which choice do you think would enhance your future earnings more? What does your answer say about the debate over signaling versus human capital in the role of education?
6. When recording devices were first invented more than 100 years ago, musicians could suddenly supply their music to large audiences at low cost. How do you suppose this development affected the income of the best musicians? How do you suppose it affected the income of average musicians?
7. A current debate in education is whether teachers should be paid on a standard pay scale based solely upon their years of training and teaching experience or whether part of their salary should be based upon their performance (called "merit pay").
   a. Why might merit pay be desirable?
   b. Who might be opposed to a system of merit pay?
   c. What is a potential challenge of merit pay?
   d. A related issue: Why might a school district decide to pay teachers significantly more than the salaries offered by surrounding districts?
8. When Alan Greenspan (an economist who would later chair the Federal Reserve) ran a consulting firm in the 1960s, he primarily hired female economists. He once told the *New York Times*, "I always valued

men and women equally, and I found that because others did not, good women economists were cheaper than men." Is Greenspan's behavior profit-maximizing? Is it admirable or despicable? If more employers were like Greenspan, what would happen to the wage differential between men and women? Why might other economic consulting firms at the time not have followed Greenspan's business strategy?

QuickQuiz Answers

1. **b** 2. **a** 3. **d** 4. **c** 5. **c** 6. **c** 7. **a**

Chapter

21

Income Inequality and Poverty

The great British Prime Minister Winston Churchill once said, "The inherent vice of capitalism is the unequal sharing of blessings. The inherent virtue of socialism is the equal sharing of miseries." In 1945, when he made those remarks, Churchill was both a famous critic of socialism and an eloquent celebrant of the British Empire and the economic system that helped fuel it. His observation underscores two facts. First, thanks to Adam Smith's invisible hand, nations that use market mechanisms to allocate resources usually achieve greater prosperity than those that do not. Second, prosperity in market economies is not shared equally. Incomes can differ greatly between the top and bottom of the economic ladder.

The previous two chapters analyzed why people's incomes differ so much. Incomes arise from payments to the factors of production—labor, land, and capital. When markets are competitive, factor prices adjust to balance supply and demand. In equilibrium, the factors of production are paid the value of their marginal product.

ROMAN SAMBORSKYI/SHUTTERSTOCK.COM; ARNO VAN ENGELAND/EYEEM/GETTY IMAGES

Because labor earnings make up about two-thirds of all income in the U.S. economy, labor is the most important factor for determining households' standard of living. Labor supply and demand depend on ability, effort, human capital, compensating differentials, discrimination, and so on. For the most part, these things determine how much workers earn. But sometimes, because of minimum-wage laws, unions, efficiency wages, and monopsony power, the payments to labor depart from their competitive equilibrium levels. All these diverse forces influence whether a person ends up rich, poor, or somewhere in between.

The distribution of income raises fundamental questions about the economy and public policy. This chapter addresses the topic in three steps. First, we consider how inequality is measured and how great it is in the United States and around the world. Second, we examine different views about the role that government should play in altering the income distribution. Third, we discuss public policies aimed at helping those members of society most in need.

21-1 Measuring Inequality

We begin by asking four questions:

- How much economic inequality is there in U.S. society?
- How many people live in poverty?
- What problems arise in measuring inequality and poverty?
- How often do people move between income classes?

BOB MANKOFF/CARTOON COLLECTIONS

"As far as I'm concerned, they can do what they want with the minimum wage, just as long as they keep their hands off the maximum wage."

21-1a U.S. Income Inequality

Imagine that you lined up all the families in the economy according to annual incomes. Then you divided them into five equal groups, called **quintiles**. Table 1 shows the income ranges for each quintile in 2019 and for the top 5 percent (a subset of the highest quintile). You can use this table to find where your family lies in the income distribution.

For studying the income distribution, economists find it useful to calculate the share of total income that each quintile received. Table 2 shows that information in selected years. In 2019, the lowest quintile earned 3.9 percent of all

Table 1

The Distribution of Family Income in the United States: 2019

Group	Annual Family Income
Lowest Quintile	$40,000 and below
Second Quintile	$40,001–$69,000
Middle Quintile	$69,001–$105,038
Fourth Quintile	$105,039–$164,930
Highest Quintile	$164,931 and above
Top 5 percent	$304,153 and above

Source: U.S. Bureau of the Census, Historical Income Tables, Table F-1.

Table 2

Income Inequality in the United States

This table shows the percentage of total before-tax income received by families in each fifth of the income distribution and by families in the top 5 percent.

Source: U.S. Bureau of the Census, Historical Income Tables, Table F-2.

Year	Lowest Quintile	Second Quintile	Middle Quintile	Fourth Quintile	Highest Quintile	Top 5%
2019	3.9%	9.2%	14.8%	22.5%	49.5%	21.9%
2010	3.8	9.4	15.4	23.5	47.9	20.0
2000	4.3	9.8	15.4	22.7	47.7	21.1
1990	4.6	10.8	16.6	23.8	44.3	17.4
1980	5.3	11.6	17.6	24.4	41.1	14.6
1970	5.4	12.2	17.6	23.8	40.9	15.6
1960	4.8	12.2	17.8	24.0	41.3	15.9
1950	4.5	12.0	17.4	23.4	42.7	17.3
1935	4.1	9.2	14.1	20.9	51.7	26.5

income, and the highest quintile earned 49.5 percent. Even though all the quintiles include the same number of families, the highest quintile had almost 13 times as much income as the lowest quintile.

The last column shows the income share of the richest families. In 2019, the top 5 percent of families earned 21.9 percent of all income, which was greater than the total income of the bottom 40 percent.

Table 2 also shows the distribution of income in various years beginning in 1935. At first glance, the distribution appears stable. Over many decades, the lowest quintile has earned about 4 to 5 percent of income, while the highest quintile has earned about 40 to 50 percent of income. Closer inspection reveals some trends in the degree of inequality. From 1935 to 1970, the distribution gradually became more equal. The share of the lowest quintile rose from 4.1 to 5.4 percent, and the share of the highest quintile fell from 51.7 to 40.9 percent. In more recent years, this trend has reversed itself. From 1970 to 2019, the share of the lowest quintile fell from 5.4 to 3.9 percent, and the share of the highest quintile rose from 40.9 to 49.5 percent.

The previous chapter discussed some reasons for the recent rise in inequality. Increased trade with low-wage countries (such as China) and skill-biased technological changes (such as robots) have tended to decrease the demand for unskilled labor and increase the demand for skilled labor. These shifts in demand have reduced the wages of unskilled workers relative to those of skilled workers, and the change in relative wages has increased inequality in family incomes.

21-1b Inequality around the World

How does inequality in the United States compare with that in other countries? This question is interesting, but answering it is problematic. Some countries don't have reliable data, and those that do collect it in different ways. Some measure individual incomes, others measure family incomes, and still others measure expenditure as a crude approximation of income. We can never be sure whether an observed difference in inequality reflects a true difference in the economies or merely a difference in the way data are collected.

FYI

Incomes of the Super-Rich

Tables 1 and 2 present some information about the affluent—the top 20 percent and the top 5 percent of the income distribution. But what about the very rich, such as the top 1 percent? Or the very, very rich, such as the top 0.01 percent?

Standard data derived from the Current Population Survey are less reliable about the extremes of the income distribution. One problem is sample size. If the government surveys 60,000 households, it will get only 600 households in the top 1 percent and only 6 households in the top 0.01 percent. Another problem is that participation in the survey is voluntary. When approached by government data collectors, the rich may be more likely to say, "no thanks."

To study the super-rich, economists turn to income tax returns. Because everyone is subject to income taxes, the sample sizes are large, and nonparticipation is not an option. Yet tax return data, while better in some ways, are worse in others. The tax code is designed to raise revenue, not to collect consistent data over time. As lawmakers revise the tax code, the incentives for people to receive and report income in any particular form can change. (For example, business income is especially important for the super-rich. How this income is reported depends on whether the business is organized as a partnership or a corporation, and business owners can change that.) Nonetheless, the tax data may be the best lens to see into the lives of the super-rich.

Two economists who have studied the U.S. tax data to gauge inequality are Thomas Piketty and Emmanuel Saez. Here are some of their findings:

- To be in the top 1 percent of the income distribution in 2018, a taxpayer had to have an income above $441,970. This group's share of total income increased from 7.8 percent in 1970 to 18.3 percent in 2018.
- To be in the top 0.1 percent of the income distribution in 2018, a taxpayer had to have an income above $1,753,300. This group's share of total income increased from 1.9 percent in 1970 to 7.9 percent in 2018.
- To be in the top 0.01 percent of the income distribution in 2018, a taxpayer had to have an income above $7,879,500. This group's share of total income increased from 0.5 percent in 1970 to 3.4 percent in 2018.

The bottom line: The increase in inequality over the past half century documented in Table 2 appears to be highly concentrated among the super-rich. ■

With this warning in mind, consider Figure 1, which compares inequality in two dozen major countries. The inequality measure used here is the **quintile ratio**, which is the income of the highest quintile divided by the income of the lowest quintile. The most equality is found in Pakistan and Sweden, where the highest quintile receives about 4.5 times as much income as the lowest quintile (though that relative equality occurs at a low level of average income in Pakistan and a high level of average income in Sweden). The least equality is found in South Africa, where the top group receives 28 times as much income as the bottom group. All countries have significant disparities between rich and poor, but the degree of inequality varies substantially.

The United States has more inequality than most countries and much more than other economically advanced countries, such as Germany, France, and Japan. But it has a more equal income distribution than some developing countries, such as South Africa, Venezuela, and Brazil. The United States has about the same degree of inequality as China, the world's most populous nation, which claims to practice "socialism with Chinese characteristics" but has, in recent years, increasingly embraced market forces.

poverty rate
the percentage of the population whose family income falls below an absolute level called the poverty line

poverty line
an absolute level of income set by the federal government for each family size below which a family is deemed to be in poverty

21-1c The Poverty Rate

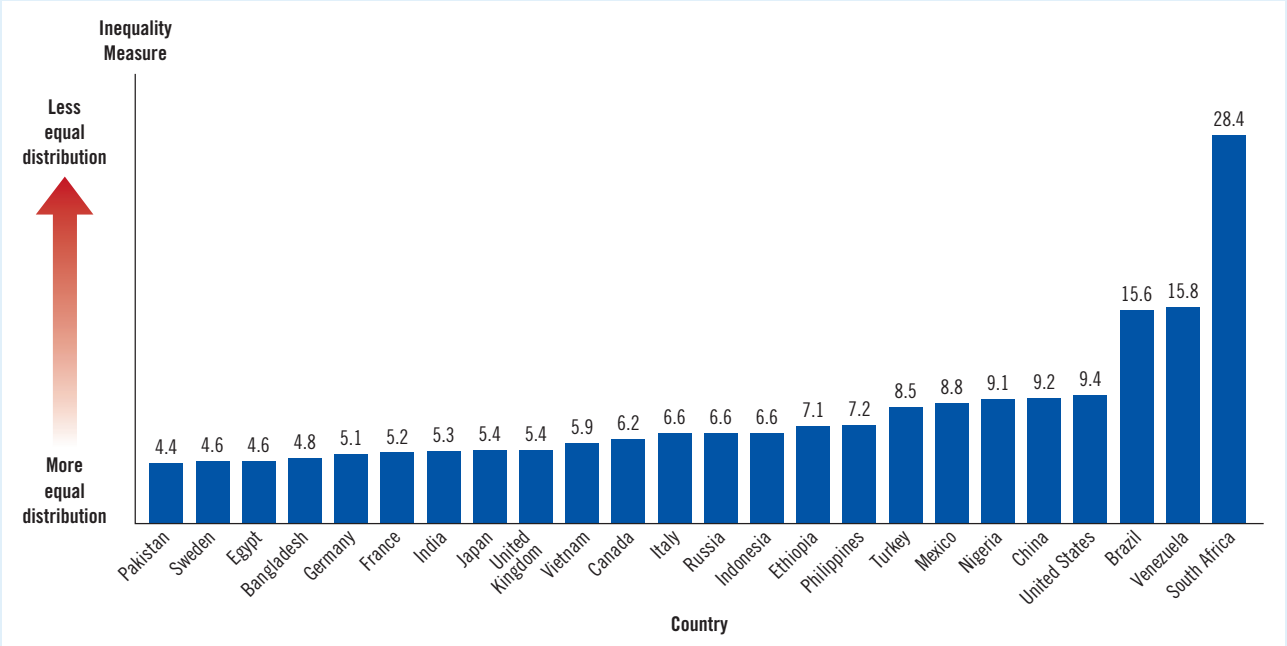
A commonly cited gauge of the distribution of income is the **poverty rate**, which is the percentage of the population whose family income falls below an absolute level called the **poverty line**. The poverty line is set by the federal government at

Figure 1

Inequality around the World

This figure shows the ratio of the income of the highest quintile to the income of the lowest quintile. Among these nations, Sweden and Pakistan have the most equal income distribution, while South Africa and Venezuela have the least equal.

Source: *Human Development Report 2018 Statistical Update.*



roughly three times the cost of providing an adequate diet. This line depends on family size and adjusts every year to account for price changes.

Consider the data for 2019. In that year, the median family in the United States had an income of $86,011, and the poverty line for a family with two adults and two children was $25,926. The poverty rate was 10.5 percent. In other words, 10.5 percent of the U.S. population were living below the poverty line for their family size.

Figure 2 shows the poverty rate since 1959, when the official data begin. You can see that the poverty rate fell from 22.4 percent in 1959 to 11.1 percent in 1973. This decline came as average income (adjusted for inflation) rose more than 50 percent. Because the poverty line is an absolute rather than a relative standard, more families tend to be pushed above the poverty line as economic growth pushes the entire income distribution upward. As President John F. Kennedy once put it, "a rising tide lifts all boats."

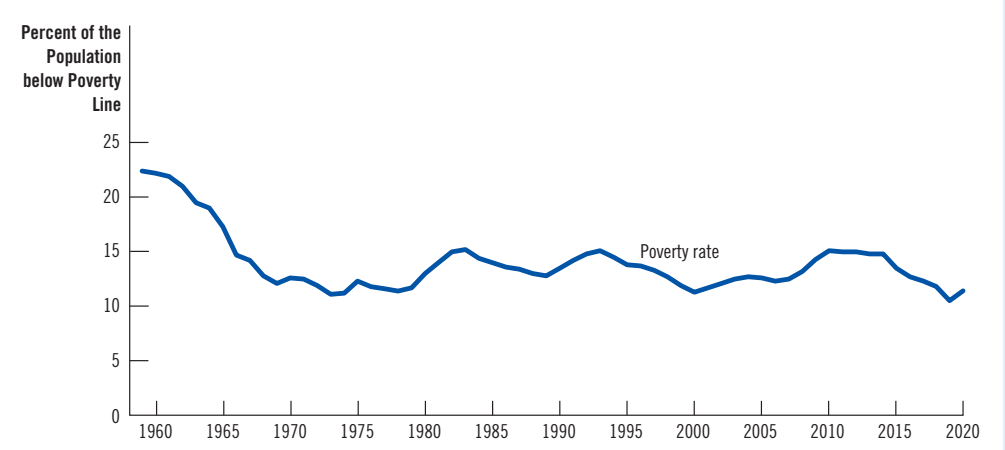
Since the early 1970s, however, the economy's rising tide has left smaller boats behind. Despite continued growth in average income, the poverty rate has not substantially changed since 1973. This lack of progress in reducing poverty in recent decades is closely related to the increasing inequality shown in Table 2. While economic growth has raised the income of the typical family, the increase in inequality has prevented the poorest families from sharing in this greater prosperity.

Figure 2

The Poverty Rate

The poverty rate measures the percentage of the population with incomes below an absolute level called the poverty line.

Source: U.S. Bureau of the Census.



Poverty affects all groups within the population but not with equal frequency. Table 3 shows the poverty rates for several groups, and it reveals three striking facts:

- Poverty is correlated with race. Blacks and Hispanics are more than twice as likely to live in poverty as Whites.
- Poverty is correlated with age. Children are more likely than average to be members of poor families, and older adults are less likely than average to be poor.
- Poverty is correlated with family composition. Families headed by a single mother are about five times as likely to live in poverty as families headed by a married couple.

These correlations have described American society for many years, and they show which people are most likely to be poor. These effects also work together: Children in Black and Hispanic female-headed households have especially high rates of poverty.

Table 3

Who Lives in Poverty?

This table shows that the poverty rate varies greatly among different groups within the population.

Source: U.S. Bureau of the Census. Data are for 2019.

Group	Poverty Rate
All persons	10.5%
White, not Hispanic	7.3
Black	18.8
Hispanic	15.7
Asian	7.3
Children (under age 18)	14.4
Older adults (over age 64)	8.9
Married-couple families	4.6
Female household, no spouse present	24.3

21-1d Problems in Measuring Inequality

Data on the income distribution and poverty rate give some idea about the degree of inequality in American society, but their interpretation is not always straightforward. The data are based on annual family incomes. What most people care about, however, is not their incomes in a single year but their ability to maintain a good standard of living. For several reasons, data on the income distribution and the poverty rate give an incomplete picture of inequality in living standards.

Taxes and In-Kind Transfers Standard measures of the income distribution and poverty rate are based on families' **pre-tax** incomes. They don't account for the greater taxes that high-income households typically pay and the tax credits that low-income households often receive. In particular, the earned income tax credit (discussed later in this chapter) gives cash payments to many low-wage workers. Because these payments are made through the income tax system, the incomes used to calculate the official poverty rate do not reflect the impact of this antipoverty program.

This data problem was significant during the pandemic of 2020, when Congress enacted large programs to support household income while many businesses were temporarily closed. Some of the new payments, such as expanded unemployment insurance, were included in the income measure used to calculate the poverty rate. But the so-called stimulus checks received by most families were excluded because they were administered through the tax system. As a result, the official poverty rate rose in 2020, though a statistic called the Supplemental Poverty Measure, which accounts for tax rebates such as the stimulus checks, showed a decline in poverty.

Another problem with the data on the income distribution and the official poverty rate is that they are based on **monetary** incomes. Yet various government programs provide low-income households with nonmonetary items, such as free food, housing vouchers, and medical services. Transfers in the form of goods and services rather than cash are called **in-kind transfers**. The Supplemental Poverty Measure takes account of some of these in-kind transfers, but standard measures of inequality and poverty do not.

in-kind transfers
transfers given in the form of goods and services rather than cash

The Economic Life Cycle Incomes vary predictably over most people's lives. Young adults, especially those in school, often have low incomes. Incomes tend to rise as people gain maturity and experience in the workforce, peak at around age 50, and then fall sharply around age 65 when many people retire. This typical pattern of income variation is called the **life cycle**.

life cycle
the regular pattern of income variation over a person's life

Because people can borrow and save to smooth out life cycle changes in income, their standard of living in any year need not depend only on that year's income. The young often borrow, perhaps to go to school or buy a house, and then repay these loans later as their incomes rise. People tend to have their highest saving rates when they are middle-aged. These savings can be drawn down in retirement, so the standard of living of older adults need not decline as much as their earnings.

Most data sets lump together the incomes of college students, people in their prime working years, and retirees. Because of the common life cycle pattern, some of the inequality in annual income does not represent true inequality in living standards.

Transitory versus Permanent Income Incomes vary over people's lives not only because of predictable life cycle variation but also because of random and transitory forces. One year, a frost kills off the Florida orange crop, and Florida orange growers see their incomes fall temporarily. At the same time, the Florida frost drives up the price of oranges, and California orange growers see their incomes temporarily rise. The next year, the reverse might happen. Or a pandemic may reduce the income of restaurant workers but increase the revenue of home-delivery pizza businesses. When public health is restored, the pattern shifts.

Just as people can borrow and save to smooth out life cycle variations in income, they can also borrow and save to smooth out transitory variations in income. To the extent that a family saves in good years and borrows (or depletes its savings) in bad ones, transitory changes in income need not affect its standard of living. A family's ability to buy goods and services would then depend on its **permanent income**, which is its normal, or average, income over several years.

permanent income
a person's normal income

To gauge inequality of living standards, the distribution of permanent income may be more relevant than the distribution of annual income. Because permanent income is less affected by transitory events, it is more equally distributed than current income.

21-1e Economic Mobility

People sometimes speak of "the rich" and "the poor" as if these groups consisted of the same families year after year. But this is not the case. Economic mobility, the movement of people among income classes, is significant in the U.S. economy. Movements up the income ladder can be due to good luck or hard work, and movements down can be due to bad luck or laziness. Some of this mobility reflects transitory variations in income, while some reflects more persistent changes in income.

Because family income changes over time, temporary poverty is more common than the poverty rate suggests, but persistent poverty is less common. In a typical 10-year period, about one in four families falls below the poverty line in at least one year. Yet fewer than 3 percent of families are poor for eight or more years. Because it is likely that the temporarily poor and the persistently poor face different problems, policies that aim to combat poverty often try to distinguish between these groups.

Another way to gauge mobility is with the persistence of economic success from generation to generation. According to studies of this topic, having an above-average income tends to carry over from parents to children, but the persistence is far from perfect, indicating substantial mobility among income classes. If a father earns 20 percent above his generation's average income, his son will most likely earn 8 percent above his generation's average income. The correlation between the income of a grandfather and the income of his grandson is small.

Because of this intergenerational economic mobility, the U.S. economy is filled with newly minted millionaires (as well as with heirs who have squandered the fortunes they inherited). According to one study, about four out of five millionaires made their money on their own rather than inheriting it. Often, they started and built a business or climbed the corporate ladder.

Studies have documented that the degree of intergenerational mobility varies from country to country and that mobility is negatively correlated with inequality. Nations with greater inequality than the United States, such as Brazil, tend to have lower mobility. Nations with lower inequality than the United States, such

as Sweden, tend to have greater mobility. Whether these international differences mainly reflect disparities in populations, policies, institutions, culture, or other factors isn't clear.

Case Study

A Lifetime Perspective on Income Inequality

For the reasons we have discussed, the standard data on the income distribution can give a misleading picture about how much inequality there is. In a 2021 study titled "U.S. Inequality and Fiscal Progressivity," the economists Alan Auerbach, Laurence Kotlikoff, and Darryl Koehler tried to correct for these problems. Rather than lumping all ages together—college students, full-time workers, and retirees—they examined inequality within specific cohorts, such as those aged 40 to 49. In addition, they looked not at a single year's income but at the total resources people have available to spend over their remaining years of life, which includes wealth, current income, and expected future income. And they accounted for taxes and transfer payments attributable to federal, state, and local government policies, including in-kind transfers.

Here are their estimates of the distribution of lifetime resources for those aged 40 to 49:

- Lowest quintile: 6.6 percent
- Second quintile: 9.8 percent
- Middle quintile: 14.0 percent
- Fourth quintile: 19.7 percent
- Highest quintile: 49.8 percent
- Top 5 percent: 24.4 percent
- Top 1 percent: 11.8 percent

Comparing these results with those in Table 2, you will see a large difference for the lowest quintile. The standard data show that the lowest quintile gets about 4 percent of annual income, but these new data say that the lowest quintile gets 6.6 percent of lifetime resources. At the other end of the distribution, while the data from income tax returns (discussed in the preceding FYI box) say that the top 1 percent gets 18.3 percent of annual income, this research finds that the top 1 percent gets 11.8 percent of lifetime resources.

Recall that one measure of inequality is the quintile ratio, the ratio of the income of the highest 20 percent to that of the lowest 20 percent. This measure is about 12.5 in the data on pre-tax annual income in Table 2 but only 7.5 in the data on lifetime resources.

This research also quantifies how public policy—including both taxes and transfers—affects the distribution of lifetime resources. The highest quintile pays an average tax rate over their lifetimes of 31 percent, and the top 1 percent pays an average tax rate of 35 percent. The middle class, defined here as the middle quintile of lifetime resources, pays an average tax rate of 19 percent. But because of the earned income tax credit and other antipoverty programs, those in the lowest quintile face an average tax rate of **negative** 44 percent. The negative sign means that people in the lowest quintile receive more in government transfer payments than they pay in taxes.

The bottom line: The United States has substantial inequality, but not as much as conventional measures indicate. ●

QuickQuiz

1. In the United States today, the poorest fifth of the population earns about _____ percent of all pre-tax annual income, while the richest fifth earns about _____ percent.
   a. 2; 70
   b. 4; 50
   c. 6; 35
   d. 8; 25
2. When comparing income inequality across nations, one finds that the United States is
   a. the most equal nation in the world.
   b. more equal than most nations but not the most equal.
   c. less equal than most nations but not the least equal.
   d. the least equal nation in the world.
3. Because consumption is largely determined by _____ income, consumption is _____ equally distributed than current income.
   a. permanent; more
   b. permanent; less
   c. transitory; more
   d. transitory; less

Answers are at the end of the chapter.

21-2 The Political Philosophy of Redistributing Income

We have seen how the economy's income is distributed and have considered some of the problems in interpreting measured inequality. This discussion was **positive** in the sense that it described the world as it is. We now turn to a **normative** question: What should society do about economic inequality?

Economic analysis alone cannot answer this question. People's differing views on it are largely a matter of political philosophy. Yet because the government's role in redistributing income is central to so many debates over economic policy, let's digress from economic science and consider a bit of what political philosophers have to say.

21-2a The Utilitarian Tradition

utilitarianism
the political philosophy according to which the government should choose policies to maximize the total utility of everyone in society

utility
a measure of satisfaction

One school of thought in political philosophy is **utilitarianism**, founded by Jeremy Bentham (1748–1832) and John Stuart Mill (1806–1873). To a large extent, utilitarians aim to apply the logic of individual decision making to questions concerning morality and public policy.

The starting point of utilitarianism is the notion of **utility**—the satisfaction that people receive from their circumstances. Utility is a measure of well-being and, according to utilitarians, the ultimate objective of all private and public actions. The proper goal of the government, they claim, is to achieve "the greatest good for the greatest number." In more technical language, the goal is to maximize the sum of utility achieved by everyone in society.

From this perspective, the case for redistributing income is based on the assumption of **diminishing marginal utility**. That means, for example, that an extra dollar of income increases the well-being of a poor person more than it increases the well-being of a rich person. Put another way, as a person's income rises, the extra utility from an additional dollar of income falls. This plausible assumption, together with the goal of maximizing total utility, implies that the government should take measures to make the distribution of income more equal.

The argument is simple. Imagine that Peter and Paula are the same, except that Peter earns $150,000 and Paula earns $50,000. In this case, taking a dollar from

Peter to pay Paula will reduce Peter's utility and raise Paula's utility. But because of diminishing marginal utility, Peter's utility falls by less than Paula's utility rises. This redistribution increases total utility, which is the utilitarian's objective.

At first, this utilitarian argument might seem to imply that the government should continue redistributing income until everyone has the same income. And that would be the case if the total amount of income—$200,000 in this example—were fixed. But in fact, it is not. Utilitarians reject complete equalization of incomes because they acknowledge one of the **Ten Principles of Economics** from Chapter 1: People respond to incentives.

To take from Peter to pay Paula, the government must pursue policies that redistribute income. The U.S. federal income tax and welfare system are examples. Under these policies, people with high incomes pay high taxes, and people with low incomes receive income transfers. These income transfers are phased out: As people earn more, they receive less from the government. Yet when Peter faces a higher income tax rate and Paula faces a system of phased-out transfers, both have less incentive to work hard because each gets to keep only a fraction of any additional earnings. But if they work less, society's income falls, and so does total utility. The utilitarian government has to balance the gains from greater equality against the losses from distorted incentives. To maximize total utility, the government stops short of making society fully egalitarian.

A famous parable sheds light on utilitarian logic. Imagine that Peter and Paula are thirsty travelers trapped at different places in the desert. Peter's oasis has a lot of water; Paula's has only a little. If the government could transfer water from one oasis to the other without cost, it would maximize total utility from water by equalizing the amount in the two places. But suppose that the government has only a leaky bucket. As it tries to move water from one place to the other, some water is lost in transit. In this case, a utilitarian government might still try to redistribute water from Peter to Paula, depending on Paula's thirst and the bucket's leak. But with only a leaky bucket at its disposal, a utilitarian government will stop short of trying to reach full equality.

21-2b The Liberal Contractarian Tradition

A second way of thinking about inequality might be called **liberal contractarianism**. An influential work in this school of thought is the book *A Theory of Justice* by the philosopher John Rawls (1921–2002). It is closely related to earlier work by the economists and Nobel laureates William Vickrey (1914–1996) and John Harsanyi (1920–2000).

liberal contractarianism the political philosophy according to which the government should choose policies deemed just, as evaluated by impartial observers behind a "veil of ignorance"

Rawls begins with the premise that a society's institutions, laws, and policies should be just. He then addresses the natural question: How can we, the members of society, ever agree on what justice means? Everyone's point of view is inevitably based on their own circumstances—whether they are talented or inept, diligent or lazy, educated or less educated, born to a wealthy family or a poor one, part of a privileged majority or an oppressed minority. Could any of us ever **objectively** determine what a just society would look like?

To answer this question, Rawls proposes a thought experiment. Imagine that before any of us is born, we all get together in the beforelife (the pre-birth version of the afterlife) to discuss and agree on a contract for the rules that will govern society. At this point, we don't know the station in life each of us will end up filling. In Rawls's words, we are sitting in an "original position" behind a "veil of ignorance." In this original position, Rawls argues, we can choose a just set of rules for society because we must consider how those rules will affect every person. As Rawls puts

it, "Since all are similarly situated and no one is able to design principles to favor his particular conditions, the principles of justice are the result of fair agreement or bargain."

Designing public policies and institutions in this way fosters objectivity about what policies are just. In many ways, this amounts to a formal, philosophical entreaty to abide by the Golden Rule, an ethic that has been found in many religions and cultures throughout history: Treat others as you would like others to treat you. When writing a social contract in the original position, you have no choice but to follow this rule because you don't know yet who you will be.

In his book, Rawls considers what public policy designed behind this veil of ignorance would try to achieve. In particular, he considers what income distribution people would consider fair if they did not know whether they would end up at the top, bottom, or middle of the distribution. Rawls says that people in the original position would be chiefly concerned about the possibility of being at the **bottom**. When society designs public policies, therefore, the welfare of the worst-off person in society should be the main concern. Rather than maximizing the sum of everyone's utility as a utilitarian would, Rawls would strive to maximize the minimum utility. Rawls's rule is called the **maximin criterion**.

maximin criterion
the claim that the government should aim to maximize the well-being of the worst-off person in society

The maximin criterion emphasizes the lot of the least fortunate, but it does not lead to a completely egalitarian society. Like the utilitarians, Rawls recognizes that people respond to incentives. If the government promised to fully equalize incomes, people would have no incentive to work hard, society's total income would fall substantially, and the least fortunate person would be worse off. The maximin criterion allows disparities in income if they improve incentives and thereby raise society's ability to help the poor. But Rawls is not in complete agreement with the utilitarians: Because his philosophy puts weight on only the least fortunate members of society, it calls for more income redistribution than does utilitarianism.

Rawls's views are hotly debated, but his thought experiment has much appeal. It allows us to consider the redistribution of income as a form of **social insurance**. From the perspective of the original position behind the veil of ignorance, income redistribution is like an insurance contract. Homeowners buy fire insurance to protect themselves from the risk of their house burning down. Similarly, when we as a society choose policies that tax the rich to supplement the incomes of the poor, we are all insuring ourselves against the possibility of being members of poor families. Because people generally dislike risk, we should be happy to be born into a society that provides this insurance.

social insurance
government policy aimed at protecting people against the risk of adverse events

Yet rational people behind the veil of ignorance might not be so risk averse that they would follow the maximin criterion. Instead, because people in the original position might end up anywhere in the distribution of outcomes, they might treat all possible outcomes equally when designing public policies. In this case, the best policy behind the veil of ignorance would be to maximize the average utility of members of society, and the resulting notion of justice would be more utilitarian than Rawlsian.

21-2c The Libertarian Tradition

A third view of inequality is called **libertarianism**. The two views we have considered so far—utilitarianism and liberal contractarianism—both view the total income of society as a shared resource that a social planner can redistribute to achieve some social goal. By contrast, according to libertarians, society itself earns no income; only individuals do. In their view, the government has no grounds for taking from some individuals and giving to others to achieve any particular distribution of income.

libertarianism
the political philosophy according to which the government should punish crimes and enforce voluntary agreements but not redistribute income

For instance, the philosopher Robert Nozick (1938–2002) writes the following in his book *Anarchy, State, and Utopia*:

> We are not in the position of children who have been given portions of pie by someone who now makes last minute adjustments to rectify careless cutting. There is no **central** distribution, no person or group entitled to control all the resources, jointly deciding how they are to be doled out. What each person gets, he gets from others who give to him in exchange for something, or as a gift. In a free society, diverse persons control different resources, and new holdings arise out of the voluntary exchanges and actions of persons.

While utilitarians and liberal contractarians try to judge what amount of inequality is desirable in society, Nozick denies the validity of this very question.

The libertarian alternative to evaluating economic **outcomes** is to evaluate the **process** by which these outcomes arise. Libertarians agree with most everyone else that when the distribution of income is achieved unfairly—for instance, when one person steals from or defrauds another—the government should remedy the problem. But this school of thought says that if the process is just, the resulting distribution is fair, no matter how unequal.

Nozick criticizes Rawls's approach by drawing an analogy between the distribution of income in society and the distribution of grades in a course. Suppose you were asked to judge the fairness of the grades in the economics course you are now taking. Would you imagine yourself behind a veil of ignorance and choose a grade distribution without knowing the talents, efforts, and performance of each student? Or would you ensure that the process of assigning grades to students is fair without regard for whether the resulting distribution is equal or unequal? For the case of grades, the libertarian emphasis on process over outcomes may be persuasive. Whether a similar logic applies to income is open to debate.

Libertarians conclude that equality of opportunities is more important than equality of outcomes. They believe that the government should enforce individual rights to ensure that everyone has the same opportunity to use their talents and achieve success. Once these rules of the game are established and enforced, libertarians say, the government has no reason to alter the resulting distribution of income.

QuickQuiz

4. A utilitarian believes that the redistribution of income is worthwhile as long as
   a. the worst-off members of society benefit from it.
   b. those contributing to the system are in favor of it.
   c. everyone's income, after taxes and transfers, reflects their marginal product.
   d. the distortionary effect on work incentives is not too large.
5. Rawls's thought experiment of the "original position" behind the "veil of ignorance" is meant to draw attention to the fact that
   a. most people with low income were not sufficiently educated when young.
   b. the station of life each of us was born into is largely a matter of chance.
   c. the rich have so much money that they don't know how to spend it all.
   d. outcomes are efficient only if everyone begins with equal opportunity.
6. Libertarians believe that
   a. the government should aim to improve the well-being of the worst-off person in society.
   b. policy should aim for an income distribution that maximizes the total happiness of all members of society.
   c. people should be free to engage in voluntary transactions, even if large income disparities result.
   d. large income disparities are likely to threaten political liberty.

Answers are at the end of the chapter.

21-3 Policies to Reduce Poverty

As we have just seen, political philosophers hold various views about income redistribution. Political debate reflects a similar disagreement. Nonetheless, most people believe that, at the very least, society should try to help those most in need and that the government needs to step in when private charity fails. According to a popular metaphor, society should provide a "safety net" to prevent any citizen from falling too far.

Poverty is a challenging problem. Families with low incomes are more likely than the overall population to experience homelessness, drug dependence, health problems, teenage pregnancy, illiteracy, unemployment, and low educational attainment. Members of these families are more likely both to commit crimes and to be victims of crimes. It is sometimes hard to distinguish the causes of poverty from the effects, but there is no doubt that poverty is associated with serious economic and social ills.

Suppose that you were a policymaker in the government and your goal was to reduce the number of people living in poverty. What would you do? Here we examine some of the policy options. Each helps some people escape poverty, but none are perfect, and deciding on the best combination is not easy.

21-3a Minimum-Wage Laws

Laws setting a minimum wage that employers can pay workers are a perennial source of debate. Advocates view the minimum wage as a way of helping the working poor without any cost to the government. Critics view it as hurting those it is intended to help.

The minimum wage is easily understood using the tools of supply and demand, as discussed in Chapter 6. For workers with low levels of skill and experience, a high minimum wage forces the wage above the level that balances supply and demand. That raises the cost of labor to firms and reduces the quantity of labor that those firms demand. The result is higher unemployment among those groups of workers affected by the minimum wage. Those workers who remain employed benefit from a higher wage, but those who might have been employed at a lower wage are worse off.

The magnitude of these effects depends crucially on the elasticity of labor demand. Advocates of a high minimum wage argue that the demand for unskilled labor is relatively inelastic, so a high minimum wage depresses employment only slightly. Critics argue that labor demand is more elastic, especially in the long run when firms can adjust employment and production more fully. They also note that because many minimum-wage workers are teenagers from middle-class families, a high minimum wage is not well-targeted as a policy for helping the poor.

21-3b Welfare

welfare
government programs that supplement the incomes of the needy

One way for the government to raise the living standards of the poor is to supplement their incomes. The primary way the government does this is through the welfare system. **Welfare** is a broad and imprecise term that encompasses a variety of programs. Temporary Assistance for Needy Families (TANF) assists families with children and no adult able to support the family. Supplemental Security Income (SSI) helps people who have low incomes because of sickness or disability. For both programs, people cannot qualify for assistance simply by having limited financial resources. They must also establish some additional "need," such as having small children or a disability.

A common criticism of public assistance programs is that they create perverse incentives for those who might qualify. For example, it is said that these programs encourage families to break up (because many families qualify for financial assistance only if the father is absent) and encourage women to give birth out of wedlock (because

many poor single women qualify for assistance only if they have children). Because poor single mothers are such a large part of the poverty problem, these policies are said to exacerbate the very problems they are supposed to cure. Such arguments led to an overhaul of the welfare system in 1996 with a law that limited the amount of time recipients could stay on welfare. That change in policy remains controversial.

Those who rebut these arguments and support more generous welfare point out that being a poor single mother on welfare is a difficult existence at best, and they do not believe that many people would choose such a life if it were not thrust upon them. Moreover, trends do not support the view that the decline of the two-parent family is connected to public assistance programs. The reduction in public assistance benefits in 1996 did not lead to a decline in the percentage of children living with only one parent.

21-3c Negative Income Tax

When the government collects taxes, it affects the distribution of income. This is clear in the case of a progressive income tax: Families with higher incomes pay a larger percentage of their income in taxes than do families with lower incomes. As Chapter 13 discussed, tax progressivity is a policy tool aimed at achieving vertical equity.

Many economists have advocated helping low-income households with a **negative income tax**. According to this policy, a progressive income tax would include not just rising average tax rates but also subsidies for families at the bottom of the income distribution. In other words, these families would "pay" a "negative tax."

negative income tax a tax system that collects revenue from high-income households and gives subsidies to low-income households

For example, suppose the government used the following formula to compute a family's tax liability:

$$\text{Taxes owed} = (1/3 \text{ of income}) - \$15{,}000$$

In this case, a family that earned $180,000 would pay $45,000 in taxes, and a family that earned $90,000 would pay $15,000 in taxes. A family that earned $45,000 would owe nothing. And a family that earned $15,000 would "owe" –$10,000. In other words, the government would send this family a check for $10,000.

A negative income tax provides what is sometimes called a **universal basic income**. In this example, a family that earns nothing on its own would receive $15,000 from the government. This way, no family would have an after-tax income below $15,000. Such a system can be viewed as a proportional tax of one-third of income, along with a grant to all families of $15,000.

Under a negative income tax, the only qualification required to receive government assistance is a low income. Depending on one's point of view, this feature can be either an advantage or a disadvantage. On the one hand, a negative income tax would establish a minimum standard of living for everyone, regardless of circumstance. On the other hand, a negative income tax would subsidize not only the unfortunate but also those who simply don't want to work and are, in some people's eyes, undeserving of public support.

Some actual tax provisions work much like a negative income tax. One is the Earned Income Tax Credit (EITC). This credit allows poor working families to receive income tax refunds greater than the taxes they paid during the year. Because the EITC applies only to the working poor, it does not discourage recipients from working as other antipoverty programs may. For the same reason, however, it also does not help alleviate poverty due to unemployment, sickness, or other inability to work.

Another tax provision that works like a negative income tax is the child tax credit. In 2021, this credit was increased so families were eligible to receive $3,600 for every

child under age 6 and $3,000 for every child ages 6 to 17. Though the expanded child tax credit was passed as part of a pandemic relief bill and was scheduled to be temporary, President Biden proposed making it permanent as a policy to reduce childhood poverty.

21-3d In-Kind Transfers

Another way to help low-income households is to provide them directly with some of the goods and services they need to raise their living standards. For example, charities provide the poor with food, clothing, shelter, and toys at Christmas. And the government gives poor families food through the Supplemental Nutrition Assistance Program, or SNAP. This program, which replaced a similar one called food stamps, gives families a plastic card, like a debit card, that can be used to buy food at stores. The government also provides healthcare to many people with low income through a program called Medicaid.

Is it better to help the poor with these in-kind transfers or with cash payments? There is no clear answer.

Advocates of in-kind transfers say that such transfers are more reliable at helping those who are struggling. Among the poorest members of society, alcohol and drug

In the News — Poverty during the Pandemic

During the Covid-19 pandemic, U.S. policymakers offered a safety net that, by historical standards, was very generous.

Temporary Pandemic Safety Net Drives Poverty to a Record Low

By Jason DeParle

Washington—The huge increase in government aid prompted by the coronavirus pandemic will cut poverty nearly in half this year from prepandemic levels and push the share of Americans in poverty to the lowest level on record, according to the most comprehensive analysis yet of a vast but temporary expansion of the safety net.

The number of poor Americans is expected to fall by nearly 20 million from 2018 levels, a decline of almost 45 percent. The country has never cut poverty so much in such a short period of time, and the development is especially notable since it defies economic headwinds—the economy has nearly seven million fewer jobs than it did before the pandemic.

The extraordinary reduction in poverty has come at extraordinary cost, with annual spending on major programs projected to rise fourfold to more than $1 trillion. Yet without further expensive new measures, millions of families may find the escape from poverty brief. The three programs that cut poverty most—stimulus checks, increased food stamps and expanded unemployment insurance—have ended or are scheduled to soon revert to their prepandemic size.

While poverty has fallen most among children, its retreat is remarkably broad: It has dropped among Americans who are white, Black, Latino and Asian, and among Americans of every age group and residents of every state.

"These are really large reductions in poverty—the largest short-term reductions we've seen," said Laura Wheaton of the Urban Institute, who produced the estimate with her colleagues Linda Giannarelli and Ilham Dehry. The institute's simulation model is widely used by government agencies. The New York Times requested the analysis, which expanded on an earlier projection.

The finding—that poverty plunged amid hard times at huge fiscal costs—comes at a moment of sharp debate about the future of the safety net.

The Biden administration has started making monthly payments to most families with children through an expansion of the child tax credit. Democrats want to make the yearlong effort permanent, which would reduce child poverty on a continuing basis by giving their families an income guarantee.

Progressives said the new numbers vindicated their contention that poverty levels reflected political choices and government programs could reduce economic need.

"Wow—these are stunning findings," said Bob Greenstein, a longtime proponent of safety net programs who is now at the Brookings Institution. "The policy response since the start of the pandemic goes beyond anything we've ever done, and the antipoverty effect dwarfs what most of us thought was possible."

Conservatives say that pandemic-era spending is unsustainable and would harm

addiction are more common than in society as a whole. By providing food, shelter, and healthcare directly, society can be more confident that it is not supporting such addictions but instead providing what the recipients truly need.

Advocates of cash payments say that in-kind transfers are inefficient and disrespectful. The government does not know what goods and services low-income households need most. Many of the poor are ordinary people down on their luck. Despite their misfortune, they are in the best position to decide how to raise their own living standards. Rather than giving people in-kind transfers of goods and services they may not want, it may be better to give them cash and allow them to buy what they think they need most.

21-3e Antipoverty Programs and Work Incentives

Policies aimed at combating poverty can sometimes have the unintended effect of discouraging people from escaping poverty on their own. For example, suppose that people need an income of $25,000 to maintain a minimally decent standard of living, and the government, out of concern for the less fortunate, guarantees everyone that

the poor in the long run, arguing that unconditional aid discourages work and marriage. The child tax credit offers families up to $300 per child a month whether or not parents have jobs, which critics call a return to failed welfare policies.

"There's no doubt that by shoveling trillions of dollars to the poor, you can reduce poverty," said Robert Rector of the Heritage Foundation. "But that's not efficient and it's not good for the poor because it produces social marginalization. You want policies that encourage work and marriage, not undermine it."

Poverty rates had reached new lows before the pandemic, Mr. Rector added, under policies meant to discourage welfare and promote work.

To understand how large the recent aid expansion has been, consider the experience of Kathryn Goodwin, a single mother of five in St. Charles, Mo., who managed a group of trailer parks before the pandemic eliminated her $33,000 job.

Without the pandemic-era expansions—passed in three rounds under both the Trump and Biden administrations—Ms. Goodwin's job loss would have caused her income to plunge to about $29,000 (in jobless benefits, food stamps and other aid), leaving her officially poor.

Instead, her income rose above its prepandemic level, though she has not worked for a year. She received about $25,000 in unemployment benefits (about three times what she would have received before the pandemic) and $12,000 in stimulus checks. With increased food stamp benefits and other help, her income grew to $67,000—almost 30 percent more than when she had a job.

"Without that help, I literally don't know how I would have survived," she said. "We would have been homeless."

Still, Ms. Goodwin, 29, has mixed feelings about large payments with no stipulations.

"In my case, yes, it was very beneficial," she said. But she said that other people she knew bought big TVs and her former boyfriend bought drugs. "All this free money enabled him to be a worse addict than he already was," she said. "Why should taxpayers pay for that?"

The Urban Institute's projections show poverty falling to 7.7 percent this year from 13.9 percent in 2018. That decline, 45 percent, is nearly three times the previous three-year record, according to historical estimates by researchers at Columbia University. The projected drop in child poverty, to 5.6 from 14.2 percent, amounts to a decline of 61 percent. That exceeds the previous 50 years combined, the Columbia figures show….

Jessica Moore of St. Louis said the expanded aid helped her make a fresh start.

A single mother of three, Ms. Moore, 24, lost work as a banquet server at the pandemic's start but received enough in unemployment insurance and stimulus checks to buy a car and enroll in community college. She is studying to become an emergency medical technician, which promises to raise her earnings 50 percent.

"When you lose your job, you don't expect benefits that are more than you were making," she said. "It was a pure blessing." ■

Questions to Discuss

1. Do you think the unusual circumstances of the Covid pandemic called for a more generous safety net than is available in normal times? Why or why not?
2. Do you think the experiences described in the article argue for a more expansive safety net in normal times? Why or why not?

Source: *New York Times*, July 29, 2021.

income. Whatever a person earns, the government makes up the difference between that income and $25,000. What effect would you expect this policy to have?

The incentive effects of this policy are obvious: Anyone who would make less than $25,000 by working has little incentive to find and keep a job. For every dollar a person would earn, the government would reduce the income supplement by a dollar. In effect, the government taxes 100 percent of additional earnings. An effective marginal tax rate of 100 percent is a policy with a large deadweight loss.

This antipoverty program is hypothetical, but it is not entirely unrealistic. Welfare, Medicaid, SNAP, and the EITC are all programs aimed at helping those living in poverty, and they are all tied to income. As people's incomes increase, they become ineligible for these programs. When all these programs are taken together, effective marginal tax rates can be very high, sometimes even exceeding 100 percent, so low-income households are worse off when they earn more. By trying to help the less fortunate, the government discourages people from working. According to critics of antipoverty programs, these programs alter work attitudes and create a "culture of poverty."

The problem of disincentives might seem to have an easy solution: Reduce benefits to recipients more gradually as their incomes rise. For example, if people lose 30 cents of benefits for every dollar they earn, they face an effective marginal tax rate of 30 percent. This effective tax reduces work effort to some extent, but it does not eliminate the incentive to work completely.

The drawback of this solution is that it greatly increases the cost of programs to combat poverty. If benefits are phased out gradually as a person's income rises, then those just above the income threshold will also be eligible for substantial benefits. The more gradual the phase-out, the more people are eligible, and the more the program costs. Policymakers face a trade-off between burdening the poor with high effective marginal tax rates and burdening taxpayers with costly programs to reduce poverty.

There are other ways to reduce the work disincentive of antipoverty programs. One is to require any person collecting benefits to be employed or accept a government-provided job—a system sometimes called **workfare**. But this approach raises the question of whether the government is the best institution to serve as the employer of last resort. Another possibility is to provide benefits for only a limited period. This route was taken in the 1996 welfare reform bill, which imposed a five-year lifetime limit on benefits for welfare recipients. When President Clinton signed the bill, he said that welfare should be "a second chance, not a way of life." Yet cutting people off from benefits isn't a solution to poverty if they lack the skills needed to land well-paying jobs. That's why these issues are still with us, despite decades of efforts to solve them.

QuickQuiz

7. A negative income tax is a policy under which
   a. all people with low income get government transfers.
   b. the government raises tax revenue without distorting incentives.
   c. everyone pays less than under a conventional income tax.
   d. some taxpayers are on the wrong side of the Laffer curve.

8. If the benefits from an antipoverty program are phased out as an individual's income increases, the program will
   a. encourage greater work effort from the poor.
   b. lead to an excess supply of labor among unskilled workers.
   c. cost the government more than a program that benefits everyone.
   d. increase the effective marginal tax rate that the poor face.

Answers are at the end of the chapter.

21-4 Conclusion

People have long reflected on the distribution of income in society. Plato, the ancient Greek philosopher, said that in an ideal society, the highest income would be no more than four times the lowest. Measuring inequality is difficult, but it is clear that most nations around the world, especially the United States, have much more inequality than Plato recommended.

One of the **Ten Principles of Economics** in Chapter 1 is that governments can sometimes improve market outcomes. This principle is important when considering the distribution of income. Even when the allocation of resources reached by the invisible hand is efficient, it is usually far from equal, and it is not necessarily fair. Yet there is no broad consensus about what fairness means or how much the government should redistribute income. Lawmakers often debate the progressivity of the tax code and the generosity of the social safety net. Economics alone cannot settle the disagreement.

Two other of the **Ten Principles of Economics** in Chapter 1 are that people face trade-offs and that people respond to incentives. These principles are intertwined in discussions of economic inequality. When the government enacts policies to partly equalize incomes, it may distort incentives, alter behavior, and make the allocation of resources less efficient. As a result, policymakers face a trade-off between equality and efficiency. The more equally they slice the economic pie, the smaller it may become. This doesn't mean that policymakers should necessarily refrain from income redistribution. But it does suggest that they approach redistributive policies aware of their potential costs.

Chapter in a Nutshell

- Data on the distribution of income show a wide disparity in U.S. society. The richest fifth of families earns more than twelve times as much as the poorest fifth.
- Because in-kind transfers, tax credits, the economic life cycle, transitory income, and economic mobility are so important for understanding variation in living standards, it is hard to gauge the degree of inequality in society using data on the distribution of income in a single year. When these other factors are considered, they tend to suggest that economic well-being is more equally distributed than annual income.
- Political philosophers differ in their views about the role of government in altering the distribution of income. Utilitarians (such as John Stuart Mill) would choose the distribution of income that maximizes the sum of utility of everyone in society. Liberal contractarians (such as John Rawls) would determine the distribution of income as if we were behind a "veil of ignorance" that prevented us from knowing our stations in life. Libertarians (such as Robert Nozick) would have the government enforce individual rights to ensure a fair process but then would not be concerned about inequality in the resulting distribution of income.
- Various policies aim to help people with low income—minimum-wage laws, welfare, negative income taxes, and in-kind transfers. While these policies help alleviate poverty, they can have unintended side effects. Because financial assistance declines as income rises, the poor often face very high effective marginal tax rates, which discourage them from escaping poverty on their own.

Key Concepts

poverty rate, p. 424
poverty line, p. 424
in-kind transfers, p. 427
life cycle, p. 427
permanent income, p. 428
utilitarianism, p. 430
utility, p. 430
liberal contractarianism, p. 431
maximin criterion, p. 432
social insurance, p. 432
libertarianism, p. 432
welfare, p. 434
negative income tax, p. 435

Questions for Review

1. Does the richest fifth of the U.S. population earn closer to three, six, or twelve times the income of the poorest fifth?
2. What has happened to the income share of the richest fifth of the U.S. population over the past 50 years?
3. What groups in the U.S. population are most likely to live in poverty?
4. When gauging the amount of inequality, why do transitory and life cycle variations in income cause difficulties?
5. How would a utilitarian, a liberal contractarian, and a libertarian each determine how much income inequality is permissible?
6. What are the pros and cons of in-kind (rather than cash) transfers to low-income households?
7. Describe how antipoverty programs can discourage people from working. How might you reduce this disincentive? What are the disadvantages of your proposed policy?

Problems and Applications

1. Table 2 shows that income inequality in the United States has increased since 1970. Some factors contributing to this increase were discussed in Chapter 20. What are they?
2. Table 3 shows that the percentage of children in families with income below the poverty line far exceeds the percentage of the elderly in such families. How might the allocation of government money across different social programs contribute to this phenomenon?
3. This chapter discusses the importance of economic mobility.
   a. What policies might the government pursue to increase economic mobility from year to year within a generation?
   b. What policies might the government pursue to increase economic mobility from generation to generation?
   c. Do you think we should reduce spending on current welfare programs to increase spending on programs that enhance economic mobility? What are some of the advantages and disadvantages of doing so?
4. Consider two communities. In one community, ten families have incomes of $100,000 each, and ten families have incomes of $20,000 each. In the other community, ten families have incomes of $250,000 each, and ten families have incomes of $25,000 each.
   a. In which community is the distribution of income more unequal? In which community is the problem of poverty likely to be worse?
   b. Which distribution of income would Rawls prefer? Explain.
   c. Which distribution of income do you prefer? Explain.
   d. Why might someone have the opposite preference?
5. This chapter uses the analogy of a "leaky bucket" to explain one constraint on the redistribution of income.
   a. What elements of the U.S. system for redistributing income create the leaks in the bucket? Be specific.
   b. Between Republicans and Democrats, who do you think generally believes that the bucket used for redistributing income is leakier? How does that belief affect their views about the amount of income redistribution the government should undertake?

6. Suppose there are two possible income distributions in a society of ten people. In the first distribution, nine people have incomes of $60,000, and one person has an income of $20,000. In the second distribution, all ten people have incomes of $50,000.
   a. If the society had the first income distribution, what would be the utilitarian argument for redistributing income?
   b. Which income distribution would Rawls consider more equitable? Explain.
   c. Which income distribution would Nozick consider more equitable? Explain.
7. The poverty rate would be substantially lower if the market value of in-kind transfers were added to family income. The largest in-kind transfer is Medicaid, the government health program for those with low incomes. Let's say the program costs $10,000 per recipient family.
   a. If the government gave each recipient family a $10,000 check instead of enrolling them in the Medicaid program, do you think that most of these families would spend that money to purchase health insurance? Why? (Recall that the poverty level for a family of four is about $25,000.)
   b. How does your answer to part (a) affect your view about whether we should determine the poverty rate by valuing in-kind transfers at the price the government pays for them? Explain.
   c. How does your answer to part (a) affect your view about whether we should provide assistance to low-income families in the form of cash transfers or in-kind transfers? Explain.
8. Consider two of the income security programs in the United States: Temporary Assistance for Needy Families (TANF) and the Earned Income Tax Credit (EITC).
   a. When a woman with children and very low income earns an extra dollar, she receives less in TANF benefits. What do you think is the effect of this feature of TANF on the labor supply of low-income women? Explain.
   b. The EITC provides greater benefits as low-income workers earn more income (up to a point). What do you think is the effect of this program on the labor supply of low-income individuals? Explain.
   c. What are the disadvantages of eliminating TANF and allocating the savings to the EITC?

QuickQuiz Answers

1. **b** 2. **c** 3. **a** 4. **d** 5. **b** 6. **c** 7. **a** 8. **d**

Chapter

22

The Theory of Consumer Choice

When you walk into a big store or click on an online shopping site, you are confronted with thousands of goods you might buy. Many things may appeal to you, but because your financial resources are limited, you probably can't buy all of them. So you look at the prices and buy a selection of items that, given your resources, best suits your needs and desires.

This chapter develops a theory that describes how people make these decisions. So far, this book has summarized consumer behavior with the demand curve. As we have seen, the demand curve reflects consumers' willingness to pay for a good. When the price of the good rises, consumers are willing to pay for fewer units, so the quantity demanded falls. We now look more closely at the decisions that lie behind the demand curve. The theory of consumer choice presented here provides a deeper understanding of demand, just as the theory of the competitive firm in Chapter 15 provides a deeper understanding of supply.

One of the **Ten Principles of Economics** in Chapter 1 is that people face trade-offs. This principle is the essence of the theory of consumer choice. When consumers buy more of one good, they can afford less of other goods they want. When they spend more time enjoying leisure and less time working, they earn less and consume less. When they spend more of their income now, they reduce their saving and have less to spend in the future. The theory of consumer choice examines how people facing these trade-offs make decisions and how they respond to changes in their environment.

The theory is useful in analyzing a wide range of issues. After developing the basic framework, we ask:

- Do all demand curves slope downward?
- How do wages affect labor supply?
- How do interest rates affect household saving?

These questions might seem unrelated, but the theory of consumer choice can help answer each of them.

22-1 The Budget Constraint: What a Consumer Can Afford

Other things being equal, most people would like to consume more—to drive nicer cars, wear trendier clothes, eat at better restaurants, or take grander vacations. People consume less than they desire because their spending is **constrained**, or limited, by their income. Our study of consumer choice begins with this constraint.

22-1a Representing Consumption Opportunities in a Graph

Consider Consuela, a consumer who buys only two goods: pizza and Pepsi. In the real world, people buy hundreds of different kinds of goods, and a diet of only pizza and Pepsi would not be healthy. But assuming only two goods simplifies the model without altering the basic insights about consumer choice.

Consuela's income is $1,000 per month, and because she is insatiable, she spends all of it on pizza and Pepsi. The price of a pizza is $10, and the price of a liter of Pepsi is $2. Her income and these market prices constrain her spending.

The table in Figure 1 shows some of the combinations of pizza and Pepsi that Consuela can buy. The first row shows that if she spends all her income on pizza, she can eat 100 pizzas during the month, but she would not be able to buy any Pepsi at all. The second row shows another possible consumption bundle: 90 pizzas and 50 liters of Pepsi. And so on. Each consumption bundle in the table costs exactly $1,000.

The graph in Figure 1 illustrates the consumption bundles that Consuela can choose. The vertical axis measures the number of liters of Pepsi, and the horizontal axis measures the number of pizzas. Three points are marked on this figure. At point A, Consuela buys no Pepsi and consumes 100 pizzas. At point B, she buys no pizza and consumes 500 liters of Pepsi. At point C, she buys 50 pizzas and 250 liters of Pepsi. Point C, which is exactly at the middle of the line from A to B, is the point at which Consuela spends an equal amount ($500) on the two goods. These are only three of the many combinations of pizza and Pepsi that she can choose. All the points on the line from A to B are possible. This line, called the **budget constraint**, shows the consumption bundles that a consumer can afford. In this case, it shows the trade-off between pizza and Pepsi that Consuela faces.

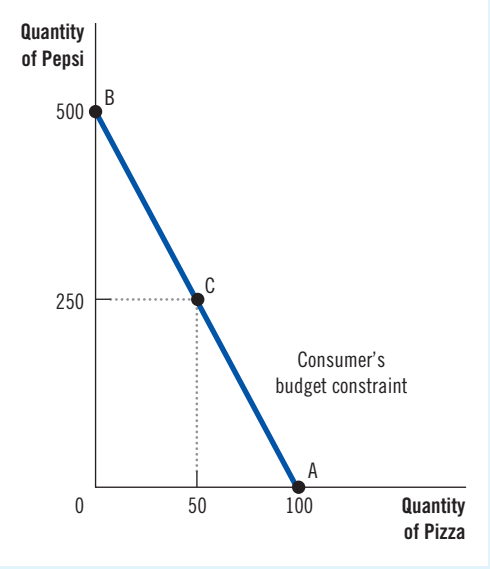
budget constraint the limit on the consumption bundles that a consumer can afford

Figure 1

The Consumer's Budget Constraint

The budget constraint shows the bundles of goods that the consumer can buy with a given income. Here, she buys bundles of pizza and Pepsi. The table and graph show what the consumer can afford if her income is $1,000, the price of pizza is $10, and the price of Pepsi is $2.

Number of Pizzas	Liters of Pepsi	Spending on Pizza	Spending on Pepsi	Total Spending
100	0	$1,000	$ 0	$1,000
90	50	900	100	1,000
80	100	800	200	1,000
70	150	700	300	1,000
60	200	600	400	1,000
50	250	500	500	1,000
40	300	400	600	1,000
30	350	300	700	1,000
20	400	200	800	1,000
10	450	100	900	1,000
0	500	0	1,000	1,000



The slope of the budget constraint measures the rate at which the consumer can trade one good for the other. Recall that the slope between two points is calculated as the change in the vertical distance divided by the change in the horizontal distance ("rise over run"). From point A to point B, the vertical distance is 500 liters, and the horizontal distance is 100 pizzas, so the slope is 5 liters per pizza. (Actually, because the budget constraint slopes downward, the slope is a negative number. But for our purposes, we can ignore the minus sign.)

Notice that the slope of the budget constraint equals the **relative price** of the two goods—the price of one good compared with the price of the other. A pizza costs five times as much as a liter of Pepsi, so the opportunity cost of a pizza is 5 liters of Pepsi. The budget constraint's slope of 5 reflects the trade-off the market is offering Consuela: 1 pizza for 5 liters of Pepsi.

22-1b Shifts in the Budget Constraint

The budget constraint shows the opportunities available to Consuela. It is drawn given her income and the prices of the two goods. If her income or the prices change, the budget constraint shifts. Consider three examples of how such a shift might occur.

Suppose first that Consuela's income increases from $1,000 to $2,000 while prices remain the same. With higher income, she can afford more of both goods. The increase in income shifts the budget constraint outward, as in panel (a) of Figure 2. Because the relative price of the two goods has not changed, the slope of the new budget constraint is the same as the slope of the initial budget constraint. That is, an increase in income leads to a parallel shift in the budget constraint.

Now suppose that the price of Pepsi falls from $2 to $1 while Consuela's income remains at $1,000 and the price of pizza remains at $10. If she spends her entire income on pizza, the price of Pepsi is irrelevant. In this case, she can still buy only 100 pizzas, so the point on the horizontal axis representing 100 pizzas and 0 liters of Pepsi stays the same. But as long as she is buying some Pepsi, the lower price of Pepsi expands her set of opportunities. The budget constraint shifts outward, as shown in panel (b) of Figure 2. The lower price allows her to buy the same amount of pizza as before and more Pepsi, the same amount of Pepsi as before and more pizza, or more of both goods.

Because the slope reflects the relative price of pizza and Pepsi, it changes when the price of Pepsi falls. With the lower price of Pepsi, Consuela can now trade a pizza for 10 liters of Pepsi rather than 5. As a result, the new budget constraint is steeper. The expansion in her opportunities is represented by a rotational shift rather than a parallel shift.

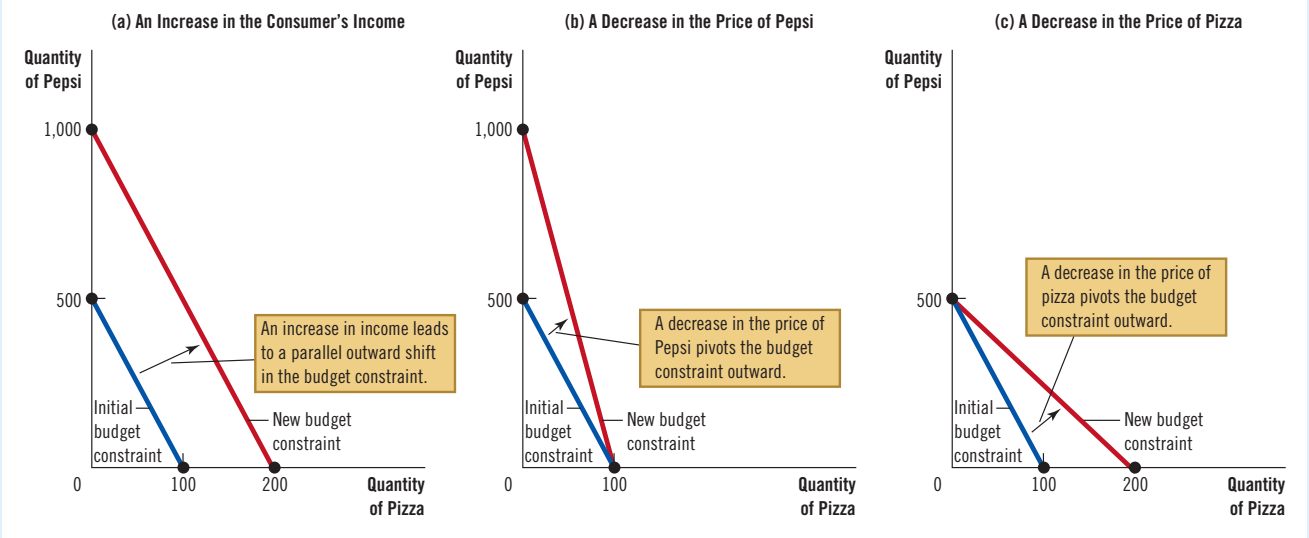
For the third example, suppose that the price of pizza falls from $10 to $5 while Consuela's income remains at $1,000 and the price of Pepsi remains at $2. Once again, the lower price expands her set of opportunities and leads to a rotational outward shift in the budget constraint, as shown in panel (c) of Figure 2. Now, with the lower price of pizza, Consuela can trade a pizza for 2.5 liters of Pepsi rather than 5, and the budget constraint becomes flatter.

Figure 2 shows what happens when a higher income or a lower price expands Consuela's set of opportunities. The opposite occurs when a lower income or a higher price reduces her opportunities. The pictures look much the same as Figure 2, but

Figure 2

Shifts in the Consumer's Budget Constraint

In panel (a), an increase in the consumer's income shifts the budget constraint outward. The slope remains the same because the relative price of pizza and Pepsi has not changed. In panel (b), a decrease in the price of Pepsi shifts the budget constraint outward, while in panel (c), a decrease in the price of pizza shifts the budget constraint outward. In these two cases, the slope changes because the relative price of pizza and Pepsi has changed.



with the direction of the arrows reversed. And when more than one change occurs at the same time, we can analyze the overall impact by graphing and comparing the initial and final budget constraints.

QuickQuiz

1. Homer buys hamburgers for $10 and root beer for $2. He has income of $100. His budget constraint will shift inward if
   a. the price of hamburgers rises to $12.
   b. the price of root beer falls to $1.
   c. his income rises to $150.
   d. the price of hamburgers, the price of root beer, and his income all rise by 50 percent.
2. Marge also buys hamburgers for $10 and root beer for $2. She has income of $200. Her budget constraint will experience a **parallel** outward shift if
   a. the price of hamburgers falls to $5, the price of root beer falls to $1, and her income falls to $100.
   b. the price of hamburgers rises to $20, the price of root beer rises to $4, and her income remains the same.
   c. the price of hamburgers falls to $8, the price of root beer falls to $1, and her income rises to $240.
   d. the price of hamburgers rises to $20, the price of root beer rises to $4, and her income rises to $500.

Answers are at the end of the chapter.

22-2 Preferences: What a Consumer Wants

The goal of this chapter is to understand how consumers make choices. The budget constraint is one piece of the analysis: It shows the combinations of goods that consumers can afford given their income and the prices of the goods. Their choices, however, depend not only on their budget constraint but also on their preferences.

22-2a Representing Preferences with Indifference Curves

Consuela's preferences allow her to choose among different bundles of pizza and Pepsi. If you offer her two different bundles, she chooses the one that best suits her tastes. If the two bundles suit her tastes equally well, we say that Consuela is **indifferent** between the two bundles.

Just as we have represented Consuela's budget constraint graphically, we can also represent her preferences graphically. We do this with indifference curves. An **indifference curve** shows the various bundles of consumption that make a consumer equally happy. In this case, the indifference curves show the combinations of pizza and Pepsi with which Consuela is equally satisfied.

indifference curve
a curve that shows consumption bundles that give the consumer the same level of satisfaction

Figure 3 shows two of Consuela's many indifference curves. We can see that she is indifferent among combinations A, B, and C because they are all on the same curve. So if her consumption of pizza decreases, say, from point A to point B, her consumption of Pepsi must increase to keep her equally happy. If her consumption of pizza decreases again, from point B to point C, the amount of Pepsi consumed must increase yet again.

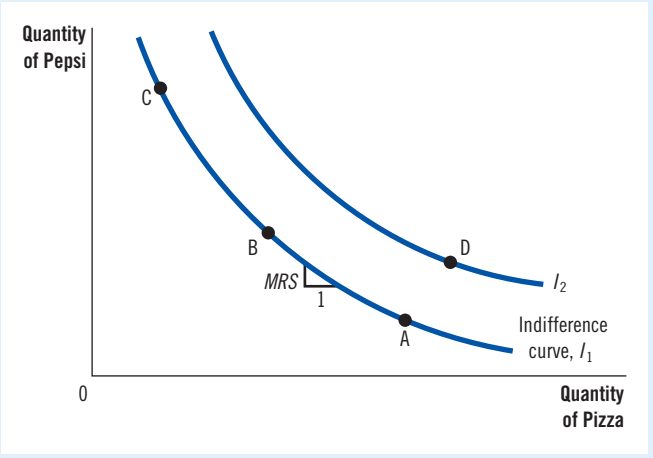
The slope at any point on an indifference curve equals the rate at which Consuela is willing to substitute one good for the other. (The slope is negative, but we can ignore the minus sign.) This rate is called the **marginal rate of substitution** (*MRS*). In this case, it measures how much additional Pepsi Consuela requires to be compensated for a one-unit reduction in pizza consumption.

marginal rate of substitution
the rate at which a consumer is willing to trade one good for another

Figure 3

The Consumer's Preferences

The consumer's preferences are represented with indifference curves, which show the combinations of pizza and Pepsi that make the consumer equally satisfied. Because the consumer prefers more of a good, points on a higher indifference curve (I_2) are preferred to points on a lower indifference curve (I_1). The marginal rate of substitution (*MRS*) shows the rate at which the consumer is willing to trade Pepsi for pizza. It measures the quantity of Pepsi the consumer must receive in exchange for 1 pizza.



Because the indifference curves are not straight lines, the marginal rate of substitution is not the same at all points on a given indifference curve. The rate at which Consuela is willing to trade one good for the other depends on the quantities of the goods she is already consuming. In other words, the rate at which she is willing to trade pizza for Pepsi depends on whether she is hungrier or thirstier, and her hunger and thirst, in turn, depend on her current consumption of pizza and Pepsi.

Consuela is equally happy at all points on a given indifference curve, but she prefers some indifference curves to others. Because she prefers more consumption to less, higher indifference curves are preferred to lower ones. In Figure 3, any point on curve I_2 is preferred to any point on curve I_1.

Consuela's set of indifference curves gives a complete ranking of her preferences. That is, we can use the indifference curves to rank any two bundles of goods. For example, the indifference curves tell us that Consuela prefers the bundle at point D to the bundle at point A because point D is on a higher indifference curve than point A. (That may be obvious, however, because point D offers more of both goods.) The indifference curves also tell us that Consuela prefers the bundle at point D to the bundle at point C because point D is on a higher indifference curve. Even though point D has less Pepsi than point C, it has more than enough extra pizza to make her prefer it. By seeing which point is on the higher indifference curve, we can use the set of indifference curves to rank any combinations of pizza and Pepsi.

22-2b Four Properties of Indifference Curves

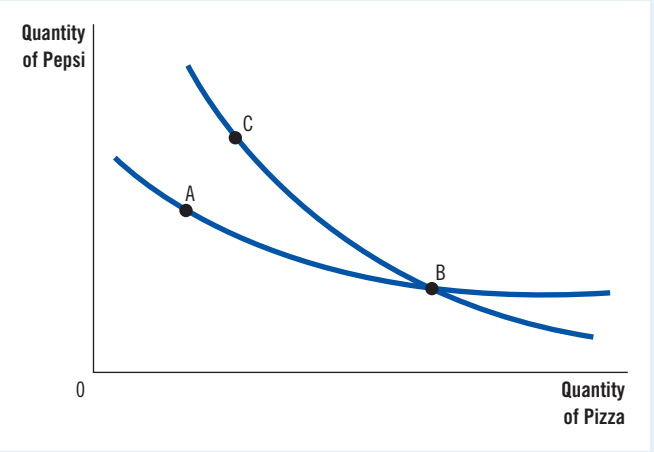
Because indifference curves represent a consumer's preferences, their properties reflect those preferences. Here are four properties that describe most indifference curves:

- **Property 1: Higher indifference curves are preferred to lower ones.** People usually prefer to consume more rather than less. This preference for greater quantities is reflected in the indifference curves. As Figure 3 shows, higher indifference curves represent larger quantities of goods

Figure 4

The Impossibility of Intersecting Indifference Curves

This situation can never happen. According to these indifference curves, the consumer would be equally satisfied at points A, B, and C, even though point C has more of both goods than point A.



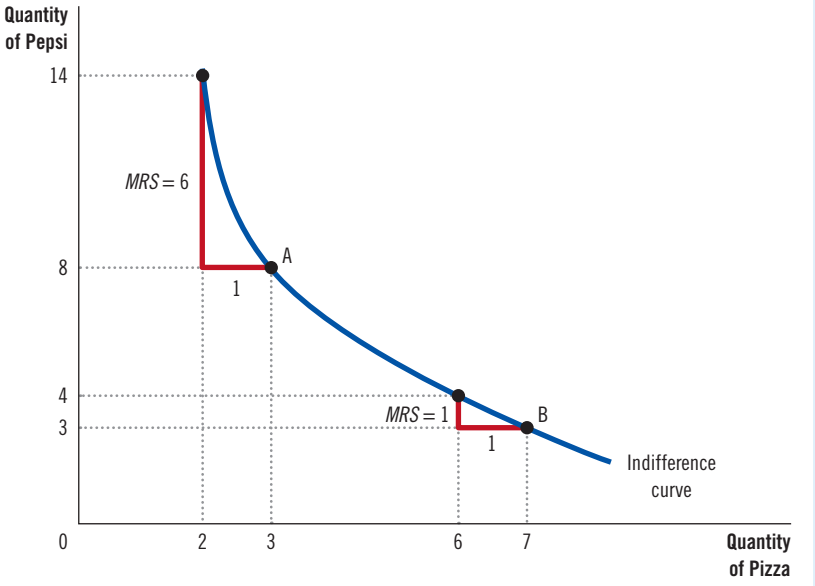
than lower indifference curves. Thus, a consumer prefers being on higher indifference curves.

- **Property 2: Indifference curves slope downward.** The slope of an indifference curve reflects the rate at which a consumer is willing to substitute one good for the other. In most cases, the consumer likes both goods, so if the quantity of one good decreases, the quantity of the other good must increase for the consumer to be equally happy. For this reason, most indifference curves slope downward.
- **Property 3: Indifference curves do not cross.** To see why this is true, suppose that two indifference curves did cross, as in Figure 4. Then, because point A is on the same indifference curve as point B, the two points would make the consumer equally happy. In addition, because point B is on the same indifference curve as point C, these two points would make the consumer equally happy. But this means that points A and C would also make the consumer equally happy, even though point C has more of both goods. This contradicts our assumption that the consumer always prefers more of both goods to less. Thus, indifference curves cannot cross.
- **Property 4: Indifference curves are bowed inward.** The slope of an indifference curve is the marginal rate of substitution—the rate at which the consumer will trade one good for the other. The marginal rate of substitution (*MRS*) usually depends on the amount of each good being consumed. Because people are more willing to trade away goods that they have in abundance and are less willing to trade away goods of which they have little, the indifference curves are bowed inward toward the graph's origin. As an example, consider Figure 5. At point A, Consuela has a lot of Pepsi and only a little pizza, so she is hungry but not thirsty. To give up 1 pizza, she would have to receive 6 liters of Pepsi: The *MRS* is 6 liters of Pepsi per pizza. By contrast, at point B, Consuela has a little Pepsi and a lot of pizza, so she is thirsty but not hungry. At this point, she would be willing to give up 1 pizza to get 1 liter of Pepsi: The *MRS* is

Figure 5

Bowed Indifference Curves

Indifference curves are usually bowed inward. This shape implies that the marginal rate of substitution (*MRS*) depends on the quantity of the two goods the consumer is currently consuming. At point A, the consumer has a little pizza and a lot of Pepsi, so she requires a lot of extra Pepsi to induce her to give up one of the pizzas: The *MRS* is 6 liters of Pepsi per pizza. At point B, the consumer has a lot of pizza and a little Pepsi, so she requires only a little extra Pepsi to induce her to give up one of the pizzas: The *MRS* is 1 liter of Pepsi per pizza.



1 liter of Pepsi per pizza. Thus, the bowed shape of the indifference curve reflects Consuela's greater willingness to give up something that she already has in abundance.

22-2c Two Extreme Examples of Indifference Curves

The shape of an indifference curve reveals a consumer's willingness to trade one good for the other. When the goods are easy to substitute for each other, the indifference curves are less bowed; when the goods are hard to substitute, the indifference curves are very bowed. To see why, consider two extreme cases.

Perfect Substitutes Suppose that someone offered you bundles of nickels and dimes. How would you rank the different bundles?

One possibility is that you would care only about the total monetary value of each bundle. If so, you would always be willing to trade 2 nickels for 1 dime. Your marginal rate of substitution between nickels and dimes would be a fixed number: *MRS* = 2, regardless of the number of nickels and dimes in the bundle.

We can represent these preferences for nickels and dimes with the indifference curves in panel (a) of Figure 6. Because the marginal rate of substitution is constant, the indifference curves are straight lines. In the case of straight indifference curves, the two goods are said to be **perfect substitutes**.

perfect substitutes
two goods with straight-line indifference curves

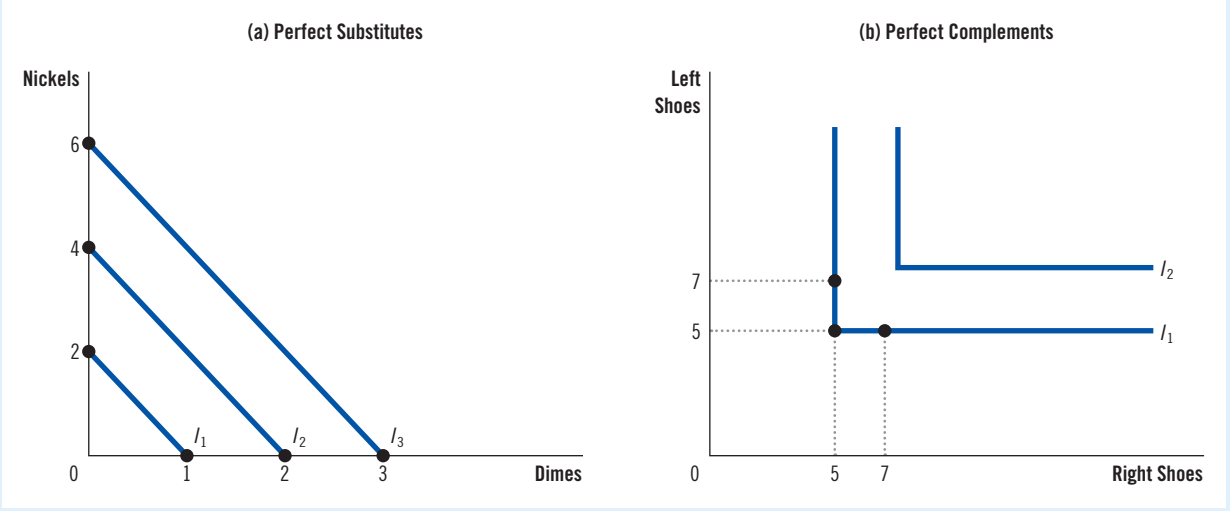
Perfect Complements Suppose now that someone offered you bundles of shoes. Some of the shoes fit your left foot, others your right foot. How would you rank these different bundles?

In this case, you might care only about how many pairs of shoes you have. Single shoes aren't much good. So you would judge a bundle based on the number of

Figure 6

Perfect Substitutes and Perfect Complements

When two goods are perfectly substitutable, such as nickels and dimes, the indifference curves are straight lines, as shown in panel (a). When two goods are perfectly complementary, such as left shoes and right shoes, the indifference curves are right angles, as shown in panel (b).



pairs you could assemble from it. A bundle of 5 left shoes and 7 right shoes yields only 5 pairs.

The indifference curves in panel (b) of Figure 6 represent these preferences. A bundle with 5 left shoes and 7 right shoes is just as good as a bundle with 7 left shoes and 5 right shoes. And these are equivalent to a bundle with 5 of each. The indifference curves are right angles. In the case of right-angle indifference curves, the two goods are said to be **perfect complements**.

perfect complements
two goods with right-angle indifference curves

In the real world, most goods are neither perfect substitutes (like nickels and dimes) nor perfect complements (like right and left shoes). Perfect substitutes and perfect complements are extreme cases. They are introduced here not because they are common but because they illustrate how indifference curves reflect a consumer's preferences. For most goods, the indifference curves are bowed inward, but not so bowed that they become right angles.

QuickQuiz

3. At two points on an indifference curve,
   a. the consumer has the same income.
   b. the consumer has the same marginal rate of substitution.
   c. the bundles of goods cost the consumer the same amount.
   d. the bundles of goods yield the consumer the same satisfaction.

4. At any point on an indifference curve, the slope of the curve measures the consumer's
   a. income.
   b. willingness to trade one good for the other.
   c. perception of the two goods as substitutes or complements.
   d. elasticity of demand.

Answers are at the end of the chapter.

22-3 Optimization: What a Consumer Chooses

We now have the building blocks of the theory of consumer choice: the budget constraint (which shows what consumers can afford) and preferences (which show what they want). Let's put them together to see what consumers choose.

22-3a The Consumer's Optimal Choices

Once again, consider Consuela, our pizza and Pepsi consumer. She would like to end up with the best bundle of pizza and Pepsi—the bundle that puts her on the highest possible indifference curve. But because Consuela's income limits her spending, she must also end up on or below her budget constraint, which measures the total resources available to her.

Figure 7 shows Consuela's budget constraint and three of her indifference curves. The highest indifference curve that she can reach (I_2 in the figure) is the one that just barely touches her budget constraint. The point at which this indifference curve and the budget constraint touch is the **optimum**. Consuela would prefer point A, but she can't afford that bundle of goods because it lies above her budget constraint. She can afford point B, but because that bundle of goods is on a lower indifference curve, it provides her with less satisfaction. The optimum represents the best bundle of pizza and Pepsi that Consuela can afford.

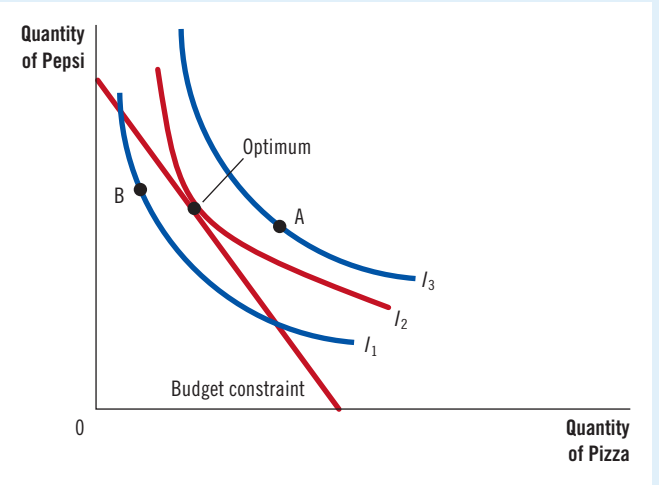
At the optimum, the slope of the indifference curve equals the slope of the budget constraint. We say that the indifference curve is **tangent** to the budget constraint. The slope of the indifference curve is the marginal rate of substitution between pizza and Pepsi, and the slope of the budget constraint is the relative price of pizza and Pepsi. This leads to an important conclusion: **The consumer chooses the quantities of the two goods so the marginal rate of substitution equals the relative price.**

Chapter 7 discussed how market prices reflect the marginal value that consumers place on goods. This analysis of consumer choice shows the same result in another way. In making her consumption choices, Consuela takes the relative price of the two goods as given and then chooses an optimum bundle of goods at which her marginal rate of substitution equals this relative price. The relative price is the rate at which the **market** is willing to trade one good for the other, while the marginal

Figure 7

The Consumer's Optimum

The consumer chooses the point on her budget constraint that lies on the highest indifference curve. Here, the highest indifference curve the consumer can reach is I_2. The consumer prefers point A, which lies on indifference curve I_3, but she can't afford this bundle of pizza and Pepsi. By contrast, point B is affordable, but because it lies on a lower indifference curve, she doesn't prefer it. At the optimum, the marginal rate of substitution equals the relative price of the two goods.



FYI

Utility: An Alternative Way to Describe Preferences and Optimization

We have used indifference curves to represent consumer preferences. Another common way to represent preferences is with the concept of **utility**. Utility is an abstract measure of the satisfaction or happiness that a consumer receives from a bundle of goods. Economists say that a consumer prefers one bundle of goods to another if it provides more utility than the other.

Indifference curves and utility are closely related. Because the consumer prefers points on higher indifference curves, bundles of goods on higher indifference curves provide higher utility. Because the consumer is equally happy with all points on the same indifference curve, all these bundles provide the same utility. You can think of an indifference curve as an "equal-utility" curve.

The **marginal utility** of any good is the increase in utility that the consumer gets from an additional unit of it. Most goods are assumed to exhibit **diminishing marginal utility**: The more of the good the consumer already has, the lower the marginal utility provided by an extra unit of that good.

The marginal rate of substitution between two goods depends on their marginal utilities. For example, if the marginal utility of good X is twice the marginal utility of good Y, then a person would need 2 units of Y to compensate for losing 1 unit of X, and the *MRS* equals 2. More generally, the marginal rate of substitution (and the slope of the indifference curve) equals the marginal utility of one good divided by the marginal utility of the other good.

Utility analysis provides another way to describe consumer optimization. Recall that, at the consumer's optimum, the marginal rate of substitution equals the ratio of prices. That is,

$$MRS = P_X/P_Y.$$

Because the marginal rate of substitution equals the ratio of marginal utilities, we can write this condition for optimization as

$$MU_X/MU_Y = P_X/P_Y.$$

Now rearrange this expression so it becomes

$$MU_X/P_X = MU_Y/P_Y.$$

This equation has a simple interpretation: At the optimum, the marginal utility per dollar spent on good X equals the marginal utility per dollar spent on good Y. If this equality did not hold, the consumer could increase her utility by spending less on the good that provided lower marginal utility per dollar and more on the good that provided higher marginal utility per dollar.

When economists discuss consumer choice, they sometimes express the theory using different words. One economist might say that the consumer's goal is to maximize utility. Another economist might say that the consumer's goal is to be on the highest possible indifference curve. The first economist would conclude that at the consumer's optimum, the marginal utility per dollar is the same for all goods, while the second would describe the optimum as the point at which the indifference curve is tangent to the budget constraint. These are two ways of saying the same thing. ■

rate of substitution is the rate at which the **consumer** is willing to trade one good for the other. At the optimum, Consuela's valuation of the two goods (as measured by the marginal rate of substitution) equals the market's valuation (as measured by the relative price). Put another way, when consumers optimize, the relative prices of goods in the marketplace reflect the relative value that consumers place on them.

22-3b How Changes in Income Affect the Consumer's Choices

Let's now examine how Consuela's consumption decision responds to changes in her income. To be specific, suppose her income increases. As we have seen, an increase in income leads to a parallel outward shift in the budget constraint, as in Figure 8. Because the relative price of the two goods has not changed, the slope of the new budget constraint is the same as the slope of the initial budget constraint.

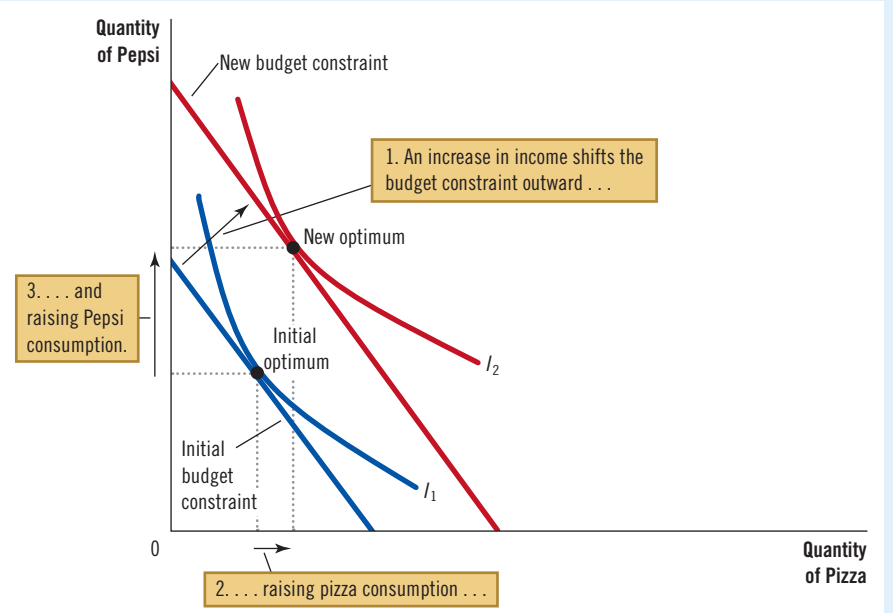
The expanded budget constraint allows Consuela to reach a higher indifference curve, reflecting a more desirable bundle of pizza and Pepsi. Given the shift in the budget constraint and Consuela's preferences, her optimum moves from the point labeled "initial optimum" to the point labeled "new optimum."

In Figure 8, Consuela chooses to consume more Pepsi **and** more pizza. The logic of the model does not require increased consumption of both goods in response to increased income, but this situation is the most common. As Chapter 4 discussed, if consumers want more of a good when their incomes rise, economists call it a

Figure 8

An Increase in Income

When the consumer's income rises, the budget constraint shifts outward. If both goods are normal goods, the consumer responds to the increase in income by buying more of both of them. Here, the consumer buys more pizza and more Pepsi.



normal good
a good for which an increase in income raises the quantity demanded

inferior good
a good for which an increase in income reduces the quantity demanded

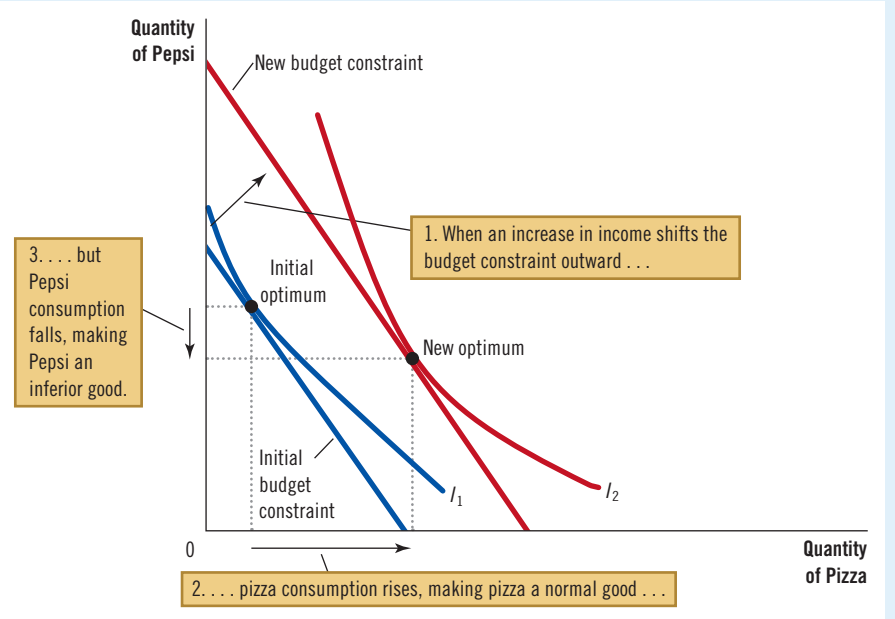
normal good. The indifference curves in Figure 8 are drawn assuming that both pizza and Pepsi are normal goods.

Figure 9 shows an example in which an increase in income induces Consuela to buy more pizza but less Pepsi. If consumers buy less of a good when their incomes rise, economists call it an **inferior good**. Figure 9 is drawn assuming that pizza is a normal good and Pepsi is an inferior good.

Figure 9

An Inferior Good

A good is inferior if the consumer buys less of it when her income rises. Here, Pepsi is an inferior good: When the consumer's income increases and the budget constraint shifts outward, the consumer buys more pizza but less Pepsi.



Although most goods in the world are normal goods, some are inferior goods. An example is bus rides. As income increases, consumers are more likely to own cars or use Uber and less likely to ride the bus. Bus rides, therefore, are an inferior good.

22-3c How Changes in Prices Affect the Consumer's Choices

Let's now use this model of consumer choice to consider how a change in the price of one of the goods alters Consuela's choices.

Suppose that the price of Pepsi falls. As noted earlier, a fall in the price of either good shifts the budget constraint outward and, by changing the relative price of the two goods, changes the slope of the budget constraint as well. Figure 10 shows how the fall in the price of Pepsi rotates the budget constraint and changes the optimum.

How such a change in the budget constraint alters the quantities of the two goods purchased depends on Consuela's preferences. For the indifference curves in this figure, she buys more Pepsi and less pizza. But it takes only a little creativity to draw indifference curves with other outcomes. A consumer could plausibly respond to the lower price of Pepsi by buying more of both goods.

22-3d Income and Substitution Effects

The impact of a change in the price of a good on the quantities purchased can be decomposed into two effects: an **income effect** and a **substitution effect**. When consumers respond to a price change, both effects are at work.

To see what these effects are, consider how Consuela might respond when she learns that the price of Pepsi has fallen. Two thoughts may occur to her:

- "Great news! Now that Pepsi is cheaper, my income has greater purchasing power. I am, in effect, richer than I was. Because I am richer, I can buy both more pizza and more Pepsi." (This is the income effect.)

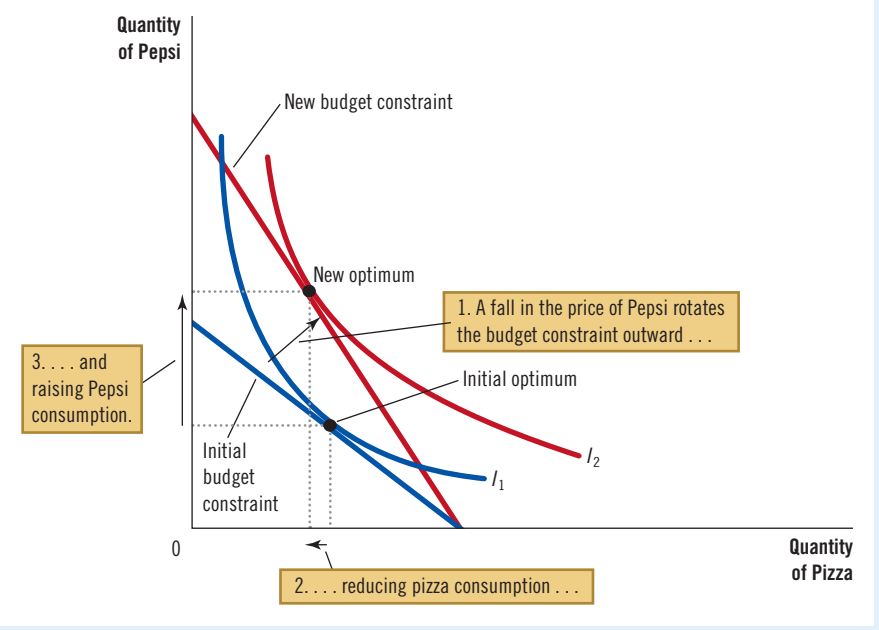
income effect
the change in consumption that results when a price change moves the consumer to a higher or lower indifference curve

substitution effect
the change in consumption that results when a price change moves the consumer along a given indifference curve to a point with a new marginal rate of substitution

Figure 10

A Change in Price

When the price of Pepsi falls, the consumer's budget constraint shifts outward and changes slope. The consumer moves from the initial optimum to the new optimum, which changes her purchases of both pizza and Pepsi. In this case, the quantity of Pepsi consumed rises, and the quantity of pizza consumed falls.



- "With the lower price of Pepsi, I get more liters of Pepsi for every pizza that I give up. Because pizza is now relatively more expensive, I should buy less pizza and more Pepsi." (This is the substitution effect.)

Both statements make sense. The decrease in the price of Pepsi makes Consuela better off. If pizza and Pepsi are both normal goods, she will want to spread this increase in her purchasing power over both goods. This income effect tends to make her buy more pizza and more Pepsi. Yet at the same time, the opportunity cost of pizza (in terms of forgone Pepsi) has risen, and the opportunity cost of Pepsi (in terms of forgone pizza) has fallen. This substitution effect tends to make Consuela choose less pizza and more Pepsi.

Now consider the result of these two effects occurring at the same time. Consuela certainly buys more Pepsi because the income and substitution effects both act to increase the consumption of Pepsi. But for pizza, the income and substitution effects push in opposite directions. As a result, it is not clear whether Consuela buys more pizza or less. The outcome could go either way, depending on the sizes of the income and substitution effects. Table 1 summarizes these conclusions.

The income and substitution effects can be interpreted using indifference curves. **The income effect is the change in consumption that results from the movement to a new indifference curve. The substitution effect is the change in consumption that results from moving to a new point on the same indifference curve with a different marginal rate of substitution.**

Figure 11 shows how to decompose the change in Consuela's decision into the income effect and the substitution effect. When the price of Pepsi falls, she moves from the initial optimum, point A, to the new optimum, point C. We can imagine this change as occurring in two steps. First, Consuela moves **along** the initial indifference curve, I_1, from point A to point B. She is equally happy at these two points, but at point B, the marginal rate of substitution reflects the new relative price. (The dashed line through point B is parallel to the new budget constraint and reflects the new relative price.) Next, Consuela **shifts** to the higher indifference curve, I_2, by moving from point B to point C. Even though point B and point C are on different indifference curves, they have the same marginal rate of substitution. That is, the slope of the indifference curve I_1 at point B equals the slope of the indifference curve I_2 at point C.

Table 1

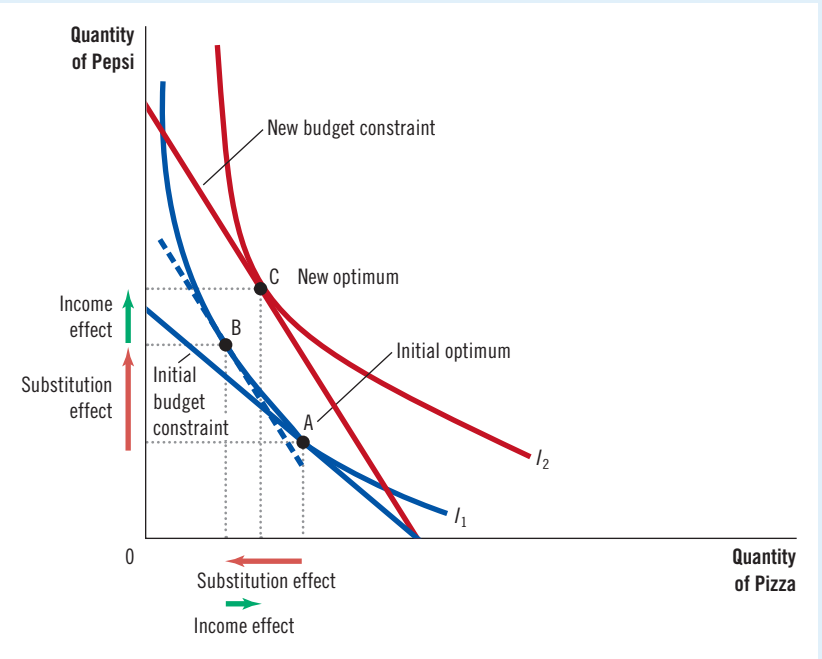
Income and Substitution Effects When the Price of Pepsi Falls

Good	Income Effect	Substitution Effect	Total Effect
Pepsi	Consumer is richer, so she buys more Pepsi.	Pepsi is relatively cheaper, so consumer buys more Pepsi.	Income and substitution effects act in the same direction, so consumer buys more Pepsi.
Pizza	Consumer is richer, so she buys more pizza.	Pizza is relatively more expensive, so consumer buys less pizza.	Income and substitution effects act in opposite directions, so the total effect on pizza consumption is ambiguous.

Figure 11

Income and Substitution Effects

The effect of a change in price can be broken down into an income effect and a substitution effect. The substitution effect—the movement along an indifference curve to a point with a different marginal rate of substitution—is the change from point A to point B along indifference curve I_1. The income effect—the shift to a higher indifference curve—is the change from point B on indifference curve I_1 to point C on indifference curve I_2.



Consuela never actually chooses point B, but this hypothetical point is useful to clarify the two effects that determine her decision. The change from point A to point B represents a pure change in the marginal rate of substitution without any change in Consuela's welfare. And the change from point B to point C represents a pure change in welfare without any change in the marginal rate of substitution. The movement from A to B shows the substitution effect, and the movement from B to C shows the income effect.

22-3e Deriving the Demand Curve

A consumer's demand curve for any good shows the quantity demanded at any price. In essence, it summarizes the optimal decisions that arise from the consumer's budget constraint and preferences.

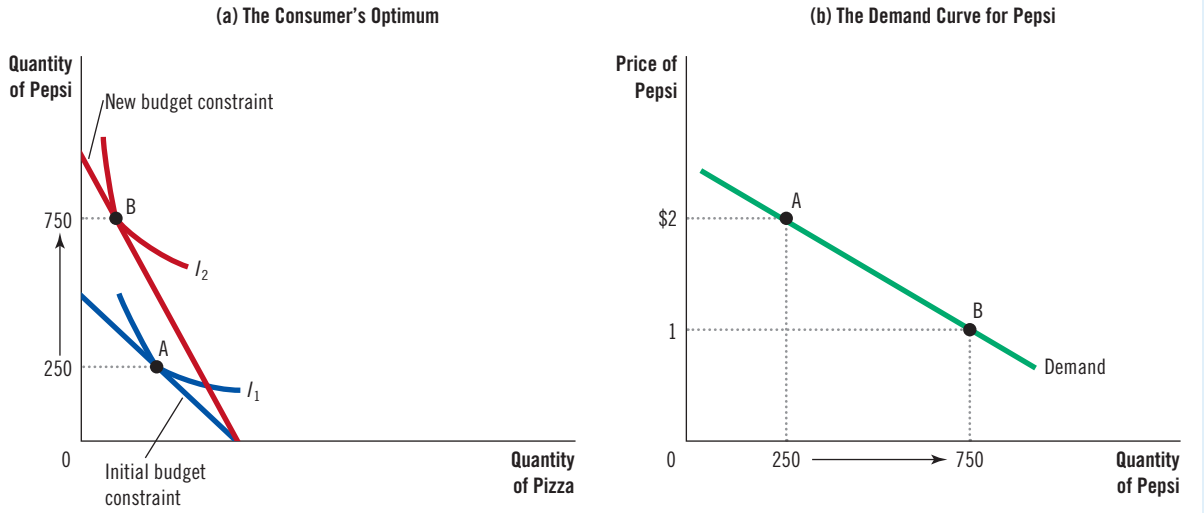
For example, Figure 12 considers Consuela's demand for Pepsi. Panel (a) shows that when the price of a liter falls from $2 to $1, her budget constraint shifts outward. Because of both income and substitution effects, she increases her purchases of Pepsi from 250 to 750 liters. Panel (b) shows the demand curve that results from these decisions. In this way, the theory of consumer choice provides the theoretical foundation for an individual's demand curve.

It may be comforting to know that the demand curve arises naturally from the theory of consumer choice, but this exercise by itself does not justify developing the theory. There is no need for a rigorous analytic framework just to establish that people respond to price changes. The theory of consumer choice is, however, useful in studying various decisions that people make as they go about their lives, as we see in the next section.

Figure 12

Deriving the Demand Curve

Panel (a) shows that when the price of Pepsi falls from $2 to $1, the consumer's optimum moves from point A to point B, and the quantity of Pepsi consumed rises from 250 to 750 liters. The demand curve in panel (b) reflects this relationship between the price and the quantity demanded.



QuickQuiz

5. Bart and Lisa are both optimizing consumers in the markets for shirts and hats, where they pay $100 for a shirt and $50 for a hat. Bart buys 8 shirts and 4 hats, while Lisa buys 6 shirts and 12 hats. From this information, we can infer that Bart's marginal rate of substitution is _____ hats per shirt, while Lisa's is _____ hats per shirt.
   a. 2; 1
   b. 2; 2
   c. 4; 1
   d. 4; 2

6. Maggie buys peanut butter and jelly, both of which are normal goods. When the price of peanut butter rises, the income effect induces Maggie to buy _____ peanut butter and _____ jelly.
   a. more; more
   b. more; less
   c. less; more
   d. less; less

7. Ned buys wine and bread. When the price of wine rises, the substitution effect induces Ned to buy _____ wine and _____ bread.
   a. more; more
   b. more; less
   c. less; more
   d. less; less

Answers are at the end of the chapter.

22-4 Three Applications

Now that we have developed the basic theory of consumer choice, let's use it to shed light on three questions about how the economy works.

22-4a Do All Demand Curves Slope Downward?

Normally, when the price of a good rises, people buy less of it. This typical behavior, called the **law of demand**, is reflected in the downward slope of the demand curve.

As a matter of economic theory, however, demand curves can sometimes slope upward. In other words, consumers can sometimes violate the law of demand and buy **more** of a good when the price rises. To see how this can happen, consider Figure 13. In this example, Conrad the consumer buys two goods: meat and potatoes. Initially, his budget constraint is the line from point A to point B, and the optimum is point C. When the price of potatoes rises, his budget constraint shifts inward and is now the line from point A to point D. His optimum moves to point E. An increase in the price of potatoes leads Conrad to buy more potatoes.

Why does Conrad respond in this strange way? In this example, meat is a normal good, but potatoes are a strongly inferior good. That is, potatoes are a good that Conrad buys a lot less of when his income rises and a lot more of when his income falls. In Figure 13, the increase in the price of potatoes makes Conrad poorer in the sense that he moves to a lower indifference curve. Because he is poorer, the income effect makes him want to buy less meat (the normal good) and more potatoes (the inferior good). At the same time, because potatoes have become more expensive relative to meat, the substitution effect makes Conrad want to buy more meat and fewer potatoes. Note that the income and substitution effects push in opposite directions. If the income effect is larger than the substitution effect, as it is in this example, Conrad responds to the higher price of potatoes by buying less meat and more potatoes.

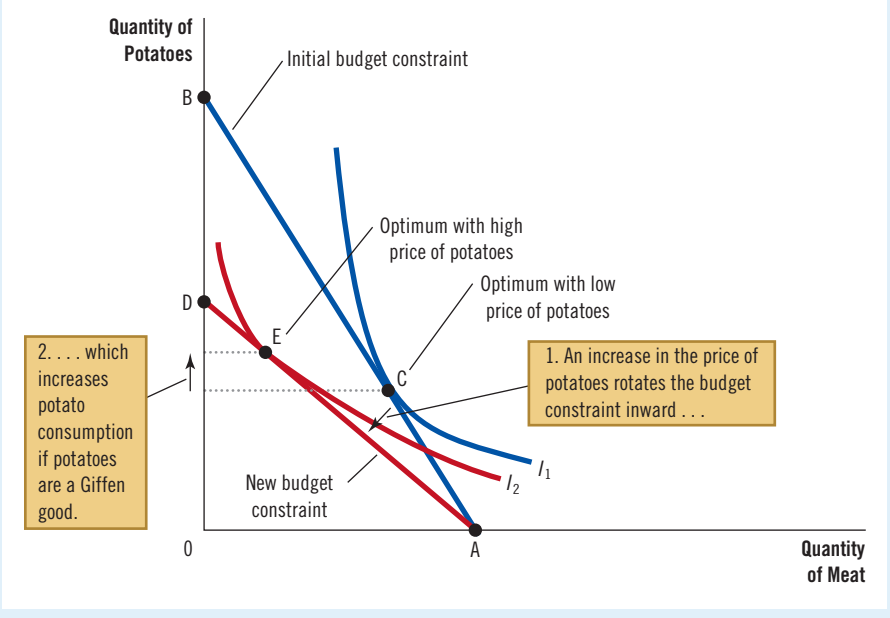
Economists use the term **Giffen good** to describe a good that violates the law of demand. The term is named for the economist Robert Giffen (1837–1910), who first noted this possibility. In this example, potatoes are a Giffen good. Giffen goods are inferior goods for which the income effect dominates the substitution effect. They therefore have demand curves that slope upward.

Giffen good
a good for which an increase in the price raises the quantity demanded

Figure 13

A Giffen Good

In this example, when the price of potatoes rises, the consumer's optimum shifts from point C to point E. In this case, the consumer responds to the higher price of potatoes by buying less meat and more potatoes.



Case Study

The Search for Giffen Goods

Have any actual Giffen goods ever been observed? Some historians suggest that potatoes were a Giffen good during the Irish potato famine of the 19th century. Potatoes were such a large part of people's diet that when the price of potatoes rose, the change had a large income effect. People responded to their reduced living standard by cutting back on the luxury of meat and buying more of the staple food of potatoes. Thus, it is sometimes argued that a higher price of potatoes actually raised the quantity of potatoes demanded.

A study by Robert Jensen and Nolan Miller, published in the *American Economic Review* in 2008, produced more concrete evidence for the existence of Giffen goods. These two economists conducted a field experiment for five months in the Chinese province of Hunan. They gave randomly selected households vouchers that subsidized the purchase of rice, a staple in local diets, and used surveys to measure how the consumption of rice responded to changes in the price. They found strong evidence that many poor households exhibited Giffen behavior. Lowering the price of rice with the subsidy voucher caused these households to reduce their rice consumption, and removing the subsidy had the opposite effect. Jensen and Miller wrote, "To the best of our knowledge, this is the first rigorous empirical evidence of Giffen behavior."

Thus, the theory of consumer choice allows demand curves to slope upward, and sometimes that strange phenomenon actually occurs. As a result, the law of demand introduced in Chapter 4 is not completely reliable. It is safe to say, however, that Giffen goods are rare. ●

22-4b How Do Wages Affect Labor Supply?

The theory of consumer choice can be used to analyze not only how people allocate their income but also how they allocate their time. Most people spend some of their time on leisurely pursuits and some of it working so they can afford to buy goods and services to consume. The essence of the time-allocation problem is the trade-off between leisure and consumption.

Consider Jasmine, a freelance software designer. She is awake for 100 hours per week. She spends some of this time enjoying leisure—playing *Minecraft*, watching *The Bachelor*, and reading this textbook. She spends the rest of this time developing software. For every hour she works coding, she earns $50, which she spends on rent, food, music downloads, and other consumption goods. Her hourly wage of $50 reflects the trade-off Jasmine faces between leisure and consumption. For every hour of leisure she gives up, she works one more hour and gets $50 of consumption.

Figure 14 shows Jasmine's budget constraint. If she spends all 100 hours enjoying leisure, she has no consumption. If she spends all 100 hours working, she has a weekly consumption of $5,000 but no time for leisure. If she works a 40-hour week, she enjoys 60 hours of leisure and has a weekly consumption of $2,000.

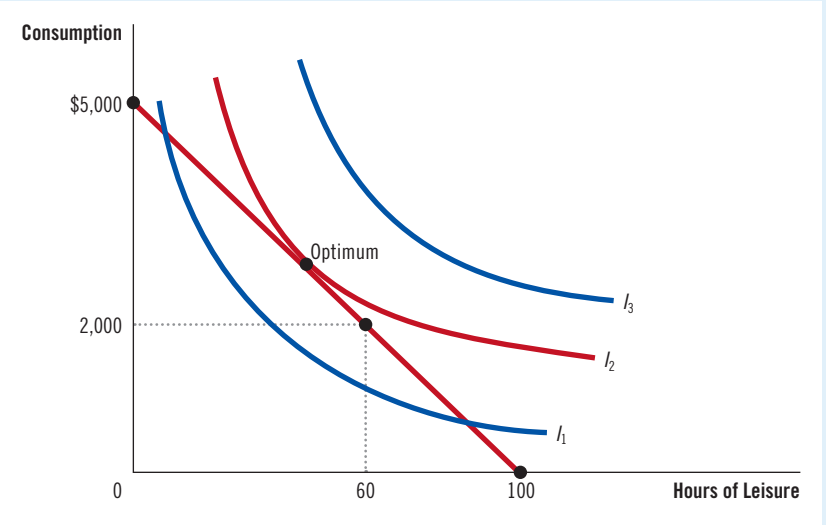
Figure 14 uses indifference curves to represent Jasmine's preferences for consumption and leisure. Here, consumption and leisure are the two "goods" between which Jasmine chooses. Because Jasmine always prefers more leisure and more consumption, she prefers points on higher indifference curves to points on lower ones. At a wage of $50 per hour, Jasmine chooses a combination of consumption and leisure represented by the point labeled "optimum." The optimum is the point on the budget constraint at which Jasmine reaches the highest possible indifference curve, I_2.

Now consider what happens when Jasmine's wage increases from $50 to $60 per hour. Figure 15 illustrates two possible outcomes. In both cases, the budget

Figure 14

The Work-Leisure Decision

This figure shows Jasmine's budget constraint for deciding how much to work, her indifference curves for consumption and leisure, and her optimum.



constraint, shown in the left graphs, shifts outward from BC_1 to BC_2. In the process, each budget constraint becomes steeper, reflecting the change in relative price: At the higher wage, Jasmine earns more consumption for every hour of leisure that she gives up.

Jasmine's preferences, as represented by her indifference curves, determine how her choice regarding consumption and leisure responds to the higher wage. In both panels, consumption rises. Yet the responses of leisure to the wage change are different in the two cases. In panel (a), Jasmine responds to the higher wage by enjoying less leisure. In panel (b), Jasmine responds by enjoying more leisure.

Jasmine's decision between leisure and consumption determines her supply of labor because the more leisure she enjoys, the less time she has left to work. In each panel of Figure 15, the right graph shows the labor-supply curve implied by Jasmine's decision. In panel (a), a higher wage induces Jasmine to enjoy less leisure and work more, so the labor-supply curve slopes upward. In panel (b), a higher wage induces Jasmine to enjoy more leisure and work less, so the labor-supply curve slopes "backward."

At first, the backward-sloping labor-supply curve is puzzling. Why would a person respond to a higher wage by working less? The answer comes from considering the income and substitution effects of a higher wage.

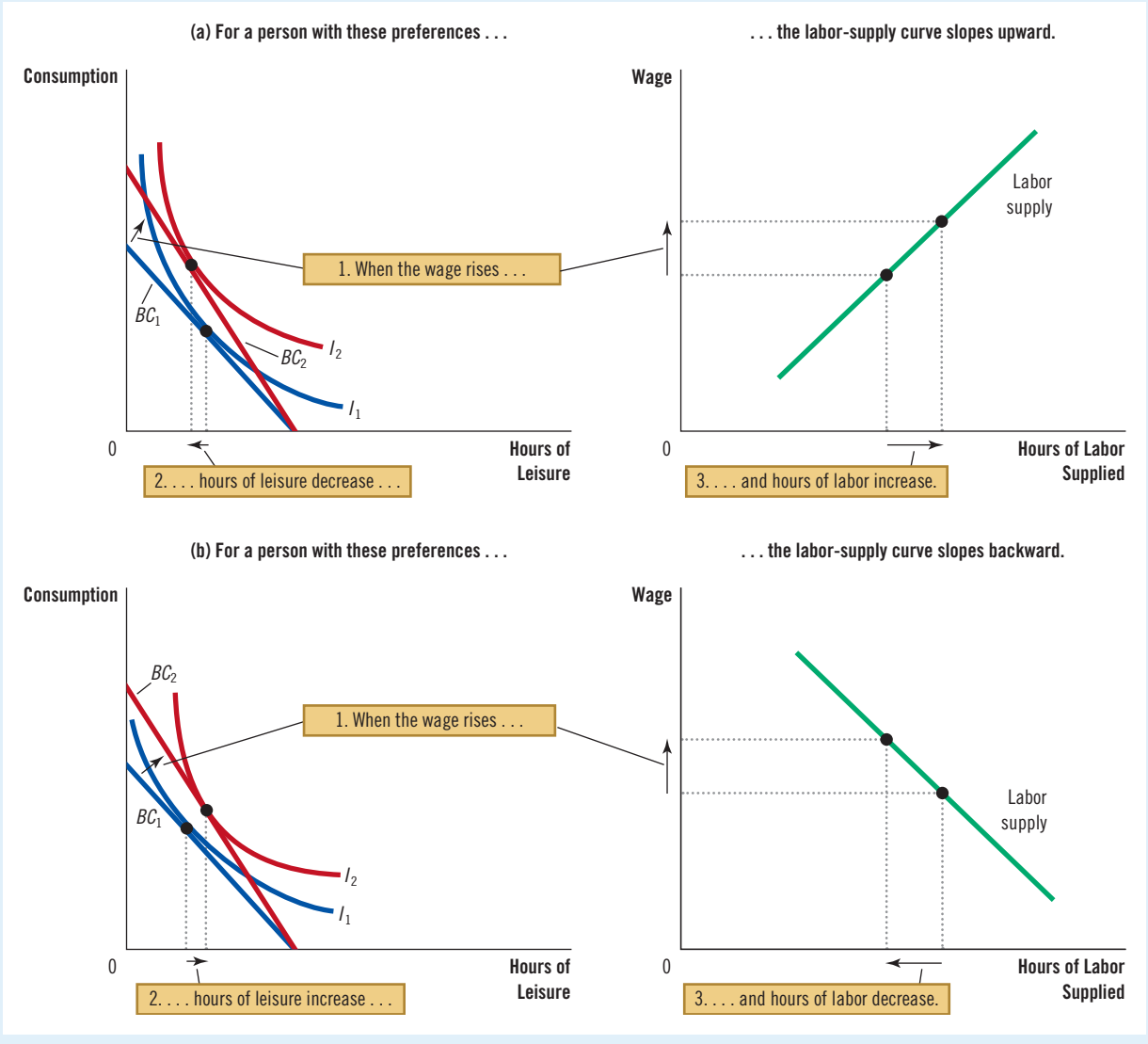
Consider first the substitution effect. When Jasmine's wage rises, leisure becomes more expensive relative to consumption, encouraging Jasmine to substitute away from leisure and toward consumption. In other words, the substitution effect induces Jasmine to work more in response to higher wages and tends to make the labor-supply curve slope upward.

Now consider the income effect. When Jasmine's wage rises, she moves to a higher indifference curve, so she is better off than before. As long as consumption and leisure are both normal goods, Jasmine will want to use her increased well-being to enjoy both higher consumption and more leisure. In other words, the income effect induces her to work less and thus tends to make the labor-supply curve slope backward.

Figure 15

An Increase in the Wage

The two panels of this figure show how a person might respond to an increase in the wage. The graphs on the left show the consumer's initial budget constraint, BC_1, and new budget constraint, BC_2, as well as the consumer's optimal choices over consumption and leisure. The graphs on the right show the resulting labor-supply curve. Because hours worked equal the total hours available minus hours of leisure, any change in leisure implies an opposite change in the quantity of labor supplied. In panel (a), when the wage rises, consumption rises, and leisure falls, resulting in a labor-supply curve that slopes upward. In panel (b), when the wage rises, both consumption and leisure rise, resulting in a labor-supply curve that slopes backward.

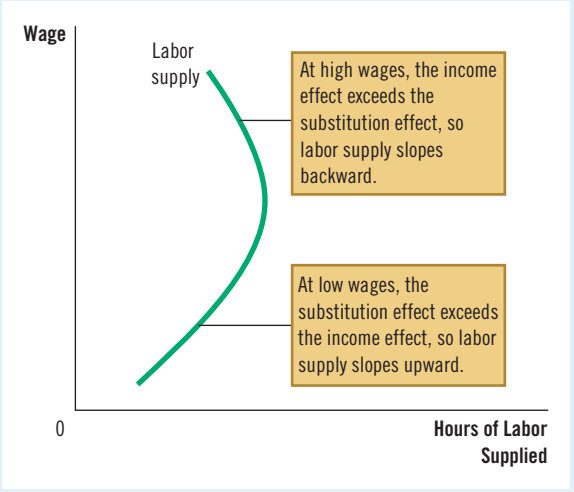


In the end, economic theory does not give a clear prediction about whether an increase in the wage induces Jasmine to work more or less. If the substitution effect exceeds the income effect, she works more. If the income effect exceeds the substitution effect, she works less. The labor-supply curve, therefore, could be either upward- or backward-sloping.

Figure 16

A Backward-Bending Labor-Supply Curve

Here, the labor-supply curve slopes upward at low wages because the substitution effect dominates the income effect. But as the wage rises, the income effect starts to dominate the substitution effect, and the labor-supply curve bends backward.



Moreover, the slope of the labor-supply curve need not be the same at all wages. For example, depending on a person's preferences, it is possible for the substitution effect to dominate the income effect at low wages and the income effect to dominate the substitution effect at high wages. In this case, as in Figure 16, the labor-supply curve starts off upward-sloping but then bends backward as the wage increases.

Case Study

Income Effects on Labor Supply: Historical Trends, Lottery Winners, and the Carnegie Conjecture

The idea of a backward-sloping labor-supply curve is not a mere theoretical curiosity. Evidence indicates that the labor-supply curve, considered over long periods, does indeed slope backward. A hundred years ago, many people worked six days a week. Today, five-day workweeks are the norm. While the length of the workweek has been falling, the wage of the typical worker (adjusted for inflation) has been rising.

Here is how economists explain this historical pattern: Over time, advances in technology increase workers' productivity and the demand for labor. This increase in labor demand raises equilibrium wages. As wages rise, so does the reward for working. Yet rather than responding to this increased incentive by working more, most workers take advantage of their greater prosperity by increasing leisure. In other words, the income effect of higher wages dominates the substitution effect.

Further evidence for a strong income effect on labor supply comes from a very different kind of data: winners of lotteries. Winners of large lottery prizes see large increases in their incomes and, as a result, large outward shifts in their budget constraints. Because the winners' wages have not changed, however, the **slopes** of their budget constraints remain the same. There is, therefore, no substitution effect. By comparing lottery winners with those who played the lottery but lost, researchers can isolate the income effect on labor supply.

A 2021 study by four University of Chicago economists did exactly that and found some striking results. For every dollar that people win, their total after-tax labor earnings decline by 40 cents. In other words, 40 percent of

DAVID EDSAM/ALAMY STOCK PHOTO

"No more 9 to 5 for me."

winnings are used to increase leisure (or perhaps to switch to a more pleasant, lower-paying job), while 60 percent are spent to increase consumption of goods and services. For example, for those near the normal retirement age, lottery winners are much more likely than lottery losers to retire early. The researchers also report that households with higher income before winning the lottery exhibited larger effects on leisure and smaller effects on consumption. The income effect on labor supply appears substantial, especially among the most fortunate.

These findings would not have surprised the 19th-century industrialist Andrew Carnegie. In his book *The Gospel of Wealth*, he warned that "the parent who leaves his son enormous wealth generally deadens the talents and energies of the son and tempts him to lead a less useful and less worthy life than he otherwise would." That is, Carnegie viewed the income effect on labor supply to be large and, from his paternalistic perspective, regrettable. That may be partly why, during his life and at his death, Carnegie gave much of his vast fortune to charity. ●

22-4c How Do Interest Rates Affect Household Saving?

An important decision that every person faces is how much income to consume today and how much to save for the future. We can use the theory of consumer choice to analyze this decision. In doing so, we examine how the amount that people save depends on the interest rate their savings will earn.

Consider the decision facing Ryder, a worker planning for retirement. To keep things simple, divide Ryder's life into two periods. In the first, Ryder is young and working. In the second, he is old and retired. When young, Ryder earns $100,000. He divides this income between current consumption and saving. When he is old, Ryder consumes what he has saved, including the interest that his savings have earned.

We can view "consumption when young" and "consumption when old" as the two goods that Ryder must choose between. The interest rate determines the relative price of these two goods. Suppose the interest rate is 10 percent. Then for every dollar that Ryder saves when young, he can consume $1.10 when old.

Figure 17 shows Ryder's budget constraint. If he saves nothing, he consumes $100,000 when young and nothing when old. If he saves everything, he consumes nothing when young and $110,000 when old. The budget constraint shows these and all the intermediate possibilities.

Figure 17 uses indifference curves to represent Ryder's preferences for consumption in the two periods. Because he prefers more consumption in both periods of his life, he prefers points on higher indifference curves to points on lower ones. Given his preferences, Ryder chooses the optimal combination of consumption in the two periods, which is the point on the budget constraint that is on the highest possible indifference curve. At this optimum, Ryder consumes $50,000 when young and $55,000 when old.

Now consider what happens when the interest rate increases from 10 percent to 20 percent. Figure 18 shows two possible outcomes. In both cases, the budget constraint shifts outward and becomes steeper. At the new, higher interest rate, Ryder gets more consumption when old for every dollar of consumption that he gives up when young.

Figure 17

The Consumption-Saving Decision

This figure shows the budget constraint for a person deciding how much to consume in the two periods of his life, the indifference curves representing his preferences, and the optimum.

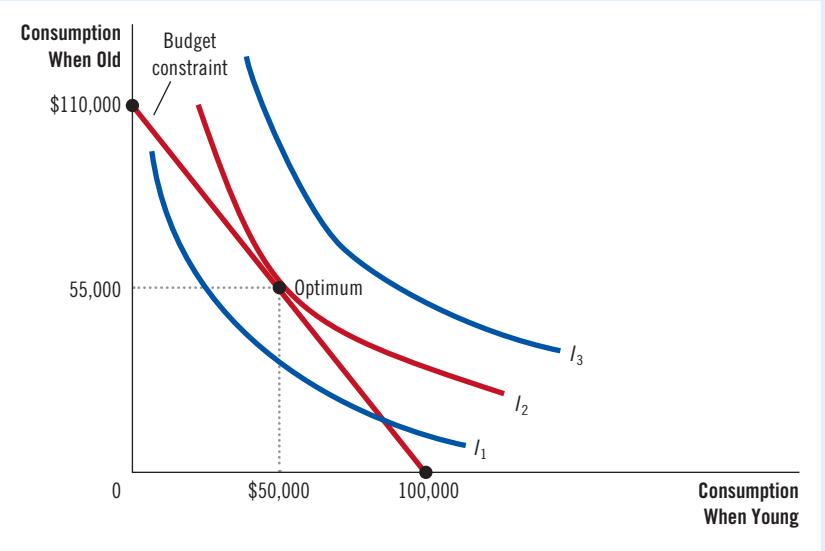
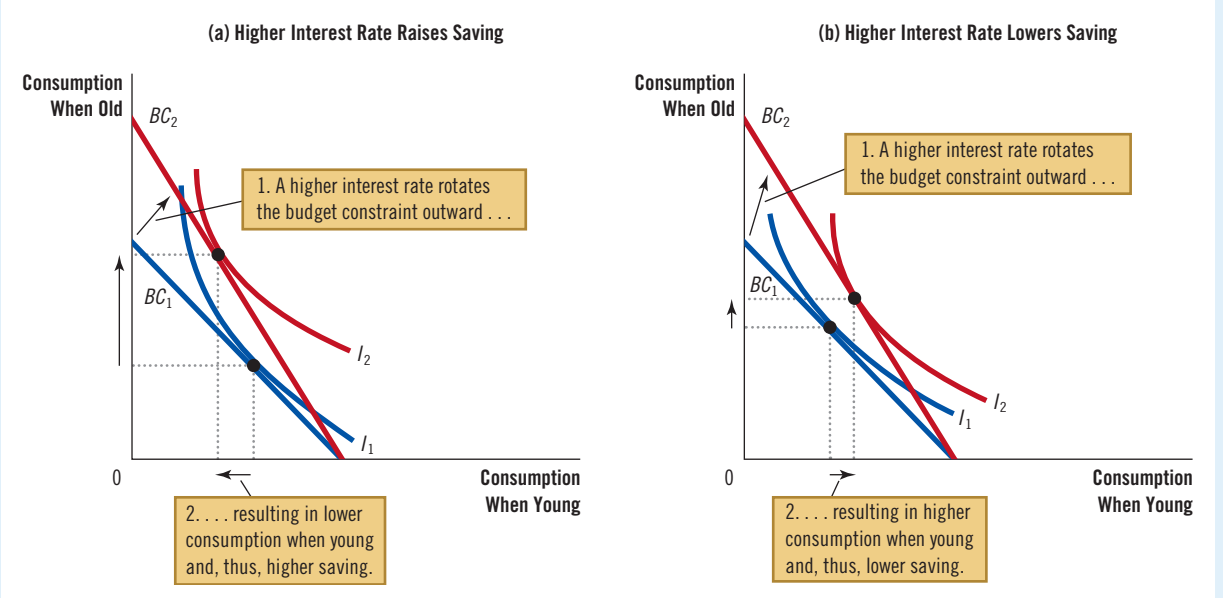


Figure 18

An Increase in the Interest Rate

In both panels, an increase in the interest rate shifts the budget constraint outward. In panel (a), consumption when young falls, and consumption when old rises. The result is an increase in saving when young. In panel (b), consumption in both periods rises. The result is a decrease in saving when young.



The two panels show the results given different preferences by Ryder. In both cases, consumption when old rises. Yet consumption when young responds differently in the two cases. In panel (a), Ryder responds to the higher interest rate by consuming less when young. In panel (b), Ryder responds by consuming more when young.

Ryder's saving is his income minus his consumption when young. In panel (a), an increase in the interest rate reduces consumption when young, so saving rises. In panel (b), an increase in the interest rate increases consumption when young, so saving falls.

The case shown in panel (b) might seem odd: Ryder responds to an increase in the return to saving by saving less. But this behavior is not as peculiar as it might seem. We can understand it by considering the income and substitution effects of a higher interest rate.

Consider first the substitution effect. When the interest rate rises, consumption when old becomes less costly relative to consumption when young. Therefore, the substitution effect induces Ryder to consume more when old and less when young. In other words, the substitution effect induces Ryder to save more.

Now consider the income effect. When the interest rate rises, Ryder moves to a higher indifference curve, so he is better off. If consumption when young and consumption when old are both normal goods, he will want to use his increased well-being to consume more in both periods. In other words, the income effect induces him to save less.

The result depends on both the income and substitution effects. If the substitution effect of a higher interest rate exceeds the income effect, Ryder saves more. If the income effect exceeds the substitution effect, Ryder saves less. The theory of consumer choice says that an increase in the interest rate could either encourage or discourage saving.

This ambiguous result is interesting from the standpoint of economic theory, but it is disappointing from the standpoint of economic policy. An important issue in tax policy hinges, in part, on how saving responds to interest rates. Some economists advocate reducing the taxation of interest and other capital income, arguing that such a policy change would raise the after-tax interest rate that savers can earn and encourage people to save more. Others argue that because of offsetting income and substitution effects, such a tax change might not increase saving and could even reduce it. Unfortunately, research has not led to a consensus. There remains disagreement about whether changes in tax policy aimed at increasing saving would, in fact, have the intended effect.

QuickQuiz

8. Mr. Burns buys only lobster and chicken. Lobster is a normal good, while chicken is an inferior good. When the price of lobster rises, Mr. Burns buys
   a. less of both goods.
   b. more lobster and less chicken.
   c. less lobster and more chicken.
   d. less lobster, but the impact on chicken is ambiguous.

9. If Edna buys more pasta when the price of pasta increases, we can infer that for Edna,
   a. pasta is a normal good for which the income effect exceeds the substitution effect.
   b. pasta is a normal good for which the substitution effect exceeds the income effect.
   c. pasta is an inferior good for which the income effect exceeds the substitution effect.
   d. pasta is an inferior good for which the substitution effect exceeds the income effect.

10. Maude's labor-supply curve slopes upward if, for Maude,
    a. leisure is a normal good.
    b. consumption is a normal good.
    c. the income effect on leisure exceeds the substitution effect.
    d. the substitution effect on leisure exceeds the income effect.

11. Consumption when young and consumption when old are both normal goods for Seymour, a worker saving for retirement. When the interest rate falls, what happens to Seymour's consumption when old?
    a. It definitely increases.
    b. It definitely decreases.
    c. It increases only if the substitution effect exceeds the income effect.
    d. It decreases only if the substitution effect exceeds the income effect.

Answers are at the end of the chapter.

22-5 Conclusion: Do People Really Think This Way?

The theory of consumer choice describes how people make decisions. As we have seen, it applies to many questions. It can explain how people choose between pizza and Pepsi, work and leisure, consumption and saving, and so on.

Now that we understand the theory, let's step back and consider whether it is credible. You might be tempted to look upon the theory of consumer choice with some skepticism. After all, you are a consumer. You decide what to buy every time you walk into a store. And you know that you do not make your purchasing decisions by writing down budget constraints and mapping out indifference curves. Doesn't this knowledge about your own decision making refute the theory?

No, it doesn't. The theory of consumer choice does not try to present a literal account of how people make decisions. It is a model. As Chapter 2 discussed, models are not supposed to be completely realistic.

The theory of consumer choice should be viewed as a metaphor for how consumers make decisions. No consumer (except an occasional economist) goes through the explicit optimization envisioned in the theory. Yet consumers know that their choices are constrained by their financial resources. And given those constraints, they do the best they can to achieve the highest level of satisfaction. The theory of consumer choice describes this intuitive process in a way that permits formal economic analysis.

Just as the proof of the pudding is in the eating, the test of a theory is in its applications. The last section of this chapter applied the theory of consumer choice to three practical issues. If you take more advanced courses in economics, you will see that this theory provides the framework for much additional analysis.

Chapter in a Nutshell

- A consumer's budget constraint shows the possible bundles of different goods she can buy given her income and the prices of the goods. The slope of the budget constraint equals the relative price of the goods.
- The consumer's indifference curves represent her preferences. An indifference curve shows the various bundles of goods that make the consumer equally happy. Points on higher indifference curves are preferred to points on lower indifference curves. The slope of an indifference curve at any point is the consumer's marginal rate of substitution—the rate at which the consumer is willing to trade one good for the other.
- The consumer optimizes by choosing the point on her budget constraint that lies on the highest indifference curve. At this point, the slope of the indifference curve (the marginal rate of substitution between the goods) equals the slope of the budget constraint (the relative price of the goods), and the consumer's valuation of the two goods (measured by the marginal rate of substitution) equals the market's valuation (measured by the relative price).
- When the price of a good falls, the impact on the consumer's choices can be broken down into an income effect and a substitution effect. The income effect is the change in consumption that arises because a lower price makes the consumer better off. The substitution effect is the change in consumption that arises because a price change encourages greater consumption of the good that has become relatively cheaper. The income effect is reflected in the movement from a lower to a higher indifference curve, while the substitution effect is reflected by a movement along an indifference curve to a point with a different slope.
- The theory of consumer choice can be applied in many situations. It explains why demand curves sometimes slope upward, why higher wages can either increase or decrease the quantity of labor supplied, and why higher interest rates can either increase or decrease saving.

Key Concepts

budget constraint, p. 444
indifference curve, p. 447
marginal rate of substitution, p. 447
perfect substitutes, p. 450
perfect complements, p. 451
normal good, p. 454
inferior good, p. 454
income effect, p. 455
substitution effect, p. 455
Giffen good, p. 459

Questions for Review

1. Hari has income of $6,000. Wine costs $6 per glass, and cheese costs $12 per pound. Draw Hari's budget constraint with wine on the vertical axis. What is the slope of this budget constraint?
2. Hari has typical indifference curves for wine and cheese. Draw them and explain their four properties.
3. Pick a point on an indifference curve for wine and cheese, and show the marginal rate of substitution. What does the marginal rate of substitution tell us?
4. Show Hari's budget constraint and indifference curves for wine and cheese, along with his optimal consumption choice. If the price of wine is $6 per glass and the price of cheese is $12 per pound, what is the marginal rate of substitution at this optimum?
5. Hari gets a raise, so his income increases from $6,000 to $8,000. Show what happens if both wine and cheese are normal goods. Next, show what happens if cheese is an inferior good.
6. The price of cheese rises from $12 to $20 per pound, while the price of wine remains $6 per glass. Assuming Hari has a constant income of $6,000, show what happens to the consumption of wine and cheese. Decompose the change into income and substitution effects.
7. Can an increase in the price of cheese possibly induce Hari to buy more cheese? Explain.

Problems and Applications

1. Maya divides her income between coffee and croissants (both of which are normal goods). An early frost in Brazil causes a large increase in the price of coffee in the United States.
   a. Show the effect of the frost on Maya's budget constraint.
   b. Show the effect of the frost on Maya's optimal consumption bundle, assuming that the substitution effect outweighs the income effect for croissants.
   c. Show the effect of the frost on Maya's optimal consumption bundle, assuming that the income effect outweighs the substitution effect for croissants.
2. Compare the following two pairs of goods:
   - Coke and Pepsi
   - Skis and ski bindings
     a. In which case are the two goods complements? In which case are they substitutes?
     b. In which case do you expect the indifference curves to be fairly straight? In which case do you expect the indifference curves to be very bowed?
     c. In which case will the consumer respond more to a change in the relative price of the two goods?
3. You consume only soda and pizza. One day, the price of soda goes up, the price of pizza goes down, and you are just as happy as you were before the price changes.
   a. Illustrate this situation on a graph.
   b. How does your consumption of the two goods change? How does your response depend on income and substitution effects?
   c. Can you afford the bundle of soda and pizza you consumed before the price changes?
4. Raj consumes only cheese and crackers.
   a. Could cheese and crackers both be inferior goods for Raj? Explain.
   b. Suppose that cheese is a normal good for Raj and crackers are an inferior good. If the price of cheese falls, what happens to Raj's consumption

of crackers? What happens to his consumption of cheese? Explain.

5. Darius buys only milk and cookies.
   a. In year 1, Darius earns $100, milk costs $2 per quart, and cookies cost $4 per dozen. Draw Darius's budget constraint.
   b. Now suppose that all prices increase by 10 percent in year 2 and that Darius's salary increases by 10 percent as well. Draw Darius's new budget constraint. How would Darius's optimal bundle of milk and cookies in year 2 compare with his optimal bundle in year 1?
6. State whether each of the following statements is true or false. Explain your answers.
   a. "All Giffen goods are inferior goods."
   b. "All inferior goods are Giffen goods."
7. Priya, a college student, has two options for meals: eating at the dining hall for $6 per meal, or eating a cup of Ramen soup for $1.50 per meal. Her weekly food budget is $60.
   a. Draw the budget constraint showing the trade-off between dining-hall meals and cups of soup. Assuming that Priya spends equal amounts on both goods, draw an indifference curve showing the optimum choice. Label the optimum as point A.
   b. Suppose the price of a cup of soup rises to $2. Using your diagram from part (a), show the consequences of this change in price. Assume that Priya now spends only 30 percent of her income on dining-hall meals. Label the new optimum as point B.
   c. What happened to the quantity of soup consumed because of this price change? What does this result say about the income and substitution effects? Explain.
   d. Use points A and B to draw Priya's demand curve for cups of soup. What is this type of good called?
8. Consider your decision about how many hours to work.
   a. Draw your budget constraint, assuming that you pay no taxes on your income. On the same diagram, draw another budget constraint, assuming that you pay a 15 percent income tax.
   b. Show how the tax might lead you to work more hours, fewer hours, or the same number of hours. Explain.
9. Anya is awake for 100 hours per week. Using one diagram, show Anya's budget constraints if she earns $12 per hour, $16 per hour, and $20 per hour. Now draw indifference curves such that Anya's labor-supply curve is upward-sloping when the wage is between $12 and $16 per hour and backward-sloping when the wage is between $16 and $20 per hour.
10. Draw the indifference curve for someone deciding how to allocate time between work and leisure. Suppose the wage increases. Is it possible that the person's consumption would fall? Is this plausible? Discuss. (Hint: Think about income and substitution effects.)
11. The economist George Stigler once wrote that, according to consumer theory, "if consumers do not buy less of a commodity when their incomes rise, they will surely buy less when the price of the commodity rises." Explain this statement using the concepts of income and substitution effects.
12. Five consumers have the following marginal utility of apples and pears:

	Marginal Utility of Apples	Marginal Utility of Pears
Claire	6	12
Phil	6	6
Haley	6	3
Alex	3	6
Luke	3	12

The price of an apple is $1, and the price of a pear is $2. Which, if any, of these consumers are optimizing their choices of fruit? For those who are not, how should they change their spending?

QuickQuiz Answers

1. **a** 2. **d** 3. **d** 4. **b** 5. **b** 6. **d** 7. **c** 8. **c** 9. **c** 10. **d** 11. **b**

Chapter

23

Frontiers of Microeconomics

Economics is a study of the choices that people make and the interactions among people as they go about their lives. As the preceding chapters demonstrate, the field has many facets. Yet the facets we have already seen do not make up a finished jewel, perfect and unchanging. Economists are always looking for new areas to study, new phenomena to explain, and new ways to see the world. This final chapter on microeconomics discusses three topics at the discipline's frontier to show how economists are trying to expand their understanding of human behavior and society.

The first topic is the economics of **asymmetric information**. In many situations, some people are better informed than others, and the imbalance in knowledge affects the choices they make and how they deal with one another. Thinking about this asymmetry sheds light on many aspects of the world, from the market for used cars to the custom of gift giving.

The second topic in this chapter is **political economy**, a term that once encompassed the entire field of economics but now refers to interdisciplinary work at the boundary of political science and

BLACKCAT/E+/GETTY IMAGES; GOWITHSTOCK/SHUTTERSTOCK.COM

economics. Throughout this book, we have seen many examples in which markets fail to deliver desirable outcomes, and government policy can potentially improve matters. But "potentially" is a necessary qualifier: Whether this potential is realized depends on political institutions. The field of political economy uses the tools of economics to study how government works and, in doing so, provides a deeper understanding of economic policy.

The third topic in this chapter is **behavioral economics**. This field brings insights from psychology into the study of economic issues. It questions whether people are entirely rational, at least in the conventional sense, and it offers a view of human behavior that is more subtle and complex, and perhaps more realistic, than the one found in standard economic theory.

This chapter covers a lot of ground. To do so, it does not offer full helpings of these three topics but, instead, gives a taste of each. One goal of this chapter is to show where economists are heading in their efforts to expand knowledge. Another is to whet your appetite for more courses in economics and allied fields.

23-1 Asymmetric Information

"I know something you don't know." This common elementary-school taunt conveys a truth about how people sometimes interact with one another. Often in life, one person knows more about what is going on than another. When that difference in knowledge is relevant to an interaction, it is called **information asymmetry**.

Examples abound. Workers know more than their employers about how much effort they put into their jobs. Sellers of used cars know more than buyers about whether the cars offered for sale are in good condition. The first is an example of a **hidden action**, while the second is an example of a **hidden characteristic**. In each case, the uninformed party (the employer, the car buyer) would like to know the relevant information, but the informed party (the worker, the car seller) may have an incentive to conceal it.

Because asymmetric information is so common, economists have devoted much effort to studying its effects. Here are some insights this study has revealed.

23-1a Hidden Actions: Principals, Agents, and Moral Hazard

moral hazard
the tendency of a person who is imperfectly monitored to engage in dishonest or otherwise undesirable behavior

agent
a person who performs an act for another person, called the principal

principal
a person for whom another person, called the agent, performs some act

Moral hazard is a problem that arises when one person, called the **agent**, performs a task on behalf of another person, called the **principal**. If the principal cannot perfectly monitor the agent's behavior, the agent may act in a way that serves the agent's self-interest but is undesirable for the principal. Economists adopted the somewhat odd phrase **moral hazard** from the insurance industry: It refers to the risk, or "hazard," of inappropriate or otherwise "immoral" behavior by the agent. In such a situation, the principal tries various ways to encourage the agent to act more responsibly.

The employment relationship is a classic example. The employer is the principal, and the worker is the agent. The moral-hazard problem is the temptation of imperfectly monitored workers to shirk their responsibilities. Employers can respond to this problem in various ways:

- **Better monitoring.** Employers may plant hidden video cameras to record workers' behavior. The employers' aim is to catch irresponsible actions that might occur when supervisors are absent.
- **High wages.** According to **efficiency-wage theories** (discussed in Chapter 20), some employers may choose to pay their workers a wage above the

level that balances supply and demand in the labor market. Workers who earn above-equilibrium wages—and are treated well in other ways—are less likely to violate their employers' trust.

- **Delayed payment.** Firms can delay part of workers' compensation, so a worker who is caught shirking and fired suffers a larger penalty. One example of delayed compensation is the year-end bonus. Similarly, a firm may pay its workers more later in their lives. The wage increases that workers usually get as they age may reflect not just the benefits of experience but also a response to moral hazard.

Employers can use any combination of these mechanisms to reduce moral hazard.

There are many examples of moral hazard beyond the workplace. A homeowner with fire insurance might buy too few fire extinguishers because the homeowner bears the cost of the extinguisher while the insurance company receives much of the benefit. A family may live near a river with a high risk of flooding because the

Corporate Management

Corporations dominate the modern economy. In some ways, corporations are like other firms: They buy inputs in markets for the factors of production, sell their output in markets for goods and services, and typically aim to maximize profit for their owners. But large corporations raise some issues that do not arise in, say, small family-owned businesses.

From a legal standpoint, a corporation is an organization that is granted a charter recognizing it as a separate legal entity with its own rights and responsibilities distinct from those of its owners and employees. An important protection that corporations have is limited liability: The owners of the corporation can lose their entire investment in the business, but not more than that. They are not personally liable for any losses the corporation suffers.

Limited liability is related to another important feature of corporations: the separation of ownership and control. One group of people, the shareholders, own the corporation and share in its profits. Another group of people, the managers, are employed by the corporation to make decisions about how to deploy its resources. This arrangement requires limited liability because shareholders would be reluctant to invest in a business and let managers control it if the shareholders could incur unbounded losses.

The separation of ownership and control creates a principal-agent problem, with the shareholders as the principals and the managers as the agents. The chief executive officer and other managers are charged with the task of running the company for the shareholders. That makes sense because the managers are in the best position to know the available business opportunities. But the situation is fraught with moral hazard. While the shareholders typically want the managers to maximize profits, the managers may have goals of their own, such as taking life easy, having a plush office and a private jet, throwing lavish parties, or presiding over a large business empire.

The corporation's board of directors is responsible for hiring and firing the top management. The board monitors the managers' performance and sets their compensation packages. These packages often include incentives aimed at aligning the interests of shareholders with the interests of management. Managers might be given performance-based bonuses or stock options, which increase in value if the company performs well.

Note that the directors are themselves agents of the shareholders. The existence of a board overseeing management only shifts the principal-agent problem. The issue then becomes how to ensure that the board of directors fulfills its own legal obligation of acting in the best interest of the shareholders. If the directors become too friendly with management, they may not provide the required oversight.

The principal-agent problem inherent in corporations flares up periodically. It became big news around 2005. The top managers of several prominent companies, including Enron, Tyco, and WorldCom, were found to be engaging in activities that enriched themselves at the expense of their shareholders. In these cases, the actions were so extreme that they were criminal, and the corporate managers were not just fired but also sent to prison. Some shareholders sued the directors for failing to monitor management sufficiently.

Criminal activity by corporate managers is rare, but in some ways, it is only the tip of the iceberg. Whenever ownership and control are separated, as they are in most large corporations, there is an inevitable tension between the interests of shareholders and the interests of management, and the problem of moral hazard can impede businesses from operating as efficiently as they could. ■

family enjoys the scenic views, while the government bears the cost of disaster relief after a flood. Regulations aim to address the problem: An insurance company may require homeowners to buy fire extinguishers, and the government may prohibit building homes on land with a high risk of flooding. But the insurance company does not have perfect information about how cautious homeowners are, and the government does not have perfect information about the risks that families undertake when choosing where to live, so the problem of moral hazard persists.

23-1b Hidden Characteristics: Adverse Selection and the Lemons Problem

adverse selection
the tendency for the mix of unobserved attributes to become undesirable from the standpoint of an uninformed party

Adverse selection is a problem that arises when one party in a transaction knows more about the attributes of the item being exchanged than the other. For example, the seller of a good might know more about it than the buyer, so the buyer runs the risk of being sold something of low quality. The "selection" of goods sold may be "adverse" from the buyer's standpoint.

The classic example of adverse selection is the market for used cars. Sellers know their vehicles' defects while buyers often do not. Because the owners of the worst cars are more likely to sell them than the owners of the best cars, buyers worry about getting a "lemon," and many people avoid buying used cars entirely. This lemons problem can explain why a used car only a few weeks old sells for thousands of dollars less than a new car of the same type. A buyer of the used car might surmise that the seller is getting rid of the car quickly because the seller knows something the buyer does not.

The labor market provides a second example of adverse selection. According to one efficiency-wage theory, workers know their own abilities better than the firms that hire them do. If a firm reduces the wage it pays, the more talented workers are more likely to quit, knowing they will be able to find other jobs. Conversely, a firm may choose to pay an above-equilibrium wage to attract better workers.

A third example of adverse selection occurs in insurance markets, as Chapter 12 discussed. Buyers of health insurance know more about their own problems than insurance companies do. If people with greater hidden health problems are more likely to buy health insurance than other people, the price of health insurance will reflect the costs of a sicker-than-average person. That may discourage healthier people from buying insurance, which only makes health insurance even more expensive.

When markets suffer from adverse selection, the invisible hand does not necessarily work its magic by ensuring that the gains from trade are fully realized. In the used car market, owners of good cars may keep them rather than sell them at the low price that skeptical buyers are willing to pay. In the labor market, wages may be stuck above the level that balances supply and demand, resulting in unemployment. In insurance markets, some potential buyers may remain uninsured because the policies they are offered fail to reflect their true characteristics.

23-1c Signaling to Convey Private Information

Markets respond to problems of asymmetric information in many ways. One is **signaling**, which refers to actions taken by an informed party for the purpose of revealing private information.

signaling
an action taken by an informed party to reveal private information to an uninformed party

Signaling has turned up in previous chapters. Chapter 17 discussed that firms may spend money on advertising to signal to potential customers that they have high-quality products. Chapter 20 considered that students may earn college degrees to signal to potential employers that they are high-ability individuals, rather than

to increase their productivity. These two examples of signaling (advertising, education) may seem very different, but below the surface, they are much the same: In both cases, the informed party (the firm, the student) uses the signal to convince the uninformed party (the customer, the employer) that the informed party is offering something of high quality.

What does it take for an action to be an effective signal? First, it must be costly. If a signal were free, everyone would use it, and it would convey no information. For the same reason, there is another requirement: The signal must be less costly, or more beneficial, to the person with the higher-quality product. Otherwise, everyone would have the same incentive to use the signal, and the signal would reveal nothing.

Consider again our two examples. In the advertising case, a firm with a good product reaps a larger benefit from advertising because customers who try the product once are more likely to become repeat customers. Thus, it is rational for the firm with a good product to pay for the cost of the signal (advertising), and it is rational for the customer to view the signal as informative about the product's quality. In the education case, a talented person can finish school more easily than a less talented one. Thus, it is rational for the talented person to pay for the cost of the signal (education), and it is rational for the employer to view the signal as informative about the person's talent.

The world is replete with instances of signaling. Magazine ads sometimes include the phrase "as seen on TV." Why stress this fact? One possibility is that the firm is trying to convey its willingness to pay for an expensive signal (an ad on television) in the hope that you will infer that its product is of high quality. For the same reason, graduates of elite schools are always sure to put that fact on their résumés.

Case Study

Gifts as Signals

A man is debating what to give his girlfriend for her birthday. "I know," he says to himself, "I'll give her cash. After all, I don't know her preferences as well as she does, and with cash, she can buy anything she wants." But when he hands her the money, she is offended. Convinced he doesn't really love her, she breaks off the relationship.

What's the economics behind this story?

In some ways, gift giving is a strange custom. As the man in our story suggests, people typically know their own tastes better than others do, so we might expect everyone to prefer cash to in-kind transfers. If your employer chose some merchandise and substituted it for your paycheck, you would likely object to this means of payment. But your reaction is very different when someone who (you hope) loves you does something similar.

One interpretation of gift giving is that it reflects asymmetric information and signaling. The man in our story has private information that the girlfriend would like to know: Does he really love her? He says he does, but should she believe him? Talk is cheap.

Choosing a good gift for her can be a signal of his love. Certainly, the act of picking out a good gift, rather than giving cash, has the right characteristics to be a signal. It is costly (it takes time), and its cost depends on private information (how much he loves her). If he really loves her, choosing a good gift is easy because he is

ADRIATICFOTO/SHUTTERSTOCK.COM

"Now we'll see how much he loves me."

thinking about her all the time. If he doesn't love her, finding the right gift is more difficult. Thus, giving a gift that suits his girlfriend is one way for him to convey the private information of his love for her. Giving cash shows that he isn't even bothering to try.

The signaling theory of gift giving is consistent with another observation: People care most about the custom when the strength of affection is most in question. Giving cash to a girlfriend or boyfriend is usually a bad move. But when college students receive a check from their parents, they are less often offended. The parents' love is less likely to be in doubt, so the recipient probably won't interpret the cash gift as a signal of insufficient affection. ●

23-1d Screening to Uncover Private Information

When an informed party takes actions to reveal private information, the phenomenon is called signaling. When an uninformed party takes actions to induce the informed party to reveal private information, the phenomenon is called **screening**.

screening
an action taken by an uninformed party to induce an informed party to reveal information

Some screening is common sense. People buying used cars often ask that the car be checked by an auto mechanic before the sale. If sellers refuse this request, they reveal their private information that the car is not in good shape. Buyers may then decide to offer a lower price or to look for another car.

Other examples of screening are more subtle. For example, consider a firm that sells car insurance. It would like to charge a lower premium to safe drivers than to risky ones. But how can it tell them apart? Drivers may know better than the insurance companies whether they are safe or risky, but the risky ones won't admit it. A driver's history is one piece of information (which insurance companies use), but because of the randomness of many car accidents, history is an imperfect indicator of future risk.

The insurance company might be able to sort out the two kinds of drivers by offering a menu of insurance policies that would induce the drivers to separate themselves. One policy would have a high premium and cover the full cost of any accidents that occur. Another policy would have low premiums but would have, say, a $2,000 deductible. (That is, the driver would be responsible for the first $2,000 of damage, and the insurance company would cover the remaining risk.) The deductible is more of a burden for risky drivers because they are more likely to have an accident. With a large enough deductible, the low-premium policy with a deductible would attract the safe drivers, while the high-premium policy without a deductible would attract the risky drivers. Faced with these two policies, the two kinds of drivers would reveal their private information by choosing different policies.

23-1e Asymmetric Information and Public Policy

We have examined two kinds of asymmetric information: moral hazard and adverse selection. And we have seen how individuals may respond to the problem with signaling or screening. Now let's consider what the study of asymmetric information suggests about the proper scope of public policy.

The tension between market success and market failure is central to microeconomics. Chapter 7 showed that the equilibrium of supply and demand in competitive markets is efficient in the sense that it maximizes the total surplus that society can obtain in a market. Adam Smith's invisible hand seemed to reign supreme.

This conclusion was then tempered by the study of externalities (Chapter 10), public goods (Chapter 11), health economics (Chapter 12), imperfect competition (Chapters 16 through 18), and poverty (Chapter 21). Those chapters examined how governments can sometimes improve market outcomes.

The study of asymmetric information provides a new reason to be wary of markets. When some people know more than others, the market may fail to put resources to their best use. People with high-quality used cars may have trouble selling them for a good price because buyers are afraid of getting a lemon. People with few health problems may have trouble getting reasonably priced health insurance because insurance companies lump them together with those who have significant health problems.

Asymmetric information may justify government action in some cases, but three facts complicate the issue. First, the market can sometimes deal with information asymmetries on its own using a combination of signaling and screening. Second, the government rarely has more information than the private parties. Even if the market's allocation of resources is not ideal, it may be the best that can be achieved. That is, when there are information asymmetries, policymakers may find it hard to improve upon the market's admittedly imperfect outcome. Third, the government is itself an imperfect institution, as we discuss in the next section.

QuickQuiz

1. Because Elaine has a family history of significant medical problems, she buys health insurance, while her friend Jerry, who has a healthier family, goes without. This is an example of
   a. moral hazard.
   b. adverse selection.
   c. signaling.
   d. screening.
2. George has a life insurance policy that pays his family $1 million if he dies. As a result, he does not hesitate to enjoy his favorite hobby of bungee jumping. This is an example of
   a. moral hazard.
   b. adverse selection.
   c. signaling.
   d. screening.
3. Before selling anyone a life insurance policy, the Kramer Insurance Company requires that applicants undergo a medical examination. Those with significant preexisting medical problems are charged more. This is an example of
   a. moral hazard.
   b. adverse selection.
   c. signaling.
   d. screening.
4. Dr. Wexler displays her medical degree in her office waiting room, hoping patients will be impressed that she attended a prestigious medical school. This is an example of
   a. moral hazard.
   b. adverse selection.
   c. signaling.
   d. screening.

Answers are at the end of the chapter.

23-2 Political Economy

When the market's outcome is deemed to be either inefficient or inequitable, the government may be able to improve the situation. Yet before embracing an active government, we need to consider one more fact: The government is also an imperfect institution. The field of **political economy** (sometimes called **public choice** or **social choice**) uses the methods of economics to study how government works.

political economy the study of government using the analytic methods of economics

23-2a The Condorcet Voting Paradox

Democratic principles undergird the complex governmental institutions in the United States and in many other countries. Perhaps the most fundamental among these principles is majority rule. It seems straightforward enough. But majority rule can be more problematic than it seems.

Consider a town deciding where to build a new park. The town council decides to let the voters choose. How will this work? If there are only two options, the answer is simple: The majority gets its way. But what if the new park could be placed in many possible locations? In this case, as the 18th-century French political theorist Marquis de Condorcet noted, democracy might have trouble figuring out which option to pick.

Suppose there are three candidate locations, A, B, and C, and three voter types with the preferences shown in Table 1. The town council wants to aggregate these individual preferences into preferences for the town as a whole. How should it do it?

At first, the council might try pairwise voting. If it asks voters to choose first between B and C, voter types 1 and 2 will vote for B, giving B the majority. If it then asks voters to choose between A and B, voter types 1 and 3 will vote for A, giving A the majority. Observing that A beats B and that B beats C, the council might conclude that A is the voters' clear choice.

But wait: Suppose the town council then asks voters to choose between A and C. In this case, voter types 2 and 3 vote for C, giving C the majority. Under pairwise majority voting, A beats B, B beats C, and C beats A. Normally, we expect preferences to exhibit a property called **transitivity**: If A is preferred to B, and B is preferred to C, then we would expect A to be preferred to C. The **Condorcet paradox** is that democratic outcomes do not always obey this property. Pairwise voting might produce transitive preferences in some cases, but it cannot be counted on to do so.

Condorcet paradox
the failure of majority rule to produce transitive preferences for society

One implication of the Condorcet paradox is that the order in which things are voted on can affect the result. If the voters choose first between A and B and then compare the winner with C, the town ends up with C. But if the voters choose first between B and C and then compare the winner with A, the town ends up with A. And if the voters choose first between A and C and then compare the winner with B, the town ends up with B.

The Condorcet paradox teaches two lessons. The narrow lesson is that when there are more than two options, setting the agenda (that is, deciding the order in which items are voted on) can have a powerful influence over the outcome of a democratic election. The broad lesson is that majority voting by itself does not always tell us what outcome a society really wants.

Table 1

The Condorcet Paradox

If voters have these preferences over candidates A, B, and C, then in pairwise majority voting, A beats B, B beats C, and C beats A.

	Voter Type		
	Type 1	**Type 2**	**Type 3**
Percent of electorate	35	45	20
First choice	A	B	C
Second choice	B	C	A
Third choice	C	A	B

23-2b Arrow's Impossibility Theorem

Since political theorists first noticed the Condorcet paradox, they have spent much energy studying voting systems. Sometimes they propose new ones.

As an alternative to pairwise majority voting, the town council could ask each voter to rank the candidates. For each voter, we could give 1 point for last place, 2 points for second to last, 3 points for third to last, and so on. The candidate with the most total points wins. This voting method is called a **Borda count** after Jean-Charles de Borda, the 18th-century French mathematician and political theorist who devised it. It is often used in polls that rank sports teams. With the preferences in Table 1, you will find that B is the winner.

Another possible system is called **ranked-choice voting**, which many cities use to elect officials. Again, the voters rank all the candidates. Each vote is first assigned to that voter's first choice. The candidate with the fewest votes is then eliminated, and those votes are reallocated to the voters' second choice. And so on. Candidates are successively eliminated until one of them has a majority. With the preferences in Table 1, C is eliminated in the first round, the type 3 votes are reallocated from C to A, and A is the winner.

Is there a perfect voting system? The economist Kenneth Arrow addressed this question in his 1951 book *Social Choice and Individual Values*. Arrow started by defining what an ideal voting system would be. He assumed that individuals have preferences over the possible outcomes: A, B, C, and so on. He then assumed that society wants a voting system to choose among these outcomes that satisfies several properties:

- **Unanimity:** If everyone prefers A to B, then A beats B.
- **Transitivity:** If A beats B, and B beats C, then A beats C.
- **Independence of irrelevant alternatives:** The ranking between any two outcomes A and B does not depend on whether some third outcome C is also available.
- **No dictators:** There is no person who always gets to pick the outcome, regardless of everyone else's preferences.

These all seem desirable and compatible. Yet Arrow proved, mathematically and incontrovertibly, that **no voting system can satisfy all these properties**. This result is called **Arrow's impossibility theorem**.

Arrow's impossibility theorem
a mathematical result showing that, under certain assumed conditions, there is no method for aggregating individual preferences into a valid set of social preferences

The mathematics needed to prove Arrow's theorem is beyond the scope of this book, but for a sense of why the theorem is true, consider some examples. We have already seen the problem with the pairwise majority rule: The Condorcet paradox shows that it fails to produce a ranking of outcomes that always satisfies transitivity.

As another example, the Borda count fails to satisfy the independence of irrelevant alternatives. Recall that, using the preferences in Table 1, B wins with a Borda count. But suppose that C is no longer available. You will find that if the Borda count method is applied only to A and B, then A wins. Eliminating C changes the ranking between A and B because the Borda count depends on the number of points that A and B receive, and the number of points each receives depends on whether the irrelevant alternative, C, is in the running.

Ranked-choice voting also fails to satisfy the independence of irrelevant alternatives. Recall that, using the preferences in Table 1, ranked-choice voting produces A as the winner. But if B disappears as an alternative, the winner becomes C. This change occurs because when B drops out, more voters have C as their first choice, so rather than being eliminated, it wins a majority in the first round.

To be sure, the three systems discussed so far—pairwise majority voting, Borda counts, and ranked-choice voting—do not do justice to the rich and complex set of rules and institutions that make up actual democracies. But Arrow's impossibility theorem is a general result: It applies not just to these three simple systems but to any way a society might aggregate individual preferences to choose an outcome. Arrow's impossibility theorem doesn't say that we should abandon democracy as a form of government. But it does say that any democracy, no matter how well designed, must be flawed as a mechanism for social choice.

23-2c The Median Voter Is King

Despite Arrow's theorem, voting is how most societies choose leaders and public policies, often by majority rule. The next step in studying government is to examine how governments run by majority rule work. That is, in a democratic society, who determines what policy is chosen? In some cases, the theory of democratic government yields a surprisingly simple answer.

Imagine that society is deciding how much money to spend on a public good, such as the army or the national parks. Each voter has an optimal budget in mind, and voters always prefer outcomes closer to their optimum to outcomes farther away. We can line up voters from those who prefer the smallest budget to those who prefer the largest. Figure 1 is an example. Here, there are 100 voters, and the budget size varies from zero to $20 billion. Given these preferences, what outcome would a democracy produce?

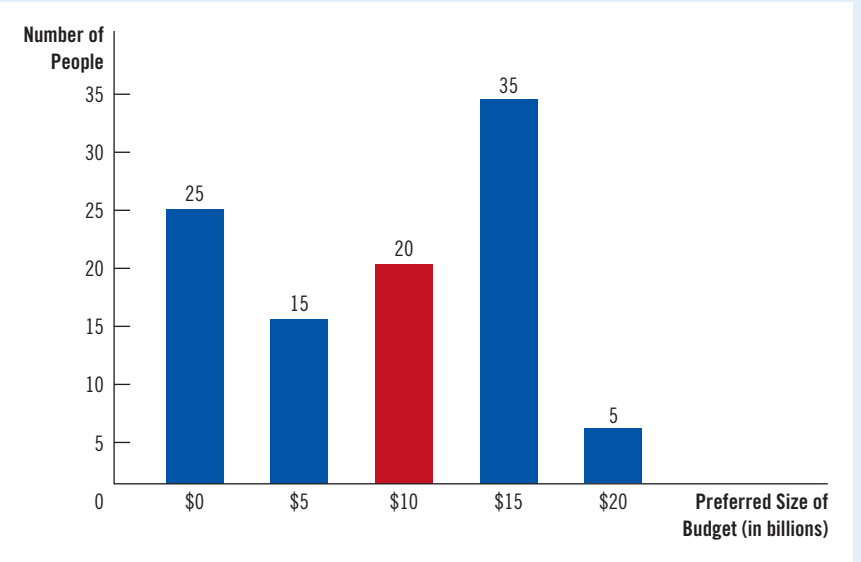
median voter theorem a mathematical result showing that if voters are choosing a point along a line and they all want the point closest to their own optimum, then majority rule will pick the optimum of the median voter

According to a result called the **median voter theorem**, majority rule produces the outcome preferred by the **median voter**, the person exactly in the middle of the distribution. In this example, if you take the line of voters ordered by their optimal budgets and count 50 voters from either end of the line, you will find that the median voter wants a budget of $10 billion. By contrast, the average preferred outcome (calculated by adding the preferred outcomes and dividing by the number

Figure 1

The Median Voter Theorem: An Example

This bar chart shows how 100 voters' optimal budgets are distributed over five options, ranging from zero to $20 billion. If society makes its choice by majority rule, the median voter, who here prefers $10 billion, determines the outcome.



of voters) is $9 billion, and the modal outcome (the one preferred by the greatest number of voters) is $15 billion.

The median voter rules because her preferred outcome beats any other proposal in a two-way race. In this example, more than half the voters want $10 billion or more, and more than half want $10 billion or less. If someone proposes, say, $8 billion instead of $10 billion, everyone who prefers $10 billion or more will vote with the median voter. Similarly, if someone proposes $12 billion instead of $10 billion, everyone who wants $10 billion or less will vote with the median voter. In either case, the median voter has more than half the voters on her side.

What about the Condorcet voting paradox? When the voters are picking a point along a line and each voter aims to be close to her own optimum, the Condorcet paradox cannot arise. The median voter's preferred outcome beats all challengers.

One implication of the median voter theorem is that if two rational political parties are each trying to win an election, they will both move their positions toward the one favored by the median voter. Suppose, for example, that the Democratic Party advocates a budget of $15 billion, while the Republican Party advocates a budget of $10 billion. The Democratic position is more popular in the sense that $15 billion has more proponents than any other single choice. Nonetheless, the Republicans get more than 50 percent of the vote: They will attract the 20 voters who want $10 billion, the 15 voters who want $5 billion, and the 25 voters who want zero. If the Democrats want to win, they will move their platform toward the median voter. This theory can explain why the parties in a two-party system are often similar: They have good reason to move toward the median voter.

Another implication of the median voter theorem is that minority views are not given much weight. Imagine that 40 percent of the population want a lot of money spent on the national parks, and 60 percent want nothing spent. In this case, the median voter's preference is zero, regardless of the intensity of the minority's view. Rather than reaching a compromise that takes everyone's preferences into account, majority rule looks only to the person in the exact middle of the distribution. That is part of the logic of democracy.

23-2d Politicians Are People Too

When economists study consumer behavior, they assume that consumers buy the bundle of goods and services that gives them the greatest level of satisfaction. When economists study firm behavior, they assume that firms produce the quantity of goods and services that yields the greatest profit. What should they assume about the objectives of people involved in politics?

It would be comforting to assume that political leaders always look out for the well-being of society, that they aim for an optimal combination of efficiency and equality. Comforting, perhaps, but not realistic. Self-interest is as powerful a motive for political actors as it is for consumers and firm owners. Some politicians, motivated by a desire for reelection, are willing to sacrifice the national interest to solidify their base of voters. Others are motivated by simple greed. If you have any doubt, look at the world's poorest nations, where rampant corruption is often an impediment to economic development. Graft among political leaders is less common in advanced nations, but scandals still occur from time to time, and those scandals reveal only the corrupt leaders who were caught.

This book is not the place to develop a full-blown theory of political behavior. But when thinking about economic policy, remember that it is made not by a benevolent king (or even by benevolent economists) but by real people

CHRIS WILDT/CARTOONSTOCK LTD

"Isn't that the real genius of democracy? . . . The VOTERS are ultimately to blame."

with all-too-human desires. Sometimes they are motivated to further the national interest, but sometimes they are moved by their own political and financial ambitions. Don't be surprised when economic policy fails to resemble the ideals derived in economics textbooks.

QuickQuiz

5. The Condorcet paradox illustrates Arrow's impossibility theorem by showing that pairwise majority voting
   a. is inconsistent with the principle of unanimity.
   b. leads to social preferences that are not transitive.
   c. violates the independence of irrelevant alternatives.
   d. makes one person, in effect, a dictator.
6. Georgette is about to win reelection as class president against her challenger Billie. But then Rossana enters the race as well, pulling votes from Georgette and allowing Billie to prevail. The school's voting system
   a. is inconsistent with the principle of unanimity.
   b. leads to social preferences that are not transitive.
   c. violates the independence of irrelevant alternatives.
   d. makes one person, in effect, a dictator.
7. Two political candidates are vying for town mayor, and the key issue is how much to spend on the annual Fourth of July fireworks. Among the 100 voters, 40 want to spend $30,000, 30 want to spend $10,000, and 30 want to spend nothing at all. What is the winning position on this issue?
   a. $10,000
   b. $15,000
   c. $20,000
   d. $30,000

Answers are at the end of the chapter.

23-3 Behavioral Economics

behavioral economics the subfield of economics that integrates the insights of psychology

Economics studies human behavior, but it is not the only field that makes that claim. The social science of psychology also sheds light on the choices that people make in their lives. The fields of economics and psychology are largely independent of one another, in part because they address different questions. But recently, a field called **behavioral economics** has emerged in which economists use psychological insights to better understand the decisions that people make.

23-3a People Aren't Always Rational

Economic theory is populated by a particular species of organism, sometimes called **Homo economicus**. Members of this species are always rational. As firm owners, they maximize profits. As consumers, they maximize utility (or equivalently, pick the point on the highest indifference curve). Given the constraints they face, they rationally weigh all the costs and benefits and always choose the best possible course of action.

Real people, however, are **Homo sapiens**. Although in many ways they resemble the rational, calculating people assumed in economic theory, they are more complex. They can be forgetful, impulsive, confused, emotional, and shortsighted. These characteristics are a central focus of psychologists but, until recently, have often been neglected by economists.

Herbert Simon, one of the first social scientists to work at the boundary of economics and psychology, suggested that humans should be viewed not as rational maximizers but as what he called **satisficers**. Instead of always choosing the best course of action, they make decisions that are good enough. Similarly, other economists have suggested that humans are only "near rational" or that they exhibit "bounded rationality."

Studies of human decision making have found that people often make the following systematic mistakes:

- **People are overconfident.** Researchers have demonstrated that people frequently believe they know more than they do. Imagine, for example, that you were asked a series of numerical questions, such as how many African countries are in the United Nations, what the height of the tallest mountain in North America is, how many gold medals China won in the last Olympics, and so on. For each question, instead of being asked for a single estimate, you were asked to give a 90 percent **confidence interval**—a range that you were 90 percent confident the true number falls within. When psychologists run experiments like this, they find that most people give ranges that are too small: The true number falls within their intervals far less than 90 percent of the time. In other words, most people are too sure of their own abilities.
- **People give too much weight to a small number of vivid observations.** Imagine that you are thinking about buying a car of a particular brand. To learn about its reliability, you read *Consumer Reports*, which has systematically surveyed 1,000 owners of the car. Then you run into a friend who owns one, and she tells you that her car is a lemon. How do you treat your friend's observation? If you think rationally, you will realize that she has only increased your sample size from 1,000 to 1,001, providing little new information. But because your friend's story is vivid—simply because it comes from your friend—you may be tempted to give it more weight in your decision making than you should. This phenomenon has been called **undue salience**.
- **People are reluctant to change their minds.** People tend to interpret evidence to confirm beliefs they already hold. In one study, subjects were asked to read and evaluate a research report on whether capital punishment deters crime. After reading the report, those who initially favored the death penalty said they were more certain of their view, and those who initially opposed the death penalty also said they were more certain of their view. The two groups interpreted the same evidence in exactly opposite ways. This behavior is called **confirmation bias**.

Think about decisions you have made in your own life. Have you exhibited any of these traits?

Deviations from rationality are important for understanding some economic phenomena. An intriguing example arises in the study of 401(k) plans, the tax-advantaged retirement savings accounts that many firms offer their workers. In some firms, workers can choose to participate in the plan by filling out a simple form. In others, workers are automatically enrolled and can opt out of the plan by filling out a simple form. Many more people participate in the second case than in the first. If they were perfectly rational maximizers, they would choose the optimal amount of retirement saving, regardless of the default offered by their employer. In fact, employees' behavior appears to exhibit substantial inertia. Understanding their behavior seems easier once we abandon the model of rational humans.

Why, you might ask, is economics built on the rationality assumption when psychology and common sense cast doubt on it? One answer is that the assumption, even if not exactly true, may be close enough that it yields workable models of behavior. For example, when we compared competitive and monopoly firms,

the assumption that firms rationally maximize profit yielded many important and valid insights. Incorporating deviations from rationality into the story might have added a bit of realism, but it would have also muddied the waters and made those insights harder to find. Recall from Chapter 2 that economic models are meant not to replicate reality but to show the essence of the problem at hand.

Another reason economists often assume rationality may be that economists are themselves not rational maximizers. Like most people, they are overconfident and reluctant to change their minds. Their choice among alternative theories of human behavior may exhibit excessive inertia. Moreover, economists may be content with a theory that is not perfect but is good enough. The model of rational humans may be the theory of choice for a satisficing social scientist.

23-3b People Care about Fairness

Another insight about human behavior is best illustrated by an experiment called the **ultimatum game**. It works like this: Two volunteers (who are strangers to each other) are told that they are going to play a game and could win a total of $100. The game begins with a coin toss, which is used to assign the volunteers to two roles: proposer and responder. The proposer must suggest a division of the $100 prize between the two players. After the proposer makes the offer, the responder decides whether to accept or reject it. If the responder accepts it, both players are paid accordingly. If the responder rejects the offer, neither player gets anything. In either case, the game then ends.

Before proceeding, think about what you would do. If you were the proposer, what division of the $100 would you offer? If you were the responder, what offers would you accept?

Conventional economic theory assumes that people in this situation are rational wealth maximizers. This assumption leads to a simple prediction: The proposer will offer to take $99 and to give the responder $1, and the responder will accept the offer. After all, once the offer is made, the responder is better off accepting it: $1 is better than nothing. And because the proposer knows that accepting the offer is in the responder's interest, the proposer has no reason to offer more than $1. In the language of game theory (discussed in Chapter 18), the 99–1 split is the Nash equilibrium.

Yet when experimental economists ask real people to play the ultimatum game, the results differ from this prediction. People in the responder's role usually reject offers of only $1 or a similarly small amount. Anticipating this, people in the proposer's role usually offer the responders much more than $1. Some people offer a 50–50 split, but it is more common for proposers to offer the responders an amount such as $30 or $40, keeping the larger share for themselves. In these cases, the responders usually accept the proposal.

What's going on here? One interpretation is that people are driven in part by some innate sense of fairness. A 99–1 split seems so wildly unfair to many people that they reject it, even to their own financial detriment. By contrast, a 70–30 split is still unfair, but it is not so bad that it induces people to abandon their normal self-interest.

Throughout our study of household and firm behavior, the innate sense of fairness has not played any role. But the results of the ultimatum game suggest that perhaps it should. For example, in Chapters 19 and 20, we discussed how wages were determined by labor supply and labor demand. Some economists have suggested that the perceived fairness of what a firm pays its workers should also enter the picture. When a firm has an especially good year, workers (like the responder) may expect

to be paid a fair share of the prize, even if the conventional market equilibrium does not dictate it. The firm (like the proposer) might well decide to give workers more than the equilibrium wage for fear that the workers might otherwise try to punish the firm with reduced effort, strikes, or even vandalism.

23-3c People Are Inconsistent over Time

Imagine a task that you find dreary. It might be doing your laundry, shoveling snow off your driveway, or filling out your income tax forms. Now consider the following questions:

1. Would you prefer (A) to spend 50 minutes doing the task right now or (B) to spend 60 minutes doing the task tomorrow?
2. Would you prefer (A) to spend 50 minutes doing the task in 90 days or (B) to spend 60 minutes doing the task in 91 days?

When asked questions like these, many people choose B for question 1 and A for question 2. When looking ahead to the future (as in question 2), they minimize the amount of time spent on the dreary task. But faced with the prospect of doing the task immediately (as in question 1), they choose to put it off.

In some ways, this behavior is not surprising: Everyone procrastinates sometimes. But from the standpoint of conventional rationality, it is puzzling. Suppose that in response to question 2, people choose to spend 50 minutes in 90 days. Then, when the 90th day arrives and they are about to start the dreary task, we allow them to change their mind. In effect, they then face question 1, so they opt to put off the task until the next day. But why should the mere passage of time affect the choices they make?

Often in life, people make plans for themselves but then fail to follow through. Smokers promise themselves they will quit, but within a few hours of smoking their last cigarette, they crave another one and renege on their commitment. People trying to lose weight promise themselves they will stop eating dessert, but when the waiter brings the dessert cart, the diet goes out the window. In both cases, the desire for instant gratification overwhelms the rational plans set in the past.

Some economists believe that the consumption–saving decision is an important instance in which people exhibit this inconsistency over time. For many people, spending provides a type of instant gratification. Saving, like passing up the cigarette or the dessert, requires a sacrifice in the present for a reward in the distant future. And just as many smokers wish they could quit and many people wish they ate less, many consumers wish they saved more of their income. According to one survey, 76 percent of Americans said they were not saving enough for retirement.

An implication of this inconsistency over time is that people might be better off if they found ways to commit their future selves to following through on their plans. Smokers trying to quit may throw away their cigarettes, and people on a diet may put locks on their refrigerators. What can people who save too little do? Find some way to lock up their money before they spend it. Some retirement accounts, such as 401(k) plans, allow people to do something much like that.

Workers can agree to have some money taken out of their paychecks before they ever see it. The money is deposited in a retirement account, and if it is withdrawn before a certain age, there is a financial penalty. Perhaps that is one reason these retirement accounts are so popular: They protect people from their own desires for instant gratification.

QuickQuiz

8. One documented deviation from rationality is that many people
   a. tend to be excessively confident in their own abilities.
   b. change their mind too quickly when they get new information.
   c. give too much weight to outcomes that will occur far in the future.
   d. make decisions by equating marginal benefits and marginal costs.

9. The experiment called the ultimatum game illustrates that people
   a. play the Nash equilibrium in strategic situations.
   b. are motivated by the desire for instant gratification.
   c. care about fairness, even to their own detriment.
   d. make decisions that are inconsistent over time.

Answers are at the end of the chapter.

Faults in Risk Assessment

Even after being fully vaccinated, many people remained fearful of Covid-19. Why is that, and what does it say about people's ability to assess risk?

Irrational Covid Fears

By David Leonhardt

Guido Calabresi, a federal judge and Yale law professor, invented a little fable that he has been telling law students for more than three decades.

He tells the students to imagine a god coming forth to offer society a wondrous invention that would improve everyday life in almost every way. It would allow people to spend more time with friends and family, see new places and do jobs they otherwise could not do. But it would also come with a high cost. In exchange for bestowing this invention on society, the god would choose 1,000 young men and women and strike them dead.

Calabresi then asks: Would you take the deal? Almost invariably, the students say no. The professor then delivers the fable's lesson: "What's the difference between this and the automobile?"

In truth, automobiles kill many more than 1,000 young Americans each year; the total U.S. death toll hovers at about 40,000 annually. We accept this toll, almost unthinkingly, because vehicle crashes have always been part of our lives. We can't fathom a world without them.

It's a classic example of human irrationality about risk. We often underestimate large, chronic dangers, like car crashes or chemical pollution, and fixate on tiny but salient risks, like plane crashes or shark attacks.

One way for a risk to become salient is for it to be new. That's a core idea behind Calabresi's fable. He asks students to consider whether they would accept the cost of vehicle travel if it did not already exist. That they say no underscores the very different ways we treat new risks and enduring ones.

I have been thinking about the fable recently because of Covid-19. Covid certainly presents a salient risk: It's a global pandemic that has upended daily life for more than a year. It has changed how we live, where we work, even what we wear on our faces. Covid feels ubiquitous.

Fortunately, it is also curable. The vaccines have nearly eliminated death, hospitalization and other serious Covid illness among people who have received shots. The vaccines have also radically reduced the chances that people contract even a mild version of Covid or can pass it on to others.

Yet many vaccinated people continue to obsess over the risks from Covid—because they are so new and salient.

To take just one example, major media outlets trumpeted new government data last week showing that 5,800 fully vaccinated Americans had contracted Covid. That may sound like a big number, but it indicates that a vaccinated person's chances of getting Covid are about one in 11,000. The chances of getting a version

23-4 Conclusion

This chapter took you on a trip to the frontier of microeconomics. You may have noticed that we sketched out ideas rather than fully developing them. One reason is that you might study these topics in more detail in advanced courses. Another is that these topics remain active areas of research and are still being fleshed out.

To see how these topics fit into the broader picture, recall the **Ten Principles of Economics** from Chapter 1. One principle states that markets are usually a good way to organize economic activity. Another says that governments can sometimes improve market outcomes. As you study economics, you can more fully appreciate these principles as well as the caveats that come with them. The study of asymmetric information may make you more wary of market outcomes. The study of political economy may make you more wary of government solutions. And the study of behavioral economics may make you wary of any institution that relies on human decision making, including both the market and the government.

If these topics have a common theme, it is that life is messy. Information is imperfect, government is imperfect, and people are imperfect. Of course, you knew this long before you started studying economics. But economists need to understand these imperfections as precisely as they can if they are to explain, and perhaps improve, the world around them.

any worse than a common cold are even more remote.

But they are not zero. And they will not be zero anytime in the foreseeable future. Victory over Covid will not involve its elimination. Victory will instead mean turning it into the sort of danger that plane crashes or shark attacks present—too small to be worth reordering our lives.

That is what the vaccines do. If you're vaccinated, Covid presents a minuscule risk to you, and you present a minuscule Covid risk to anyone else. A car trip is a bigger threat, to you and others. About 100 Americans are likely to die in car crashes today. The new federal data suggests that either zero or one vaccinated person will die today from Covid….

Coming to grips with the comforting realities of post-vaccination life is going to take some time for most of us. It's only natural that so many vaccinated people continue to harbor irrational fears. Yet slowly recognizing that irrationality will be a vital part of overcoming Covid.

"We're not going to get to a place of zero risk," Jennifer Nuzzo, a Johns Hopkins epidemiologist, told me during a virtual Times event last week. "I don't think that's the right metric for feeling like things are normal."

FLY VIEW PRODUCTIONS/E+/GETTY IMAGES

After Nuzzo made that point, Dr. Ashish Jha of Brown University told us about his own struggle to return to normal. He has been fully vaccinated for almost two months, he said, and only recently decided to meet a vaccinated friend for a drink, unmasked. "It was hard—psychologically hard—for me," Jha said.

"There are going to be some challenges to re-acclimating and re-entering," he added. "But we've got to do it."

And how did it feel in the end, I asked, to get together with his friend?

"It was awesome," Jha said. ■

Questions to Discuss

1. How do you think you would have answered the question in Calabresi's parable? Do you think your answer is rational?
2. If you were vaccinated for Covid-19, how did the vaccination change your attitudes and behavior? Do you think your response was fully rational?

Source: *New York Times*, April 19, 2021.

Chapter in a Nutshell

- In many transactions, information is asymmetric. When there are hidden actions, principals may be concerned that agents suffer from the problem of moral hazard. When there are hidden characteristics, buyers may be concerned about the problem of adverse selection among the sellers. Private markets sometimes deal with asymmetric information with signaling and screening.
- Although government policy can sometimes improve market outcomes, governments are themselves imperfect institutions. The Condorcet paradox shows that majority rule fails to produce transitive preferences for society, and Arrow's impossibility theorem shows that no voting system can be perfect. In many situations, democratic institutions will produce the outcome desired by the median voter, regardless of the preferences of the rest of the electorate. Moreover, the individuals who set government policy may be motivated by self-interest rather than the national interest.
- The study of psychology and economics reveals that human decision making is more complex than is assumed in conventional economic theory. People are not always rational, they care about the fairness of economic outcomes (even to their own detriment), and they can be inconsistent over time.

Key Concepts

moral hazard, p. 472
agent, p. 472
principal, p. 472
adverse selection, p. 474
signaling, p. 474
screening, p. 476
political economy, p. 477
Condorcet paradox, p. 478
Arrow's impossibility theorem, p. 479
median voter theorem, p. 480
behavioral economics, p. 482

Questions for Review

1. What is moral hazard? List three things an employer might do to reduce the severity of this problem.
2. What is adverse selection? Give an example of a market in which adverse selection might be a problem.
3. Define **signaling** and **screening** and give an example of each.
4. What unusual property of voting did Condorcet notice?
5. Explain why majority rule respects the preferences of the median voter rather than those of the average voter.
6. Describe the ultimatum game. What outcome from this game does conventional economic theory predict? Do experiments confirm this prediction? Explain.

Problems and Applications

1. Each of the following situations involves moral hazard. In each case, identify the principal and the agent and explain why there is asymmetric information. How does the action described reduce the problem of moral hazard?
   a. Landlords require tenants to pay security deposits.
   b. Firms compensate top executives with options to buy company stock at a given price in the future.
   c. Car insurance companies offer discounts to customers who install antitheft devices in their cars.
2. A case study in this chapter describes how a boyfriend can signal his love to a girlfriend by giving an appropriate gift. Do you think saying "I love you" can also serve as a signal? Why or why not?
3. The Affordable Care Act signed into law by President Obama in 2010 included the following two provisions:
   i. Insurance companies must offer health insurance to everyone who applies and charge them the same price regardless of a person's preexisting health condition.
   ii. Everyone must buy health insurance or pay a penalty for not doing so.
   a. Which of these policies taken on its own makes the problem of adverse selection worse? Explain.
   b. Why do you think the policy you identified in part (a) was included in the law?
   c. Why do you think the other policy was included in the law?

4. Ken walks into an ice-cream parlor.

   Waiter: "We have vanilla and chocolate today."
   Ken: "I'll take vanilla."
   Waiter: "I almost forgot. We also have strawberry."
   Ken: "In that case, I'll take chocolate."

   What standard property of decision making is Ken violating? (Hint: Reread the section on Arrow's impossibility theorem.)

5. Three friends are choosing a restaurant for dinner. Here are their preferences:

	Rachel	Ross	Joey
First choice	Italian	Italian	Chinese
Second choice	Chinese	Chinese	Mexican
Third choice	Mexican	Mexican	French
Fourth choice	French	French	Italian

   a. If the three friends use a Borda count to make their decision, where do they go to eat?
   b. On their way to their chosen restaurant, they see that the Mexican and French restaurants are closed, so they use a Borda count again to decide between the remaining two restaurants. Where do they decide to go now?
   c. How do your answers to parts (a) and (b) relate to Arrow's impossibility theorem?

6. Three friends are choosing a TV show to watch. Here are their preferences:

	Chandler	Phoebe	Monica
First choice	*NCIS*	*Ted Lasso*	*Survivor*
Second choice	*Ted Lasso*	*Survivor*	*NCIS*
Third choice	*Survivor*	*NCIS*	*Ted Lasso*

   a. If the three friends try using a Borda count to make their choice, what would happen?
   b. Monica suggests a vote by majority rule. She proposes that they first choose between *NCIS* and *Ted Lasso*, and they then choose between the winner of the first vote and *Survivor*. If they all vote their preferences honestly, what outcome would occur?
   c. Should Chandler agree to Monica's suggestion? What voting system would he prefer?
   d. Phoebe and Monica convince Chandler to go along with Monica's proposal. In round one, Chandler dishonestly says he prefers *Ted Lasso* to *NCIS*. Why might he do this?

7. Five roommates are planning to spend the weekend in their apartment watching movies, and they are debating how many movies to watch. The table below shows each roommate's willingness to pay for each of the movies:

	Ava	Ridley	Spike	Chloe	Quentin
First film	$14	$10	$8	$4	$2
Second film	12	8	4	2	0
Third film	10	6	2	0	0
Fourth film	6	2	0	0	0
Fifth film	2	0	0	0	0

   A movie on their streaming service costs $15, which the roommates split equally, so each pays $3 per movie.

   a. What is the efficient number of movies to watch (that is, the number that maximizes total surplus)?
   b. For each roommate, what is the preferred number of movies to watch?
   c. What is the preference of the median roommate?
   d. If the roommates held a vote on the efficient outcome versus the median voter's preference, how would each person vote? Which outcome would get a majority?
   e. If one of the roommates proposed a different number of movies, could the proposal beat the winning outcome from part (d) in a vote?
   f. Can majority rule be counted on to reach efficient outcomes in the provision of public goods?

8. Two ice-cream stands are deciding where to set up along a one-mile beach. The people are uniformly located along the beach, and each person sitting on the beach buys exactly one ice-cream cone per day from the nearest stand. Each ice-cream seller wants the maximum number of customers. Where along the beach will the two stands locate? Of which result in this chapter does this outcome remind you?

9. The government is considering two ways to help low-income families: giving them cash or giving them free meals at soup kitchens.
   a. Give an argument, based on the standard theory of the rational consumer, for giving cash.
   b. Give an argument, based on asymmetric information, for why free meals at soup kitchens may be better than cash handouts.
   c. Give an argument, based on behavioral economics, for why free meals at soup kitchens may be better than cash handouts.

QuickQuiz Answers

1. **b** 2. **a** 3. **d** 4. **c** 5. **b** 6. **c** 7. **a** 8. **a** 9. **c**

Chapter

24

Measuring a Nation's Income

When you finish school and look for a full-time job, prevailing economic conditions will, to a large extent, shape your experience. In some years, firms throughout the economy are expanding their production of goods and services, employment is rising, and job openings are plentiful. In other years, firms are cutting back production, employment is declining, and help-wanted ads are few and far between. College graduates who start their job hunt in a year of economic expansion have a much easier time of it than those who start in a year of economic contraction.

Because the health of the overall economy affects everyone, changes in economic conditions are widely reported. Indeed, the news is filled with economic statistics. These numbers might measure the total income of everyone in the economy (gross domestic product), the rate at which average prices are rising or falling (inflation/deflation), the percentage of the labor force that is out of work (unemployment), total spending at stores (retail sales),

SHIFT DRIVE/SHUTTERSTOCK.COM; ERDARKE+/GETTY IMAGES

or the imbalance of trade between the United States and the rest of the world (the trade deficit). All these statistics are **macroeconomic**. Rather than focusing on a particular household, firm, or market, they provide information about the entire economy.

microeconomics
the study of how households and firms make decisions and how they interact in markets

macroeconomics
the study of economy-wide phenomena, including inflation, unemployment, and economic growth

As Chapter 2 discussed, economics is divided into two branches: microeconomics and macroeconomics. **Microeconomics** studies how individual households and firms make decisions and interact in markets. **Macroeconomics** studies the economy as a whole, with the goal of explaining changes that affect many households, firms, and markets simultaneously. Macroeconomists address a broad range of questions: Why is average income high in some countries and low in others? Why are prices sometimes rapidly rising and other times more stable? Why do production and employment expand in some years and contract in others? How can the government promote rapid growth in incomes, low inflation, and stable employment? These questions concern the workings of the entire economy, so they fall under the scope of macroeconomics.

Because the overall economy is a collection of households and firms interacting in many markets, microeconomics and macroeconomics are closely linked. The tools of supply and demand, for instance, are as central to macroeconomic analysis as they are to microeconomic analysis. Yet studying the economy in its entirety raises special challenges. To understand a forest, it's not enough to analyze each of the trees.

This chapter and the next discuss some of the data that economists and policymakers use to monitor the performance of the overall economy. Our starting point is **gross domestic product** (GDP), which measures the total income of a nation. GDP is often thought to be the single best measure of a society's economic well-being.

24-1 The Economy's Income and Expenditure

If you were to judge how a family is doing economically, you might first look at its income. Higher incomes usually translate into higher standards of living—larger houses, better healthcare, fancier cars, more opulent vacations, and so on.

The same logic applies to a nation's economy. When judging whether the economy is doing well or poorly, it is natural to look at the aggregate income that everyone in the economy is earning. Gross domestic product allows us to do just that.

GDP measures two things at once: the total income of everyone in the economy and the total expenditure on the economy's output of goods and services. GDP can perform the trick of measuring both total income and total expenditure because these two things are the same. **For an economy as a whole, income must equal expenditure.**

This is because every transaction has two parties: a buyer and a seller. Every dollar of spending by some buyer is a dollar of income for some seller. If Karen pays Doug $100 to mow her lawn, Doug earns $100 as a seller of a service, and Karen spends $100 as a buyer of the service. The transaction contributes equally to the economy's income and to its expenditure. Whether GDP is measured as total income or total expenditure, it rises by $100.

Another way to see the equality of income and expenditure is with the circular-flow diagram in Figure 1. As you may recall from Chapter 2, this diagram describes all the transactions between households and firms. It simplifies matters by assuming

that all goods and services are bought by households and that households spend all of their income. In this economy, when households buy goods and services from firms, these expenditures flow through the markets for goods and services. When the firms use the money they receive from sales for workers' wages, landowners' rent, and firm owners' profit, this income flows through the markets for the factors of production. Money continuously flows from households to firms and then back to households.

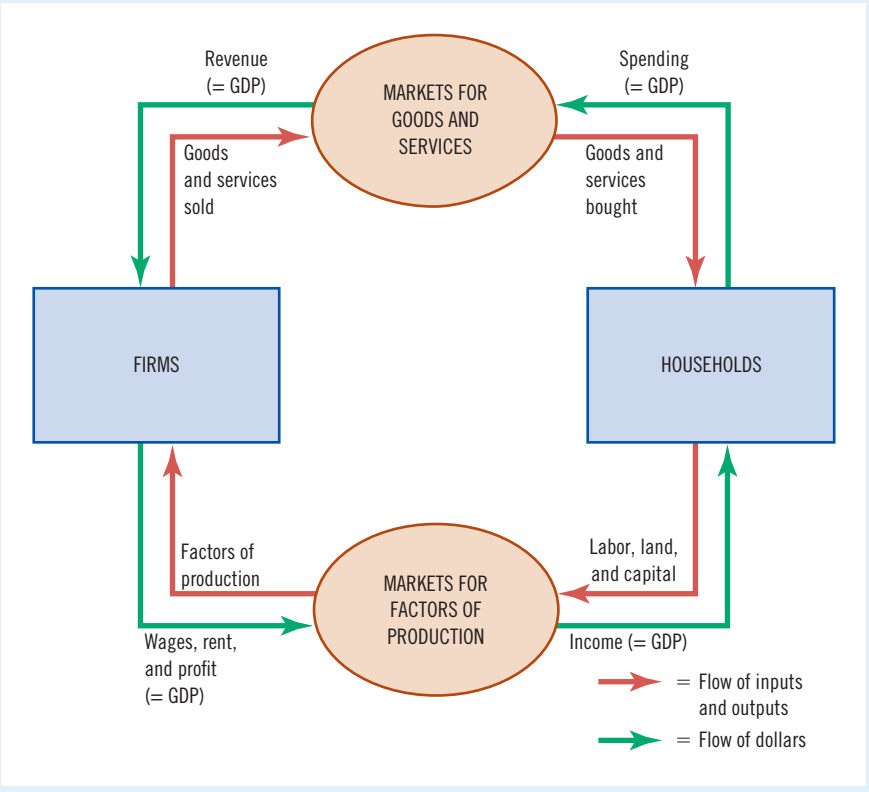
GDP measures the flow of money around the economy. We can compute it in two ways: by adding up all the expenditure by households or by adding up all the income (wages, rent, and profit) paid by firms. Because all expenditure in the economy ends up as someone's income, GDP is the same regardless of how we compute it.

The actual economy is far more complex than the one shown in Figure 1. Households do not spend all of their income. They pay some of it to the government in taxes and save some for use in the future. In addition, households do not buy all goods and services produced in the economy. Some goods and services are bought by governments for public purposes, and some are bought by firms to expand their production capacity. Yet the basic lesson remains the same: Regardless of whether a household, government, or firm buys a good or service, the transaction always has a buyer and a seller. For the economy as a whole, expenditure equals income.

Figure 1

The Circular-Flow Diagram

Households buy goods and services from firms, and firms use their revenue from sales for wages to workers, rent to landowners, and profit to firm owners. GDP equals the total amount spent by households in the markets for goods and services. It also equals the total wages, rent, and profit paid by firms in the markets for the factors of production.



QuickQuiz

1. An economy's gross domestic product is
   a. the excess of spending over income.
   b. the excess of income over spending.
   c. total income and total spending.
   d. total spending as a percent of total income.

2. Carla bakes a cake and sells it to Norm for $10. Sam pays Diane $30 to tutor him. In this economy, GDP is
   a. $10.
   b. $20.
   c. $30.
   d. $40.

Answers are at the end of the chapter.

24-2 The Measurement of GDP

Having discussed gross domestic product in general terms, let's be more precise about how it is measured. Here is a definition of GDP that focuses on expenditure:

gross domestic product (GDP) the market value of all final goods and services produced within a country in a given period

- **Gross domestic product (GDP)** is the market value of all final goods and services produced within a country in a given period.

This definition might seem simple enough. But in fact, many subtle issues arise when computing an economy's GDP. Let's consider each phrase in this definition.

24-2a "GDP Is the Market Value . . ."

"You can't compare apples and oranges." This adage is often true, yet GDP does exactly that. GDP adds together many different products into a single measure of the value of economic activity. It does this using market prices to translate quantities into a measure of value. Prices are a gauge of value because they reflect the amount people are willing to pay for different goods. If the price of an apple is twice the price of an orange, then an apple contributes twice as much to GDP as an orange.

24-2b ". . . of All . . ."

GDP tries to be comprehensive. It includes all items produced in the economy and sold legally in markets. GDP measures the market value of not just apples and oranges but also pears and grapefruit, books and movies, haircuts and healthcare, and so on.

GDP also includes the market value of the housing services that flow from the economy's stock of housing. For rental housing, this value is easy to calculate—the rent equals both the tenant's expenditure and the landlord's income. Yet many people do not pay rent because they own their homes. The government includes this owner-occupied housing in GDP by estimating its rental value. In effect, the national income accounts assume that homeowners rent their houses to themselves. This imputed rent is included both in the homeowners' expenditure and in their income, so it adds to GDP.

But GDP excludes some products whose value is hard to measure. In this category are most items produced and sold illicitly, such as illegal drugs. GDP also excludes most items that never enter the marketplace because they are produced and consumed at home. For instance, GDP includes vegetables you buy at the grocery store but not those you grow in your garden. When you are working at home, what you produce for your employer is part of GDP, but what you produce for your family is not.

These exclusions from GDP can lead to paradoxical results. When Karen pays Doug to mow her lawn, that transaction is part of GDP. But suppose Doug and

Karen marry. Even though Doug may continue to mow Karen's lawn, the value of the mowing is now left out of GDP because Doug's service is no longer sold in a market. Their marriage may well enhance their well-being and the total well-being of society, but it reduces GDP.

24-2c ". . . Final . . ."

When Intel makes a microchip that Dell uses to make a computer, the microchip is an **intermediate good**, and the computer is a **final good**. GDP includes only the value of final goods because the value of intermediate goods is already included in the prices of the final goods. Adding the market value of the microchip to the market value of the computer would be double counting: It would (incorrectly) count the microchip twice.

An exception to this principle arises when an intermediate good is produced and, rather than being used, is added to a firm's inventory of goods for use or sale later. In this case, the intermediate good is taken to be "final" for the moment, and its value as inventory investment is included in GDP. Thus, additions to inventory add to GDP, and when the goods in inventory are later used or sold, the reductions in inventory subtract from GDP.

24-2d ". . . Goods and Services . . ."

GDP includes tangible goods (food, clothing, cars) and intangible services (haircuts, house cleaning, doctor visits). When you buy a T-shirt with your favorite band's logo, you are buying a good, and the purchase price is part of GDP. When you pay to hear a concert by the same band, you are buying a service, and the ticket price is also part of GDP.

24-2e ". . . Produced . . ."

GDP includes goods and services currently produced. It does not include transactions involving items produced in the past. When Ford produces and sells a new car, the value of the car is included in GDP. But when one person sells a used car to another, the value of the used car is not included in GDP.

24-2f ". . . Within a Country . . ."

GDP measures the value of production within the geographic confines of a country. When Canadians work temporarily in the United States, their production counts toward U.S. GDP. When Americans own a factory in Haiti, the production at the factory does not contribute to U.S. GDP. (It adds to Haiti's GDP.) Items are included in a nation's GDP if they are produced domestically, regardless of the nationality of the producer.

24-2g ". . . In a Given Period."

GDP measures the value of production that takes place within a specific interval of time. Usually, that interval is a year or a quarter (three months). GDP measures the economy's flow of income, as well as its flow of expenditure, during that interval.

When the government reports the GDP for a quarter, it usually presents GDP "at an annual rate." This means that the figure reported for quarterly GDP is the amount of income and expenditure during the quarter multiplied by four. The government uses this convention so that quarterly and annual figures on GDP can be compared more easily.

In addition, when the government reports quarterly GDP, it presents the data after they have been modified by a statistical procedure called **seasonal adjustment**. The unadjusted data show that the economy produces more goods and services

during some times of the year than others. (December's holiday shopping season is a high point.) When monitoring the economy, economists and policymakers often want to look beyond these seasonal changes. Therefore, government statisticians adjust the quarterly data to take out the seasonal cycle. The GDP data reported in the news are always seasonally adjusted.

Now let's repeat the definition of GDP:

- Gross domestic product (GDP) is the market value of all final goods and services produced within a country in a given period.

This definition focuses on GDP as total expenditure in the economy. But recall that every dollar spent by a buyer of a good or service becomes a dollar of income to the seller of that good or service. Therefore, in addition to adding up total expenditure in the economy to calculate GDP, the government also adds up total income in the economy to arrive at **gross domestic income** (GDI). GDP and GDI give almost exactly the same number. Why "almost"? The two measures should be precisely the same, but data sources are not perfect. The difference between GDP and GDI is called the **statistical discrepancy**.

It should be apparent that GDP is a sophisticated measure of the value of economic activity. In advanced courses in macroeconomics, you will learn more about the nuances of its calculation. But even now, you can see that each phrase in this definition is packed with meaning.

Other Measures of Income

When the U.S. Department of Commerce computes the nation's GDP, it also calculates other measures of income to get a more complete picture of what's happening in the economy. These other measures differ from GDP by excluding or including certain categories of income. What follows is a brief description of five of these income measures, ordered from largest to smallest.

- **Gross national product** (GNP) is the total income earned by a nation's permanent residents (called **nationals**). It differs from GDP in that it includes income that domestic residents earn abroad and excludes income that foreigners earn here. For example, when Canadians work temporarily in the United States, the income from this work is part of U.S. GDP, but it is not part of U.S. GNP. (It is part of Canada's GNP.) For most countries, including the United States, domestic residents are responsible for most domestic production, so GDP and GNP are quite close.
- **Net national product** (NNP) is the total income of a nation's residents (GNP) minus losses from depreciation. **Depreciation** is the wear and tear on the economy's stock of equipment and structures, such as trucks rusting and old computer models becoming obsolete. In the national income accounts prepared by the Department of Commerce, depreciation is called the "consumption of fixed capital."
- **National income** is the total income earned by a nation's residents in the production of goods and services. It is almost identical to net national product. These two measures differ because of the **statistical discrepancy** that arises from problems in data collection.
- **Personal income** is the income that households and noncorporate businesses receive. Unlike national income, it excludes **retained earnings**, the income that corporations earn but do not pay out to their owners. It also subtracts indirect business taxes (such as sales taxes), corporate income taxes, and contributions for social insurance (mostly Social Security taxes). In addition, personal income includes the interest income that households receive from their holdings of government debt and the income that households receive from government transfer programs, such as welfare and Social Security.
- **Disposable personal income** is the income that households and noncorporate businesses have left after satisfying all their obligations to the government. It equals personal income minus personal taxes and certain nontax payments (such as traffic tickets).

Although the various measures of income differ in detail, they almost always tell the same story about economic conditions. When GDP grows rapidly, these other measures of income tend to grow rapidly. And when GDP falls, these other measures tend to fall as well. As a result, for monitoring fluctuations in the overall economy, it does not matter much which measure of income we use. ■

QuickQuiz

3. If the price of a hot dog is $2 and the price of a hamburger is $4, then 30 hot dogs contribute as much to GDP as _____ hamburgers.
   a. 5
   b. 15
   c. 30
   d. 60
4. Angus the sheep farmer sells wool to Barnaby the knitter for $20. Barnaby makes two sweaters, each of which has a market price of $40. Collette buys one of them, while the other remains on the shelf of Barnaby's store to be sold later. What is GDP here?
   a. $40
   b. $60
   c. $80
   d. $100
5. Desirae, an American college student, moves to Japan for the summer to teach English. Her salary is included
   a. only in U.S. GDP.
   b. only in Japan's GDP.
   c. in both U.S. GDP and Japan's GDP.
   d. in neither U.S. GDP nor Japan's GDP.

Answers are at the end of the chapter.

24-3 The Components of GDP

Spending in an economy takes many forms. At any moment, the Lopez family may be having lunch at Burger King, Ford may be building a car factory, the U.S. Navy may be procuring a submarine, and British Airways may be buying an airplane from Boeing. GDP includes all of these forms of spending on domestically produced goods and services.

To understand how the economy is using its scarce resources, economists study the composition of GDP among the types of spending. To do this, GDP (denoted as Y) is divided into four components: consumption (C), investment (I), government purchases (G), and net exports (NX):

$$Y = C + I + G + NX.$$

This equation is an **identity**—an equation that must be true given the definitions of the variables in the equation. In this case, because each dollar of expenditure included in GDP is placed into one of the four components, the total of the four components must be equal to GDP. Let's look at these four components more closely.

24-3a Consumption

Consumption is spending by households on goods and services except for purchases of new housing. Goods include durable goods, such as automobiles and appliances, and nondurable goods, such as food and clothing. Services include intangible items, such as haircuts and medical care. By convention, household spending on education is also included in consumption of services (although one might argue that it would fit better in the next component).

consumption
spending by households on goods and services, with the exception of purchases of new housing

24-3b Investment

Investment is the purchase of goods (called **capital goods**) that will be used in the future to produce more goods and services. Investment is the sum of purchases of business capital, residential capital, and inventories. Business capital includes business structures (such as a factory or office building), equipment (such as a

investment
spending on business capital, residential capital, and inventories

worker's computer), and intellectual property products (such as the software that runs the computer). Residential capital includes the landlord's apartment building and a homeowner's personal residence. By convention, the purchase of a new house is the one type of household spending categorized as investment rather than consumption.

As already noted, inventories deserve special attention. When Apple produces a computer and adds it to its inventory instead of selling it, Apple is assumed to have "purchased" the computer for itself. That is, the national income accounts treat the computer as part of Apple's investment spending. When Apple later sells the computer out of inventory, the sale will subtract from Apple's inventory investment, offsetting the positive expenditure of the buyer. Inventories are treated this way because GDP aims to measure the value of the economy's production, and goods added to inventory are part of that period's production.

Notice that GDP accounting uses the word **investment** in a specialized way. In everyday conversation, this word might make you might think of financial investments, such as stocks, bonds, and mutual funds—topics that we will study later in this book. By contrast, because GDP measures expenditure on goods and services, here the word **investment** means purchases of goods (such as business capital, residential structures, and inventories) that will be used to produce other goods and services in the future.

24-3c Government Purchases

government purchases
spending on goods and services by local, state, and federal governments

Government purchases measure spending on goods and services by federal, state, and local governments. This component includes the salaries of government workers as well as expenditures on public works. The U.S. national income accounts call this spending category **government consumption expenditure and gross investment**, but this book uses the shorter term **government purchases**.

The meaning of government purchases requires some clarification. When the government pays the salary of a soldier or schoolteacher, that salary is included in government purchases. But when the government pays a Social Security benefit to an older person or an unemployment insurance benefit to a recently laid-off worker, the story is different: These are called **transfer payments** because they are not made in exchange for a currently produced good or service. Transfer payments alter household income but do not reflect the economy's production. (From a macroeconomic standpoint, they are like negative taxes.) Because GDP is intended to measure income from, and expenditure on, the production of goods and services, transfer payments are not counted as government purchases.

24-3d Net Exports

net exports
spending on domestically produced goods by foreigners (exports) minus spending on foreign goods by domestic residents (imports)

Net exports equal the foreign purchases of domestically produced goods (exports) minus the domestic purchases of foreign goods (imports). A domestic firm's sale to a buyer in another country, such as Boeing's sale of an airplane to British Airways, increases net exports.

The **net** in **net exports** refers to the fact that imports are subtracted from exports. This subtraction is made because other components of GDP include imports of goods and services. For example, suppose that a household buys a $50,000 car from Volvo, the Swedish carmaker. This transaction increases consumption by $50,000 because car purchases are part of consumer spending. It also reduces net exports by $50,000 because the car is an import. In other words, net exports include goods and services produced abroad (with a minus sign) because these goods and services are included in consumption, investment, and government purchases (with a plus sign).

Thus, when a domestic household, firm, or government buys a good or service from abroad, the purchase does not affect GDP because it reduces net exports by the same amount that it raises consumption, investment, or government purchases.

Case Study

The Components of U.S. GDP

Table 1 shows the composition of U.S. GDP in 2021. In this year, the GDP of the United States was about $23 trillion. Dividing this number by the 2021 U.S. population of 331 million yields GDP per person (sometimes called GDP per capita) and reveals that the income and expenditure of the average American in 2021 was $69,386.

Consumption made up 68 percent of GDP, or $47,528 per person. Investment was $12,396 per person. Government purchases were $12,226 per person. Net exports were –$2,764 per person. This number is negative because Americans spent more on foreign goods than foreigners spent on American goods.

These data come from the Bureau of Economic Analysis, the part of the U.S. Department of Commerce that produces the national income accounts. You can find more recent data on GDP on its website, http://www.bea.gov. ●

Table 1

GDP and Its Components

This table shows total GDP for the U.S. economy in 2021 and the breakdown of GDP among its four components. When reading this table, recall the identity $Y = C + I + G + NX$.

Source: U.S. Department of Commerce. Parts may not sum to totals due to rounding.

	Total (in billions of dollars)	Per Person (in dollars)	Percent of Total
Gross domestic product, Y	$22,994	$69,386	100%
Consumption, C	15,750	47,528	68
Investment, I	4,108	12,396	18
Government purchases, G	4,052	12,226	18
Net exports, NX	−916	−2,764	−4

QuickQuiz

6. Which of the following does NOT add to U.S. GDP?
   a. Boeing manufactures and sells a plane to Air France.
   b. General Motors builds a new auto factory in North Carolina.
   c. The city of New York pays a salary to a policeman.
   d. The federal government sends a Social Security check to your grandmother.
7. An American buys a pair of shoes made in Italy. How do the U.S. national income accounts treat the transaction?
   a. Net exports and GDP both rise.
   b. Net exports and GDP both fall.
   c. Net exports fall, while GDP does not change.
   d. Net exports do not change, while GDP rises.
8. Which is the largest component of GDP?
   a. consumption
   b. investment
   c. government purchases
   d. net exports

Answers are at the end of the chapter.

24-4 Real versus Nominal GDP

As we have seen, GDP measures the total spending on goods and services in all markets in the economy. If total spending rises from one year to the next, at least one of two things must be true: (1) the economy is producing a larger output of goods and services, or (2) goods and services are being sold at higher prices. When studying changes in the economy over time, economists want to separate these two effects. In particular, they want a measure of the total quantity of goods and services the economy is producing independent of changes in the prices of those goods and services.

To do this, economists use a measure called **real GDP**. Real GDP answers a hypothetical question: What would be the value of the goods and services produced this year if they were valued using the prices that prevailed in some specific year in the past? By evaluating current production using prices fixed at past levels, real GDP shows how the economy's overall production of goods and services changes over time.

To see more precisely how real GDP is constructed, consider an example.

24-4a A Numerical Example

Table 2 shows some data for an economy that produces only two goods: hot dogs and hamburgers. The table shows the prices and quantities produced of the two goods in the years 2022, 2023, and 2024.

Table 2

Real and Nominal GDP

This table shows how to calculate real GDP, nominal GDP, and the GDP deflator for a hypothetical economy that produces only hot dogs and hamburgers.

Prices and Quantities

Year	Price of Hot Dogs	Quantity of Hot Dogs	Price of Hamburgers	Quantity of Hamburgers
2022	$1	100	$2	50
2023	$2	150	$3	100
2024	$3	200	$4	150

Calculating Nominal GDP

2022	($1 per hot dog × 100 hot dogs) + ($2 per hamburger × 50 hamburgers) = $200
2023	($2 per hot dog × 150 hot dogs) + ($3 per hamburger × 100 hamburgers) = $600
2024	($3 per hot dog × 200 hot dogs) + ($4 per hamburger × 150 hamburgers) = $1,200

Calculating Real GDP (base year 2022)

2022	($1 per hot dog × 100 hot dogs) + ($2 per hamburger × 50 hamburgers) = $200
2023	($1 per hot dog × 150 hot dogs) + ($2 per hamburger × 100 hamburgers) = $350
2024	($1 per hot dog × 200 hot dogs) + ($2 per hamburger × 150 hamburgers) = $500

Calculating the GDP Deflator

2022	($200/$200) × 100 = 100
2023	($600/$350) × 100 = 171
2024	($1,200/$500) × 100 = 240

To compute total spending in this economy, we multiply the quantities of hot dogs and hamburgers by their prices. In the year 2022, 100 hot dogs are sold at a price of $1 per hot dog, so expenditure on hot dogs equals $100. In the same year, 50 hamburgers are sold for $2 per hamburger, so expenditure on hamburgers is also $100. Total expenditure in the economy—the sum of expenditure on hot dogs and hamburgers—is $200. This amount, the production of goods and services valued at current prices, is called **nominal GDP**.

nominal GDP
the production of goods and services valued at current prices

The table shows the calculation of nominal GDP for these three years. Total spending rises from $200 in 2022 to $600 in 2023 and then to $1,200 in 2024. Part of this increase is attributable to greater quantities of hot dogs and hamburgers, and part is attributable to higher prices.

To remove the effect of price changes and obtain a measure of the amount produced, we use **real GDP**, which is the production of goods and services valued at constant prices. To calculate real GDP, we first designate one year as a **base year**. We then use the prices of hot dogs and hamburgers in that year to compute the value of goods and services in all the years. In other words, the prices in the base year provide the basis for comparing quantities in different years.

real GDP
the production of goods and services valued at constant prices

Suppose that, in our example, 2022 is the designated base year. We can then use the prices in 2022 to compute the value of production in 2022, 2023, and 2024. Table 2 shows these calculations. To compute real GDP for 2022, we multiply the prices of hot dogs and hamburgers in 2022 (the base year) by the quantities of hot dogs and hamburgers produced in 2022. (For the base year, real GDP always equals nominal GDP.) To compute real GDP for 2023, we multiply the prices of hot dogs and hamburgers in 2022 (the base year) by the quantities of hot dogs and hamburgers produced in 2023. Similarly, to compute real GDP for 2024, we multiply the prices in 2022 by the quantities in 2024. We find that real GDP has risen from $200 in 2022 to $350 in 2023 and then to $500 in 2024. This increase must be attributable to an increase in the quantities produced because the prices are being held fixed at base-year levels.

To sum up: **Nominal GDP uses current prices to value the economy's production of goods and services. Real GDP uses constant base-year prices to value the economy's production.** Because price changes do not affect real GDP, changes in real GDP reflect only changes in the quantities produced. Therefore, real GDP can be interpreted as a measure of the overall economy's production of goods and services.

Our goal in computing GDP is to gauge how well the overall economy is performing. Because real GDP measures the economy's production of goods and services, it reflects the economy's ability to satisfy people's material needs and desires. Thus, real GDP is better than nominal GDP for assessing economic well-being. When economists talk about the economy's GDP, they usually mean real GDP rather than nominal GDP. And when they talk about growth in the economy, they measure that growth as the percentage change in real GDP from one period to another.

24-4b The GDP Deflator

To recap: Nominal GDP reflects both the quantities of goods and services the economy is producing and the prices of those goods and services. Real GDP reflects only the quantities produced because it holds prices constant at base-year levels. From these two statistics, we can compute a third, called the GDP deflator, which reflects only the prices of goods and services.

The **GDP deflator** is defined as follows:

GDP deflator
a measure of the price level calculated as the ratio of nominal GDP to real GDP times 100

$$\text{GDP deflator} = \frac{\text{Nominal GDP}}{\text{Real GDP}} \times 100.$$

Because nominal GDP and real GDP must be the same in the base year, the GDP deflator for the base year always equals 100. The GDP deflator for subsequent years measures the change in nominal GDP from the base year that isn't attributable to a change in real GDP.

The GDP deflator measures the current price level relative to the price level in the base year. To see why, consider a couple of simple examples. First, imagine that the quantities produced in the economy rise over time, but prices remain the same. In this case, both nominal and real GDP rise at the same rate, so the GDP deflator is constant. Now suppose that, instead, prices rise, but the quantities produced stay the same. In this second case, nominal GDP rises, but real GDP remains the same, so the GDP deflator rises. In both cases, the GDP deflator reflects what's happening to prices but not to quantities.

Return to our numerical example in Table 2. The GDP deflator is computed at the bottom. For the year 2022, nominal GDP is $200, and real GDP is $200, so the GDP deflator is 100. (Recall that the deflator is always 100 in the base year.) For the year 2023, nominal GDP is $600, and real GDP is $350, so the GDP deflator is 171.

Economists use the term **inflation** to describe a situation in which the economy's overall price level is rising. The **inflation rate** is the percentage change in some measure of the price level from one period to the next. Using the GDP deflator, the inflation rate between two consecutive years is computed as follows:

$$\text{Inflation rate in year 2} = \frac{\text{GDP deflator in year 2} - \text{GDP deflator in year 1}}{\text{GDP deflator in year 1}} \times 100.$$

Because the GDP deflator rose in year 2023 from 100 to 171, the inflation rate is $100 \times (171 - 100)/100$, or 71 percent. In 2024, the GDP deflator rose to 240 from 171 the previous year, so the inflation rate is $100 \times (240 - 171)/171$, or 40 percent.

The GDP deflator is one measure used to monitor the average level of prices and the rate of inflation. The GDP deflator gets its name because it takes inflation out of nominal GDP—that is, it "deflates" nominal GDP for the rise that is due to price increases. The next chapter examines another measure of the economy's price level, called the **consumer price index** and discusses how the two statistics differ.

Case Study

A Half Century of Real GDP

Real GDP conveys important information. Figure 2 shows quarterly data on real GDP for the U.S. economy since 1970.

One message from these data is that real GDP grows over time. The real GDP of the U.S. economy in 2021 was about four times its 1970 level. Put differently, the output of goods and services produced in the United States has grown, on average, about 3 percent per year. Because this continued growth in real GDP exceeds the rate of population growth, average Americans today enjoy greater economic prosperity than past generations.

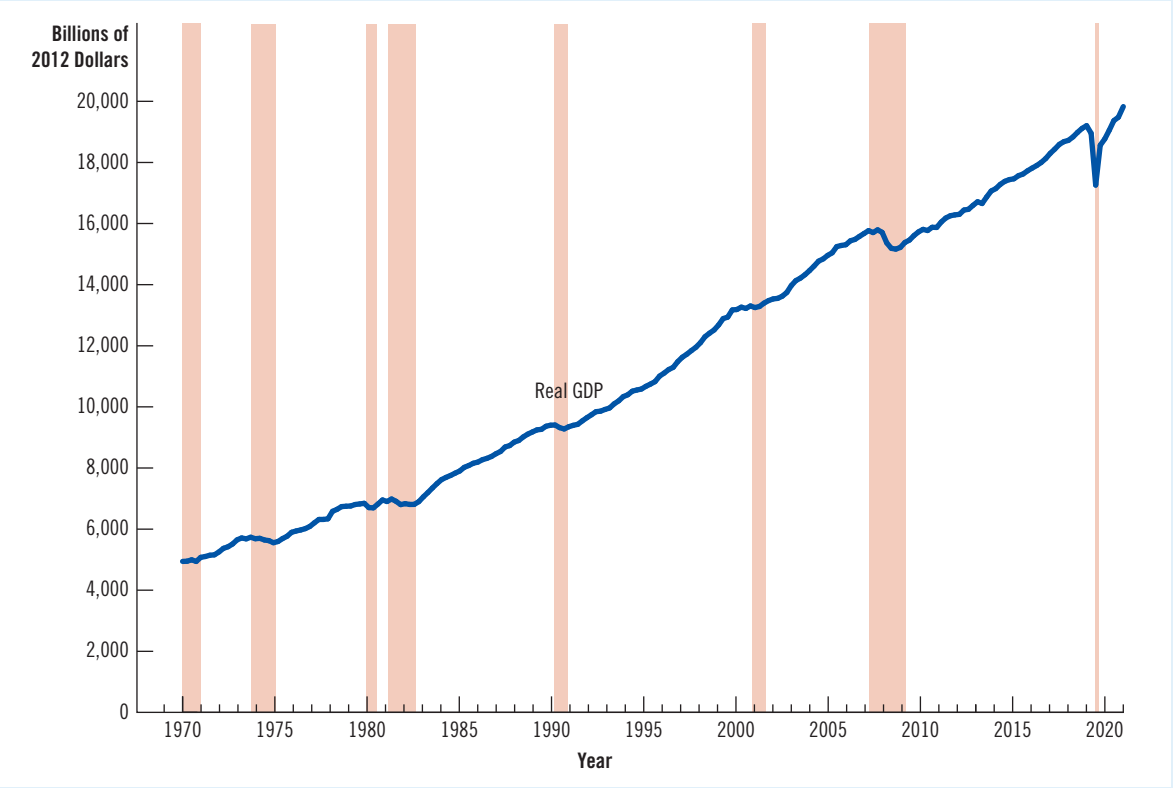
The GDP data also show that growth is not steady. The upward climb of real GDP is occasionally interrupted by periods when GDP declines, called **recessions**. Figure 2 marks recessions with shaded vertical bars. (There is no ironclad rule for when the official business cycle dating committee will declare that a recession has occurred, but an old rule of thumb is two consecutive quarters of falling real GDP. An exception is the pandemic recession of 2020, when there was an unusually large fall in GDP that lasted only one quarter.) Recessions are associated not only with lower incomes but also with other forms of distress: rising unemployment, falling profits, increased bankruptcies, and so on.

Figure 2

Real GDP in the United States

This figure shows quarterly data on real GDP for the U.S. economy since 1970. Recessions—periods of falling real GDP—are marked with the shaded vertical bars.

Source: U.S. Department of Commerce.



Much of macroeconomics aims to explain the long-run growth and short-run fluctuations in real GDP. As the coming chapters will show, different models are needed for these two purposes. Because the short-run fluctuations represent deviations from the long-run trend, we first examine the behavior of key macroeconomic variables, including real GDP, in the long run. Later chapters build on this analysis to explain short-run fluctuations. ●

QuickQuiz

9. An economy produces 10 cookies in year 1 at a price of $2 per cookie and 12 cookies in year 2 at a price of $3 per cookie. From year 1 to year 2, real GDP increases by
   a. 20 percent.
   b. 50 percent.
   c. 70 percent.
   d. 80 percent.

10. If all quantities produced rise by 5 percent and all prices fall by 5 percent, which of the following best describes what occurs?
    a. Real GDP rises by 5 percent, while nominal GDP falls by 5 percent.
    b. Real GDP rises by 5 percent, while nominal GDP is unchanged.
    c. Real GDP is unchanged, while nominal GDP rises by 5 percent.
    d. Real GDP is unchanged, while nominal GDP falls by 5 percent.

Answers are at the end of the chapter.

24-5 Is GDP a Good Measure of Economic Well-Being?

Earlier, this chapter noted that GDP is often considered the single best measure of society's economic well-being. Now that we know what GDP is, let's discuss the basis for this view and the limitations of GDP as a gauge of social welfare.

As we have seen, GDP measures both the economy's total income and its total expenditure on goods and services. GDP per person tells us the income and expenditure of the average person. Because most people would prefer to receive higher income and enjoy greater expenditure, GDP per person seems a natural measure of the economic well-being of the typical individual.

Yet some people dispute the validity of GDP as a measure of well-being. When Senator Robert Kennedy was running for president in 1968, he gave a moving critique of such statistics:

> [Gross domestic product] does not allow for the health of our children, the quality of their education, or the joy of their play. It does not include the beauty of our poetry or the strength of our marriages, the intelligence of our public debate or the integrity of our public officials. It measures neither our courage, nor our wisdom, nor our devotion to our country. It measures everything, in short, except that which makes life worthwhile, and it can tell us everything about America except why we are proud that we are Americans.

Much of what Robert Kennedy said is correct. Why, then, do we care about GDP?

The answer is that a larger GDP does, in fact, make it easier to lead better lives. GDP does not measure the health of our children, but nations with larger GDP can afford better healthcare for their children. GDP does not measure the quality of their education, but nations with larger GDP can afford better educational systems. GDP does not measure the beauty of our poetry, but nations with larger GDP can afford to teach more of their citizens to read and enjoy poetry. GDP does not take account of our intelligence, integrity, courage, wisdom, or devotion to country, but these virtues are easier to foster when people are less concerned about the material necessities of life. In short, GDP does not directly measure those things that make life worthwhile, but it does measure our ability to obtain many of the inputs for a worthwhile life.

GDP is not, however, a perfect measure of well-being. It omits some things that contribute to a good life, such as leisure. Suppose, for instance, that everyone suddenly started working every day of the week rather than taking weekends off to spend with family and friends or on personal pursuits. More goods and services would be produced, and GDP would rise. But despite the increase in GDP, we should not conclude that everyone would be better off. The loss from reduced leisure would offset the gain from producing and consuming a greater quantity of goods and services.

Because GDP uses market prices to value goods and services, it excludes the value of almost all activity that takes place outside markets. In particular, GDP omits the value of goods and services produced at home. When chefs prepare delicious meals at their restaurants, the value of those meals is part of GDP. But if chefs prepare the same meals for their own families, the value they have added to the raw ingredients is left out of GDP. Similarly, childcare provided in day care centers is part of GDP, whereas childcare provided by parents at home is not. Volunteer work also contributes to the well-being of those in society, but GDP does not reflect these contributions.

Another thing that GDP excludes is the quality of the environment. Imagine that the government eliminated all environmental regulations. Firms could then produce goods and services without considering the pollution they create, and GDP might rise. Yet well-being would most likely fall. The deterioration in the quality of the air and water would more than offset the gains from greater production.

GDP also says nothing about the distribution of income. Consider two societies, one in which 100 people have annual incomes of $50,000 and another in which 10 people earn $500,000 and 90 suffer with nothing at all. Both societies have GDP of $5 million and GDP per person of $50,000. Yet few people would consider the two situations equivalent. While GDP per person tells us what's happening to the average person, behind the average lies a large variety of individual experiences.

In the end, we can conclude that GDP is a good measure of economic well-being for many—but not all—purposes. It is important to keep in mind what GDP includes and what it leaves out.

Case Study

International Differences in GDP and the Quality of Life

One way to gauge the usefulness of GDP as a measure of economic well-being is to examine international data. Rich and poor countries have vastly different levels of GDP per person. If a large GDP leads to a higher standard of living, then GDP should be strongly correlated with various measures of the quality of life. And, in fact, they are.

Table 3 shows twelve large nations ranked in order of GDP per person. The table also shows life expectancy at birth, average years of schooling among adults, and an index of life satisfaction built by asking people how they feel about their lives on a scale of 0 to 10 (with 10 being the best). These data show a clear pattern. In rich countries, such as the United States and Germany, people have a life expectancy of about 80, acquire about 13 years of schooling, and rate their life satisfaction at about 7. In poor countries, such as Bangladesh and Nigeria, people typically die about 10 years earlier, have less than half as much schooling, and rate their life satisfaction about 2 points lower on the 10-point scale.

Data on other aspects of the quality of life tell a similar story. Countries with low GDP per person tend to have more infants with low birth weight, higher rates

Table 3

GDP and the Quality of Life

The table shows GDP per person and three other measures of the quality of life for twelve major countries.

Source: *Human Development Indices and Indicators: 2018 Statistical Update*, United Nations. Real GDP is for 2017, expressed in 2011 dollars. Average years of schooling is among adults 25 years and older.

Country	Real GDP per Person	Life Expectancy	Average Years of Schooling	Overall Life (0 to 10 scale)
United States	$54,941	80 years	13 years	7.0
Germany	46,136	81	14	7.1
Japan	38,986	84	13	5.9
Russia	24,233	71	12	5.6
Mexico	16,944	77	9	6.4
China	15,270	76	8	5.1
Brazil	13,755	76	8	6.3
Indonesia	10,846	69	8	5.1
India	6,353	69	6	4.0
Pakistan	5,311	67	5	5.8
Nigeria	5,231	54	6	5.3
Bangladesh	3,677	73	6	4.3

In the News

Sex, Drugs, and GDP

Some nations are debating what to include in their national income accounts.

No Sex, Please, We're French

By Zachary Karabell

The government of France has just made what on the face of it appears to be a nonannouncement announcement: It will not include illegal drugs and prostitution in its official calculation of the country's gross domestic product.

What made the announcement odd was that it never has included such activities, nor have most countries. Nor do most governments announce what they do not plan to do. ("The U.S. government has no intention of sending a man to Venus.") Yet the French decision comes in the wake of significant pressure from neighboring countries and from the European Union to integrate these activities into national accounts and economic output. That raises a host of questions: **Should** these activities be included, and if those are, why not others? And what exactly are we measuring—and why?

Few numbers shape our world today more than GDP. It has become the alpha and omega of national success, used by politicians and pundits as the primary gauge of national strength and treated as a numerical proxy for greatness or the lack thereof.

Yet GDP is only a statistic, replete with the limitations of all statistics. Created as an outgrowth of national accounts that were themselves only devised in the 1930s, GDP was never an all-inclusive measure, even as it is treated as such. Multiple areas of economic life were left out, including volunteer work and domestic work.

Now Eurostat, the official statistical agency of the European Union, is leading the drive to include a host of illegal activities in national calculations of GDP, most notably prostitution and illicit drugs. The argument, as a United Nations commission laid out in 2008, is fairly simple: Prostitution and illicit drugs are significant economic activities, and if they're not factored into economic statistics, then we're looking at an incomplete picture—which in turn will make it that much harder to craft smart policy. Additionally, different countries have different laws: In the Netherlands, for instance, prostitution is legal, as is marijuana. Those commercial transactions (or at least those that are recorded and taxed) are already part of Dutch GDP. Not including them in Italy's or Spain's GDPs can thus make it challenging to compare national numbers.

That is why Spain, Italy, Belgium, and the U.K. have in recent months moved to include

of infant mortality, higher rates of maternal mortality, and higher rates of child malnutrition. They also have lower rates of access to electricity, paved roads, and clean drinking water. In these countries, fewer school-age children are actually in school, those who are in school must learn with fewer teachers per student, and illiteracy among adults is more common. The citizens of these nations tend to have fewer televisions, fewer telephones, and fewer opportunities to access the Internet. International data leave no doubt that a nation's GDP per person is closely associated with its population's standard of living. ●

QuickQuiz

11. If Mr. Keating quits his job as a teacher to homeschool his own children, GDP
    a. stays the same because he is engaged in the same activity.
    b. rises because he now pays lower income taxes.
    c. falls because his market income decreases.
    d. could rise or fall, depending on the value of homeschooling.

12. GDP is an imperfect measure of well-being because it
    a. includes physical goods produced but not intangible services.
    b. excludes goods and services provided by the government.
    c. ignores the environmental degradation from economic activity.
    d. is not correlated with other measures of the quality of life.

Answers are at the end of the chapter.

illegal drugs and nonlicensed sex trade in their national accounts. The U.K. Office for National Statistics in particular approached its mandate with wonkish seriousness, publishing a 20-page précis of its methodology that explained how it would, say, calculate the dollar amount of prostitution (police records help) or deal with domestically produced drugs versus imported drugs. The result, which will be formally announced in September, will be an additional 10 billion pounds added to Great Britain's GDP.

France, however, has demurred. A nation with a clichéd reputation for a certain savoir faire when it comes to sex and other nocturnal activities has decided (or at least its bureaucrats have) that in spite of an EU directive, it will not calculate the effects of illegal activities that are often nonconsensual or nonvoluntary. That is clearly the case for some prostitution—one French minister stated that "street prostitution" is largely controlled by the Mafia—and the same could be reasonably said of the use of some hard drugs, given their addictive nature.

There is undeniably a strong moralistic component in the French decision. By averring that because they are not voluntary or consensual these exchanges should not be included in GDP, the French government is placing a moral vision of what society **should be** ahead of an economic vision of what society **is**. That in turn makes an already messy statistic far messier, and that serves no one's national interests. . . .

With all of GDP's limitations, adding a new moral dimension would only make the number that much less useful. After all, why stop at not including prostitution because it degrades women? Why not refuse to measure coal production because it degrades the environment? Why not leave out cigarette usage because it causes cancer? The list of possible exclusions on this basis is endless.

If GDP is our current best metric for national output, then at the very least it should attempt to include all measurable output. The usually moralistic United States has actually been including legal prostitution in Nevada and now marijuana sales and consumption in Colorado, California, and Washington without any strong objections based solely on the argument that these are commercial exchanges that constitute this fuzzy entity we call "the economy." . . .

Not measuring drugs and sex won't make them go away, but it will hobble efforts to understand the messy latticework of our economic lives, all in a futile attempt to excise what we do not like. ■

Source: *Slate*, June 20, 2014.

Questions to Discuss

1. Do you think illegal activities should be included in GDP? Why or why not?
2. Are there legal activities that you view as socially undesirable? If so, which ones? Do you think that GDP should include these activities? Why or why not?

24-6 Conclusion

This chapter discussed how economists measure the total income of a nation. Measurement is, of course, only a starting point. Much of macroeconomics is aimed at revealing the long-run and short-run determinants of a nation's gross domestic product. Why, for example, is GDP per person higher in the United States and Japan than in India and Nigeria? What can poor countries do to promote more rapid growth so they can catch up with rich countries? Why does GDP in all nations rise rapidly in some years and fall in others? How can policymakers reduce the severity of these fluctuations? These are the questions we will take up shortly.

At this point, it is important to acknowledge the significance of just measuring GDP. We all get some sense of how the economy is doing as we go about our lives. But to do their jobs well, economists and policymakers need concrete data. Quantifying the behavior of the economy with statistics such as GDP is the first step to developing a science of macroeconomics.

Chapter in a Nutshell

- Because every transaction has a buyer and a seller, the economy's total expenditure must equal its total income.
- Gross domestic product (GDP) measures an economy's total expenditure on newly produced goods and services and the total income earned from the production of these goods and services. More precisely, GDP is the market value of all final goods and services produced within a country in a given period.
- GDP consists of four components of expenditure: consumption, investment, government purchases, and net exports. Consumption includes spending on goods and services by households, with the exception of purchases of new housing. Investment includes spending on business capital, residential capital, and inventories. Government purchases include spending on goods and services by local, state, and federal governments. Net exports equal the value of goods and services produced domestically and sold abroad (exports) minus the value of goods and services produced abroad and sold domestically (imports).
- Nominal GDP uses current prices to value the economy's production of goods and services. Real GDP uses constant base-year prices to value this production. The GDP deflator—calculated from the ratio of nominal GDP to real GDP—measures the level of prices in the economy.
- GDP is a good measure of economic well-being because people usually prefer higher to lower incomes. But it is not a perfect measure of well-being. For example, GDP excludes the value of leisure and the value of a clean environment.

Key Concepts

microeconomics, p. 492
macroeconomics, p. 492
gross domestic product (GDP), p. 494
consumption, p. 497
investment, p. 497
government purchases, p. 498
net exports, p. 498
nominal GDP, p. 501
real GDP, p. 501
GDP deflator, p. 501

Questions for Review

1. Explain why an economy's income must equal its expenditure.
2. Which contributes more to GDP—the production of an economy car or the production of a luxury car? Why?
3. A farmer sells wheat to a baker for $2. The baker uses the wheat to make a loaf of bread, which a consumer buys for $3. What is the total contribution of these transactions to GDP?
4. Many years ago, Sophie paid $500 to put together a record collection. Today, she sold her albums at a garage sale for $100. How does this sale affect current GDP?
5. List the four components of GDP. Give an example of each.
6. Why do economists use real GDP rather than nominal GDP to gauge economic well-being?
7. In the year 2023, the economy produces 100 loaves of bread that sell for $2 each. In the year 2024, the economy produces 200 loaves of bread that sell for $3 each. Calculate nominal GDP, real GDP, and the GDP deflator for each year. (Use 2023 as the base year.) By what percentage does each of these three statistics rise from one year to the next?
8. Why is it desirable for a country to have a large GDP? Give an example of something that would raise GDP and yet be undesirable.

Problems and Applications

1. What components of GDP (if any) would each of the following transactions affect? Explain.
   a. Uncle Fester buys a new refrigerator from a domestic manufacturer.
   b. Aunt Dolly hires a local contractor to build her a new house.
   c. The Huang family buys an old Victorian house from the Ellis family.
   d. You pay a hairdresser for a haircut.
   e. Ford sells a Mustang from its inventory to the Martinez family.

f. Ford manufactures a Focus and sells it to Avis, the car rental company.
g. California hires workers to repave Highway 66.
h. The federal government sends your grandmother a Social Security check.
i. Your parents buy a bottle of French wine.
j. Honda expands its factory in Ohio.

2. Fill in the blanks:

Year	Real GDP (in 2000 dollars)	Nominal GDP (in current dollars)	GDP deflator (base year 2000)
1970	3,000	1,200	_____
1980	5,000	_____	60
1990	_____	6,000	100
2000	_____	8,000	_____
2010	_____	15,000	200
2020	10,000	_____	300
2030	20,000	50,000	_____

3. The government purchases component of GDP does not include spending on transfer payments such as Social Security. Thinking about the definition of GDP, explain why transfer payments are excluded.

4. As the chapter states, GDP does not include the value of used goods that are resold. Why would including such transactions make GDP a less informative measure of economic well-being?

5. Below are some data from the land of milk and honey.

Year	Price of Milk	Quantity of Milk	Price of Honey	Quantity of Honey
2023	$1	100 quarts	$2	50 quarts
2024	1	200	2	100
2025	2	200	4	100

a. Compute nominal GDP, real GDP, and the GDP deflator for each year, using 2023 as the base year.
b. Compute the percentage change in nominal GDP, real GDP, and the GDP deflator in 2024 and 2025 from the preceding year. For each year, identify the variable that does not change. Explain why your answer makes sense.
c. Did economic well-being increase more in 2024 or 2025? Explain.

6. Consider an economy that produces only chocolate bars. In year 1, the quantity produced is 3 bars, and the price is $4 per bar. In year 2, the quantity produced is 4 bars, and the price is $5 per bar. In year 3, the quantity produced is 5 bars, and the price is $6 per bar. Year 1 is the base year.
a. What is nominal GDP for each of these three years?
b. What is real GDP for each of these years?
c. What is the GDP deflator for each of these years?
d. What is the percentage growth rate of real GDP from year 2 to year 3?
e. What is the inflation rate as measured by the GDP deflator from year 2 to year 3?
f. In this one-good economy, how might you have answered parts (d) and (e) without first answering parts (b) and (c)?

7. Consider the following data on the U.S. economy:

Year	Nominal GDP (in billions of dollars)	GDP Deflator (base year 2012)
2020	21,141	113.6
2000	10,287	78.1

a. What was the growth rate of nominal GDP between 2000 and 2020? (Hint: The growth rate of a variable X over an N-year period is calculated as $100 \times [(X_{final}/X_{initial})^{1/N} - 1]$.)
b. What was the growth rate of the GDP deflator between 2000 and 2020?
c. What was real GDP in 2000 measured in 2012 prices?
d. What was real GDP in 2020 measured in 2012 prices?
e. What was the growth rate of real GDP between 2000 and 2020?
f. Was the growth rate of nominal GDP higher or lower than the growth rate of real GDP? Explain.

8. Revised estimates of U.S. GDP are usually released by the government near the end of each month. Find a news article that reports on the most recent release, or read the news release yourself at http://www.bea.gov, the website of the U.S. Bureau of Economic Analysis. Discuss the recent changes in real and nominal GDP and in the components of GDP.

9. A farmer grows wheat and sells the wheat to a miller for $100. The miller turns the wheat into flour and sells the flour to a baker for $150. The baker turns the flour into bread and sells the bread to consumers for $180. Consumers eat the bread.
a. What is GDP in this economy? Explain.
b. **Value added** is defined as the value of a producer's output minus the value of the intermediate goods that the producer buys to make the output. Assuming there are no intermediate goods beyond those described above, calculate the value added of each of the three producers.

c. What is the total value added of the three producers in this economy? How does it compare to the economy's GDP? Does this example suggest another way of calculating GDP?

10. Goods and services that are not sold in markets, such as food produced and consumed at home, are generally not included in GDP. How might this cause the numbers in the second column of Table 3 to be misleading in a comparison of the economic well-being of the United States and India? Explain.

11. The participation of women in the U.S. labor force has risen dramatically since 1970.
    a. How do you think this rise affected GDP?
    b. Now imagine a measure of well-being that includes time spent working in the home and taking leisure. How would the change in this measure of well-being compare to the change in GDP?
    c. Can you think of other aspects of well-being that are associated with the rise in women's labor-force participation? Would it be practical to construct a measure of well-being that includes these aspects?

12. One day, Barry the Barber, Inc., collects $400 for haircuts. Over the course of this day, his equipment depreciates in value by $50. Of the remaining $350, Barry sends $30 to the government in sales taxes, takes home $220 in wages, and retains $100 in his business to add new equipment in the future. From the $220 that Barry takes home, he pays $70 in income taxes. Based on this information, compute Barry's contribution to the following measures of income.
    a. gross domestic product
    b. net national product
    c. national income
    d. personal income
    e. disposable personal income

Quick**Quiz Answers**

1. **c** 2. **d** 3. **b** 4. **c** 5. **b** 6. **d** 7. **c** 8. **a** 9. **a** 10. **b** 11. **c** 12. **c**

Chapter

25

Measuring the Cost of Living

In 1931, as the U.S. economy was suffering through the Great Depression, the New York Yankees paid a baseball player named Babe Ruth a salary of $80,000. At the time, this pay was extraordinary, even among the stars of baseball. But Ruth was no ordinary player, and he wasn't short of self-confidence. According to one story, a reporter asked him whether he thought it was right that he made more than President Herbert Hoover, who had a salary of $75,000. Ruth replied, "I had a better year."

In 2021, the average salary of major league baseball players was about $4.2 million, and the Los Angeles Dodgers pitcher Trevor Bauer earned $38 million, making him the highest-paid player. At first, this might lead you to think that baseball has become vastly more lucrative over the past nine decades. But as everyone knows, the prices of goods and services have also risen. In 1931, a nickel would buy an ice-cream cone, and a quarter would buy a ticket at the local movie theater. Because prices were so much lower in Babe Ruth's day, it is not immediately clear whether Ruth enjoyed a higher or lower standard of living than today's players.

AZMANJAKA/E+/GETTY IMAGES; CARLOS FLORES/EYEEM/GETTY IMAGES

The preceding chapter looked at how economists use gross domestic product (GDP) to measure the quantity of goods and services the economy is producing. This chapter examines how economists measure the overall cost of living. To compare Babe Ruth's salary of $80,000 with today's salaries, we need to turn dollar figures into meaningful measures of purchasing power. That is exactly the job of a statistic called the **consumer price index**, or simply the CPI.

The CPI is used to monitor changes in the cost of living. When the CPI rises, the typical family has to spend more money to maintain the same standard of living. Economists use the term **inflation** to describe a situation in which the overall price level is rising and **deflation** for one in which the overall price level is falling. The **inflation rate** is the percentage change in the price level from the previous period. The preceding chapter showed how inflation can be measured using the GDP deflator. The inflation rate you are likely to hear on the nightly news, however, is based on the CPI, which better reflects the goods and services bought by consumers.

As we will see in the coming chapters, inflation is a closely watched aspect of macroeconomic performance and a key variable guiding macroeconomic policy. This chapter provides the background for that analysis by discussing how the CPI is constructed and how it can be used to compare dollar figures from different points in time.

25-1 The Consumer Price Index

consumer price index (CPI)
a measure of the overall cost of the goods and services bought by a typical consumer

The **consumer price index (CPI)** measures the overall cost of the goods and services bought by a typical consumer. Every month, the Bureau of Labor Statistics (BLS), which is part of the Department of Labor, computes and reports the CPI. This section examines how the CPI is calculated and what problems arise in its measurement. It also discusses how this index compares with the GDP deflator, another measure of the overall level of prices, which we examined in the preceding chapter.

25-1a How the CPI Is Calculated

When the BLS calculates the CPI and the inflation rate, it uses data on the prices of thousands of goods and services. To see how these statistics are constructed, consider a simple economy in which consumers buy only two goods: hot dogs and hamburgers. Table 1 shows the five steps that the BLS follows.

1. **Fix the basket.** Determine which prices are most important to consumers. If the typical consumer buys more hot dogs than hamburgers, then the price of hot dogs is more important than the price of hamburgers and, therefore, should be given greater weight in measuring the cost of living. The BLS sets these weights by surveying consumers to find the basket of goods and services bought by the typical consumer. In the example in the table, the basket includes 4 hot dogs and 2 hamburgers.
2. **Find the prices.** Find the prices of each of the goods and services in the basket at each point in time. The table shows the prices of hot dogs and hamburgers for three different years.
3. **Compute the basket's cost.** Use the data on prices to calculate the cost of the basket of goods and services at different times. The table shows this calculation for each of the three years. Notice that only the prices in the calculation change. By keeping the basket of goods the same (4 hot dogs and 2 hamburgers), we isolate the effects of price changes from the effects of any quantity changes that might be occurring at the same time.

Table 1

Calculating the Consumer Price Index and the Inflation Rate: An Example

This table shows how to calculate the CPI and the inflation rate for a hypothetical economy in which consumers buy only hot dogs and hamburgers.

Step 1: Survey Consumers to Determine a Fixed Basket of Goods

Basket = 4 hot dogs, 2 hamburgers

Step 2: Find the Price of Each Good in Each Year

Year	Price of Hot Dogs	Price of Hamburgers
2022	$1	$2
2023	2	3
2024	3	4

Step 3: Compute the Cost of the Basket of Goods in Each Year

2022	($1 per hot dog × 4 hot dogs) + ($2 per hamburger × 2 hamburgers) = $8 per basket
2023	($2 per hot dog × 4 hot dogs) + ($3 per hamburger × 2 hamburgers) = $14 per basket
2024	($3 per hot dog × 4 hot dogs) + ($4 per hamburger × 2 hamburgers) = $20 per basket

Step 4: Choose One Year as a Base Year (2022) and Compute the CPI in Each Year

2022	($8/$8) × 100 = 100
2023	($14/$8) × 100 = 175
2024	($20/$8) × 100 = 250

Step 5: Use the CPI to Compute the Inflation Rate from Previous Year

2023	(175 − 100)/100 × 100 = 75%
2024	(250 − 175)/175 × 100 = 43%

4. **Choose a base year and compute the index.** Designate one year as the base year, the benchmark against which other years are to be compared. (The choice of base year is arbitrary. The index is used to measure percentage changes in the cost of living, which are the same regardless of the choice of base year.) Once the base year is chosen, the index is calculated as follows:

$$\text{Consumer price index} = \frac{\text{Price of basket of goods and services in current year}}{\text{Price of basket in base year}} \times 100.$$

That is, the CPI in any given year is the price of the basket of goods and services in that year divided by the price of the basket in the base year, multiplied by 100.

In the example in Table 1, 2022 is the base year. In this year, the basket of hot dogs and hamburgers costs $8. Therefore, to calculate the CPI, the price of the basket in each year is divided by $8 and multiplied by 100. The CPI is 100 in 2022. (The index is always 100 in the base year.) The CPI is 175 in 2023. This means that the price of the basket in 2023 is 175 percent of its price in the base year. Put differently, a basket of goods that costs $100 in the base year costs $175 in 2023. Similarly, the CPI is 250 in 2024, indicating that the price level in 2024 is 250 percent of the price level in the base year.

inflation rate
the percentage change in the price index from the preceding period

5. **Compute the inflation rate.** Use the CPI to calculate the **inflation rate**, the percentage change in the price index from the preceding period. That is, the inflation rate between two consecutive years is computed as follows:

$$\text{Inflation rate in year 2} = \frac{\text{CPI in year 2} - \text{CPI in year 1}}{\text{CPI in year 1}} \times 100.$$

As shown at the bottom of Table 1, the inflation rate in our example is 75 percent in 2023 and 43 percent in 2024.

Although this example simplifies the real world by considering a basket of only two goods, it shows how the BLS computes the CPI and the inflation rate. The BLS collects and processes data on the prices of thousands of goods and services every month and, by following these five steps, determines how quickly the cost of living for the typical consumer is rising. When the BLS makes its monthly announcement of the CPI, you can usually hear the number on the evening news or see it in your newsfeed.

core CPI
a measure of the overall cost of consumer goods and services excluding food and energy

In addition to the CPI for the overall economy, the BLS calculates several other price indexes. It reports the index for some narrow categories of goods and services, such as food, clothing, and energy. It also calculates the CPI for all goods and services excluding food and energy, a statistic called the **core CPI**. Because food and energy

FYI What's in the CPI's Basket?

When constructing the CPI, the Bureau of Labor Statistics tries to include all the goods and services the typical consumer buys. Moreover, it tries to weight these goods and services according to how much consumers buy of each item.

Figure 1 shows the breakdown of consumer spending into the major categories of goods and services. By far, the largest category is housing, which makes up 42 percent of the typical consumer's budget. This includes the cost of shelter (33 percent), fuel and utilities (5 percent), and household furnishings and operation (5 percent). The next largest category, at 18 percent, is transportation, which includes spending on cars, gasoline, buses, subways, and so on. At 14 percent is the category of food and beverages; this includes food at home (8 percent), food away from home (5 percent), and alcoholic beverages (1 percent). Next are medical care at 8 percent, education and communication at 6 percent, and recreation at 5 percent. Apparel, which includes clothing, footwear, and jewelry, makes up 2 percent of the typical consumer's budget.

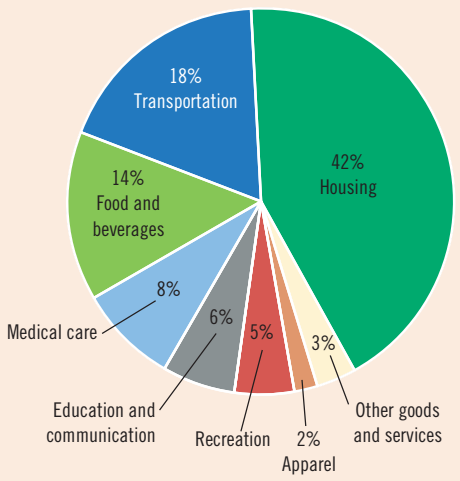
Finally, at 3 percent of spending is a category for other goods and services. This is a catchall for consumer purchases (such as cigarettes, haircuts, and funeral expenses) that do not naturally fit into the other categories. ■

Figure 1

The Typical Basket of Goods and Services

This figure shows how the typical consumer divides spending among various categories of goods and services. The Bureau of Labor Statistics calls each percentage the "relative importance" of the category.

Source: Bureau of Labor Statistics.



prices show substantial short-run volatility, the core CPI better reflects underlying inflation trends. Finally, the BLS also calculates the **producer price index (PPI)**, which measures the prices of the output of domestic producers. Formerly known as the wholesale price index, it dates back to a Senate report in 1893, making it one of the oldest prices indexes for the U.S. economy.

producer price index (PPI)
a measure of the cost of a basket of goods and services sold by domestic firms

25-1b Problems in Measuring the Cost of Living

The goal of the consumer price index is to measure changes in the cost of living. In other words, the CPI tries to gauge how much incomes must rise to maintain a constant standard of living. The CPI, however, is not a perfect measure. Three problems with the index are widely acknowledged but hard to solve.

The first problem is **substitution bias**. When prices change from one year to the next, they do not all change proportionately: Some prices rise more than others, and some even fall. Consumers respond by buying less of the goods with large price increases and by buying more of those with small price increases or price declines. That is, consumers substitute toward goods that become relatively less expensive. If a price index is computed assuming a fixed basket of goods, it ignores consumer substitution and overstates the increase in the cost of living from one year to the next.

Consider an example. Imagine that apples are initially cheaper than pears, so consumers buy more apples than pears. When setting the CPI's basket of goods, the BLS will observe consumers' buying habits and include more apples than pears. Now suppose that next year, pears are cheaper than apples. Consumers will likely buy more pears and fewer apples. But because the basket of goods is fixed, the CPI is calculated as if consumers continued to buy the now expensive apples in the same quantities as before. For this reason, the index measures a larger increase in the cost of living than consumers actually experience.

A second problem with the CPI arises from the **introduction of new goods**. When a new good is introduced, consumers have more variety from which to choose, which in turn reduces the cost of maintaining the same level of economic well-being. To see why, suppose you could choose between a $100 gift certificate at a large store that offered a wide array of goods and a $100 gift certificate at a small store with the same prices but a more limited selection. Which would you prefer? Most people would pick the store with greater variety. In essence, with more choices, each dollar is more valuable. The same is true for the economy: As new goods are introduced, consumers have more choices, and each dollar is worth more. But because the CPI is based on a fixed basket of goods and services, it doesn't reflect the increase in the value of the dollar that results from the introduction of new goods.

For example, in 2001, Apple introduced the iPod, a small music-playing device that was a precursor to the iPhone. Devices to play music were already available, but they weren't nearly as portable, powerful, and easy to use. The iPod was a new option that increased consumers' opportunities. For any given number of dollars, the introduction of the iPod made people better off; conversely, achieving the same level of well-being required fewer dollars. And, of course, this was true again with the subsequent introduction of the iPhone and other smartphones, which could do everything the iPod did and more. A perfect cost-of-living index would have reflected the decrease in the cost of living from the introduction of these gadgets. But because the CPI uses a fixed basket, it does not decrease when new goods are introduced. Eventually, the BLS revised the basket of goods to include the iPod and iPhone, and subsequently, the index reflected changes in their prices. But the reduction in the cost of living associated with the initial introduction of these devices never showed up in the index.

The third problem with the CPI is **unmeasured quality change**. If the quality of a good deteriorates from one year to the next while its price remains the same, you are getting a lesser good for the same amount of money, so the value of a dollar falls. Similarly, if the quality rises from one year to the next, the value of a dollar rises. The BLS does its best to account for quality change. When the quality of a good in the basket changes—for example, when a car model has more horsepower or gets better gas mileage from one year to the next—the Bureau adjusts the price of the good to account for the quality change. In doing so, it is trying to compute the price of a basket of goods of constant quality. Despite these efforts, changes in quality remain a problem because quality is hard to measure.

There is much debate about how severe these measurement problems are and what should be done about them. Studies put the upward bias in measured inflation at about 0.5 to 1.0 percent per year. The issue is important because many government programs use the CPI to adjust for changes in the overall level of prices. Recipients of Social Security, for instance, get annual increases in benefits that are tied to the CPI. Some economists have suggested modifying these programs to correct for the measurement problems by, for instance, reducing the magnitude of the automatic benefit increases. Others, however, say that would be a mistake because older people tend to spend more on healthcare, which often increases in price more rapidly than the standard CPI basket.

25-1c The GDP Deflator versus the Consumer Price Index

The preceding chapter examined another measure of the overall level of prices in the economy—the GDP deflator. The GDP deflator is the ratio of nominal GDP to real GDP. Because nominal GDP is current output valued at current prices and real GDP is current output valued at base-year prices, the GDP deflator reflects the current level of prices relative to the level of prices in the base year.

Economists and policymakers monitor both the GDP deflator and the CPI—along with several other indicators—to gauge how quickly prices are rising. Usually, these two statistics tell a similar story. Yet two important differences can cause them to diverge.

The first difference is that the GDP deflator reflects the prices of all goods and services **produced domestically**, while the CPI reflects the prices of all goods and services **bought by consumers**. For example, suppose that the price of a fighter jet produced by Boeing and sold to the Air Force rises. Even though the plane is part of GDP, it is not part of the basket of goods and services bought by consumers. This price increase shows up in the GDP deflator but not in the CPI.

As another example, suppose that Fiat raises the price of its cars. Because Fiats are made in Italy, the car is not part of U.S. GDP. But U.S. consumers buy Fiats, so the car is part of the CPI's basket of goods. A price increase in an imported consumption good, such as a Fiat, shows up in the CPI but not in the GDP deflator.

Historically, this first difference between the CPI and the GDP deflator has been particularly important when the price of oil changes. The United States has long produced some oil, but it consumed even more, resulting in substantial oil imports. As a result, oil and oil products such as gasoline and heating oil made up a larger share of consumer spending than of GDP. So when the price of oil increased, the CPI increased by much more than the GDP deflator. This phenomenon is less important today. Since 2008, U.S. oil production has increased substantially, reducing the nation's dependence on oil imports.

"The price may seem a little high, but you have to remember that's in today's dollars."

The second and subtler difference between the GDP deflator and the CPI concerns how various prices are weighted to yield a single number for the overall level of prices. The CPI compares the price of a **fixed** basket of goods and services with the price of the basket in the base year. Only occasionally does the BLS change what's in the basket. By contrast, the GDP deflator compares the price of **currently produced** goods and services with the price of those goods and services in the base year. For this reason, the group of goods and services used to compute the GDP deflator changes automatically over time. This difference is not important when all prices are changing proportionately. But if the prices of different goods and services are changing by varying amounts, the weighting of the various prices affects the calculation of the overall inflation rate.

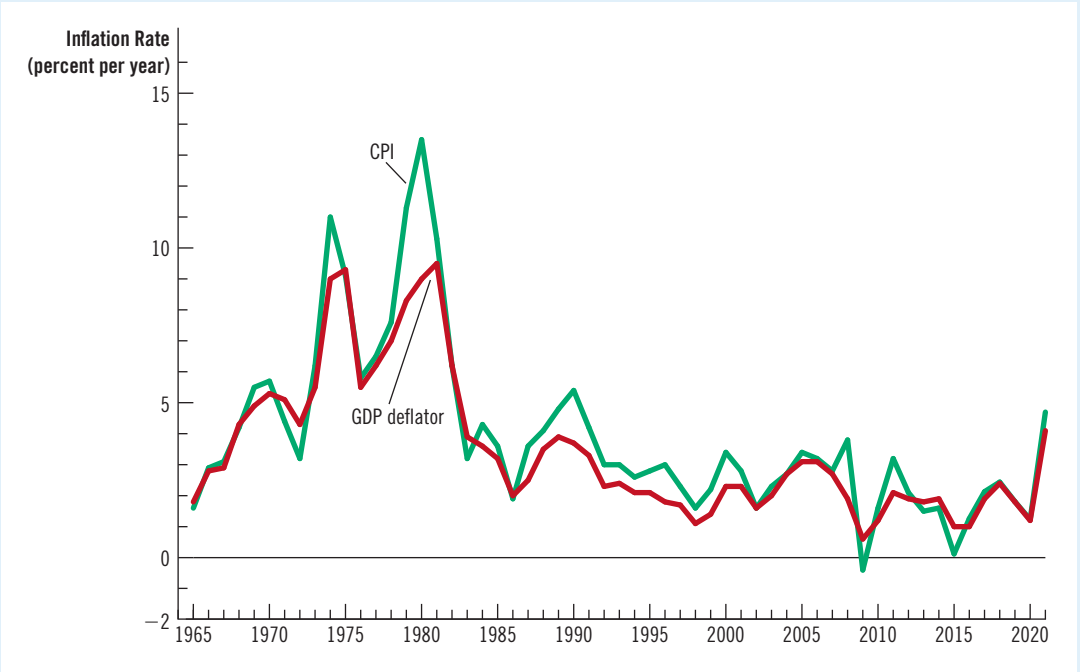
Figure 2 shows the inflation rate as measured by both the GDP deflator and the CPI for each year since 1965. You can see that sometimes the two measures diverge. When they do, it is possible to go behind these numbers and explain the divergence with the two differences we have discussed. For example, in 1979 and 1980, CPI inflation spiked up by more than inflation as measured by the GDP deflator largely because oil prices more than doubled during these two years. Conversely, in 2009 and 2015, CPI inflation fell well below inflation as gauged by the GDP deflator because of plummeting oil prices. Yet divergence between these two measures is the exception rather than the rule.

Figure 2

Two Measures of Inflation

This figure shows the inflation rate—the percentage change in the level of prices—as measured by the GDP deflator and the CPI using annual data since 1965. Notice that the two measures of inflation generally move together.

Source: U.S. Department of Labor; U.S. Department of Commerce.



QuickQuiz

1. The CPI measures approximately the same economic phenomenon as
   a. nominal GDP.
   b. real GDP.
   c. the GDP deflator.
   d. the unemployment rate.
2. The largest component in the basket of goods and services used to compute the CPI is
   a. food and beverages.
   b. housing.
   c. medical care.
   d. apparel.
3. If a Pennsylvania gun manufacturer raises the price of rifles it sells to the U.S. Army, its price hike will increase
   a. both the CPI and the GDP deflator.
   b. neither the CPI nor the GDP deflator.
   c. the CPI but not the GDP deflator.
   d. the GDP deflator but not the CPI.
4. Because consumers can sometimes substitute cheaper goods for those that have risen in price,
   a. the CPI overstates inflation.
   b. the CPI understates inflation.
   c. the GDP deflator overstates inflation.
   d. the GDP deflator understates inflation.

Answers are at the end of the chapter.

25-2 Correcting Economic Variables for the Effects of Inflation

Now that we know how price indexes are calculated, let's see how we might use such an index to compare a dollar figure from the past with a dollar figure in the present.

25-2a Dollar Figures from Different Times

Return to the issue of Babe Ruth's income. Was his salary of $80,000 in 1931 high or low compared with the salaries of today's players?

To answer this question, we need to know the level of prices in 1931 and the level of prices today. Part of the increase in baseball salaries compensates players for higher prices. To compare Ruth's salary with the salaries of today's players, we must inflate Ruth's salary to turn 1931 dollars into today's dollars.

The formula for turning dollar figures from year T into today's dollars is the following:

$$\text{Amount in today's dollars} = \text{Amount in year } T \text{ dollars} \times \frac{\text{Price level today}}{\text{Price level in year } T}.$$

A price index such as the CPI measures the price level and determines the size of the inflation correction.

Let's apply this formula to Ruth's salary. Government statistics show a CPI of 15.2 for 1931 and 271 for 2021. That means the overall level of prices has risen by a factor of 17.8 (calculated from 271/15.2). We can use these numbers to measure Ruth's salary in 2021 dollars as follows:

$$\begin{aligned}\text{Salary in 2021 dollars} &= \text{Salary in 1931 dollars} \times \frac{\text{Price level in 2021}}{\text{Price level in 1931}} \\ &= \$80{,}000 \times \frac{271}{15.2} \\ &= \$1{,}426{,}316\end{aligned}$$

We find that Babe Ruth's 1931 salary is equivalent to a salary today of over $1.4 million. That is a high income, but it is about a third of the average player's salary

FYI

Mr. Index Goes to Hollywood

What is the most popular movie of all time? The answer might surprise you.

Movie popularity is often gauged by box office receipts. By that measure, *Star Wars: The Force Awakens*, released in 2015, is the number-one movie of all time with domestic receipts of $937 million, followed by *Avengers: Endgame* ($853 million), *Avatar* ($761 million), and *Black Panther* ($700 million). But this ranking ignores an important fact: Prices, including those of movie tickets, have risen over time. Inflation gives an advantage to newer films.

LUCASFILM/PHOTO 12/ALAMY STOCK PHOTO

"May the force of inflation be with you."

When box office receipts are corrected for the effects of inflation, the story is very different. The number-one movie is now *Gone with the Wind* ($1,851 million in 2019 dollars), followed by the original *Star Wars* ($1,629 million) and *The Sound of Music* ($1,304 million). *Star Wars: The Force Awakens* ($989 million) falls to number 11.

Gone with the Wind was released in 1939, before everyone had televisions and when about 90 million Americans went to the cinema each week, compared with about 25 million recently. But the movies from that era don't appear in conventional popularity rankings because ticket prices were only a quarter. And indeed, in the ranking based on nominal box office receipts, *Gone with the Wind* does not make the top 100 films. Scarlett and Rhett fare a lot better once we correct for the effects of inflation. ■

today and less than 4 percent of what the star pitcher Trevor Bauer earns. Various forces, including overall economic growth and the increasing income shares earned by superstars, have substantially raised the living standards of the best athletes.

Let's also examine President Hoover's 1931 salary of $75,000. To translate that figure into 2021 dollars, we again multiply it by the ratio of the price levels in the two years. We find that Hoover's salary is equivalent to $75,000 × (271/15.2), or $1,337,171 in 2021 dollars. This is well above President Joe Biden's salary of $400,000. It seems that President Hoover did have a pretty good year after all.

Case Study

Regional Differences in the Cost of Living

When you graduate from college, you may have several job offers from which to choose. If the jobs are in different places, however, be careful when comparing how much they pay. The cost of living varies not only over time but also across locations. What seems like a larger paycheck might not turn out to be once you account for regional price differences.

The Bureau of Economic Analysis uses the data collected for the CPI to compare prices around the United States. It produces a useful statistic called **regional price parities**. Just as the CPI measures variation in the cost of living from year to year, regional price parities measure differences in the cost of living from state to state.

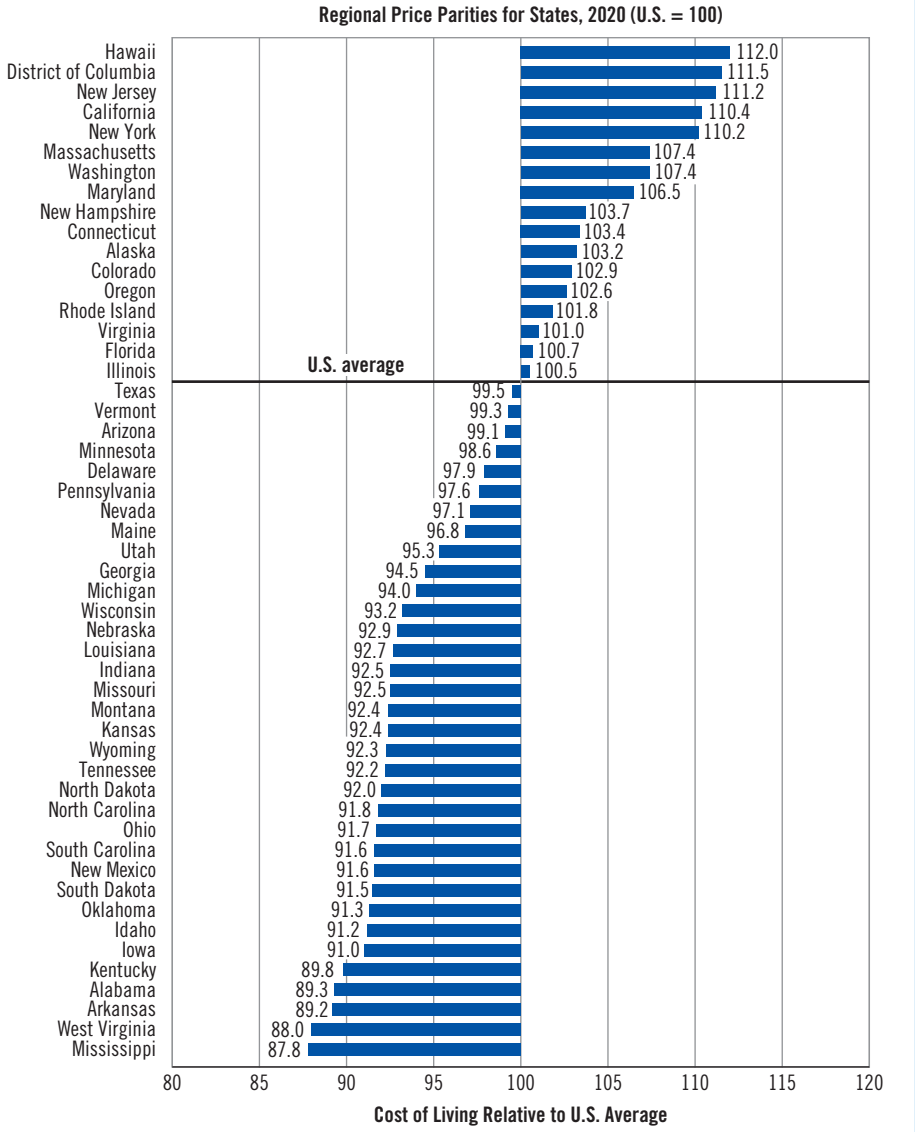
Figure 3 shows the regional price parities for 2020. For example, living in Hawaii costs 112.0 percent of what it costs to live in the typical place in the United States (that is, Hawaii is 12.0 percent more expensive than average). Living in Mississippi

Figure 3

Regional Variation in the Cost of Living

This figure shows how the costs of living in the 50 U.S. states and the District of Columbia compare to the U.S. average.

Source: U.S. Department of Commerce.



costs 87.8 percent of what it costs to live in the typical place (that is, Mississippi is 12.2 percent less expensive than average).

What accounts for these differences? The prices of goods, such as food and clothing, explain only a small part of these regional differences. Most goods are tradable: They can be easily transported from one state to another. Because of regional trade, large price disparities are unlikely to persist for long.

Services explain a larger part of regional differences. A haircut, for example, can cost more in one state than in another. If barbers were willing to move to places where the price of a haircut is high, or if customers were willing to fly across the country in search of cheap haircuts, then the prices of haircuts across regions

might well converge. But because transporting haircuts is so costly, large price disparities persist.

Housing services are particularly important for understanding regional differences in the cost of living. Such services represent a large share of a typical consumer's budget. And once built, a house or apartment building can't easily be moved, while the land on which it sits is completely immobile. As a result, differences in housing costs can be persistently large. For example, rents in Hawaii are about twice those in Mississippi.

Keep these facts in mind when it comes time to compare job offers. Look not only at the dollar salaries but also at the local prices of goods and services, especially housing. ●

25-2b Indexation

As we have seen, price indexes are used to correct for the effects of inflation when comparing dollar figures from different times. When some dollar amount is automatically corrected for changes in the price level by law or contract, the amount is said to be **indexed** for inflation.

indexation
the automatic correction by law or contract of a dollar amount for the effects of inflation

For example, some long-term contracts between firms and unions include partial or complete indexation of the wage to the CPI. Such a provision, called a **cost-of-living allowance** (or COLA), automatically raises the wage when the CPI rises.

Indexation is also a feature of many laws. Social Security benefits, for instance, are adjusted every year to compensate recipients for increases in prices. The brackets of the federal income tax—the income levels at which the tax rates change—are also indexed for inflation. There are, however, many ways in which the tax system is not indexed for inflation, even when perhaps it should be. We discuss these issues more fully later in this book.

25-2c Real and Nominal Interest Rates

Correcting economic variables for the effects of inflation is particularly important, and somewhat tricky, when we look at data on interest rates. The very concept of an interest rate necessarily involves comparing amounts of money at different points in time. When you deposit your savings in a bank account, you give the bank some money now, and the bank returns your deposit with interest in the future. Similarly, when you borrow from a bank, you get some money now, but you will have to repay the loan with interest in the future. In both cases, to fully understand the deal between you and the bank, it is crucial to acknowledge that future dollars could have a different value than today's dollars. In other words, you have to correct for the effects of inflation.

Consider an example. Suppose Sara Saver deposits $1,000 in a bank account that pays an annual interest rate of 10 percent. A year later, after Sara has accumulated $100 in interest, she withdraws her $1,100. Is Sara $100 richer than she was when she made the deposit a year earlier?

The answer depends on what we mean by "richer." Sara does have $100 more than she had before. In other words, the number of dollars in her possession has risen by 10 percent. But Sara does not care about the amount of money itself: She cares about what she can buy with it. If prices have risen while her money was in the bank, each dollar now buys less than it did a year ago. In this case, her purchasing power—the amount of goods and services she can buy—has not risen by 10 percent.

To keep things simple, suppose that Sara is a film buff and spends all her money on movie tickets. When Sara made her deposit, a ticket cost $10. Her deposit of $1,000 was equivalent to 100 tickets. A year later, after getting her 10 percent interest, she

has $1,100. How many tickets can she buy now? The answer depends on what has happened to the price of a ticket. Here are a few scenarios:

- Zero inflation: If the price of a ticket remains at $10, the amount she can buy has risen from 100 to 110 tickets. The 10 percent increase in the number of dollars means a 10 percent increase in her purchasing power.
- Six percent inflation: If the price of a ticket rises from $10 to $10.60, then the number of tickets she can buy has risen from 100 to approximately 104. Her purchasing power has increased by about 4 percent.
- Ten percent inflation: If the price of a ticket rises from $10 to $11, she can still buy only 100 tickets. Even though Sara's dollar wealth has risen, her purchasing power is the same as it was a year earlier.
- Twelve percent inflation: If the price of a ticket increases from $10 to $11.20, the number of tickets she can buy has fallen from 100 to approximately 98. Even with her greater number of dollars, her purchasing power has decreased by about 2 percent.

And if Sara were living in an economy with deflation—negative inflation or, more simply, falling prices—another possibility could arise:

- Two percent deflation: If the price of a ticket falls from $10 to $9.80, then the number of tickets she can buy rises from 100 to approximately 112. Her purchasing power increases by about 12 percent.

These examples show that the higher the rate of inflation, the smaller the increase in Sara's purchasing power. If the rate of inflation exceeds the rate of interest, her purchasing power actually falls. And if there is deflation, her purchasing power rises by more than the rate of interest.

To understand how much a person earns in a savings account, we need to consider both the interest rate and the change in prices. The interest rate that measures the change in dollar amounts is called the **nominal interest rate**, and the interest rate corrected for inflation is called the **real interest rate**. The nominal interest rate, the real interest rate, and inflation are related approximately as follows:

nominal interest rate
the interest rate as usually reported without a correction for the effects of inflation

real interest rate
the interest rate corrected for the effects of inflation

$$\text{Real interest rate} = \text{Nominal interest rate} - \text{Inflation rate}.$$

The real interest rate is the difference between the nominal interest rate and the rate of inflation. The nominal interest rate tells you how fast the number of dollars in your bank account rises over time, while the real interest rate tells you how fast the purchasing power of your bank account rises over time.

Interest Rates in the U.S. Economy

Figure 4 shows real and nominal interest rates in the U.S. economy since 1965. The nominal interest rate in this figure is the rate on three-month Treasury bills (although data on other interest rates would be similar). The real interest rate is computed by subtracting the rate of inflation from this nominal interest rate. Here the inflation rate is measured as the percentage change in the CPI.

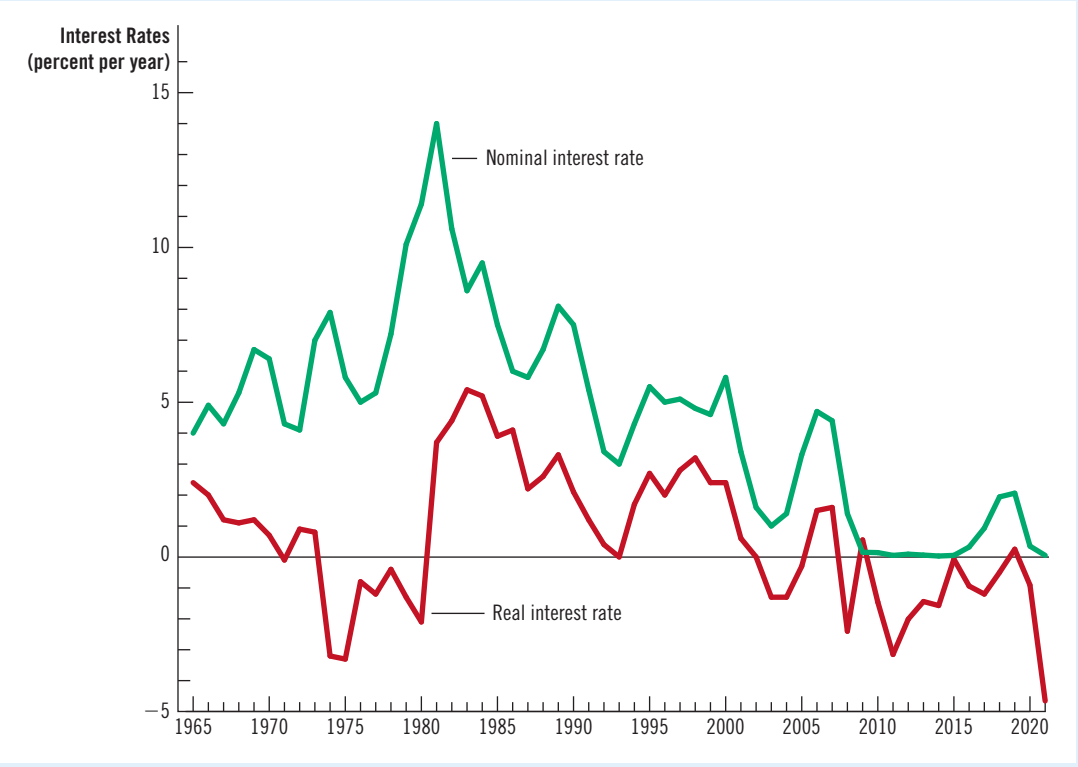
One feature of this figure is that the nominal interest rate usually exceeds the real interest rate. This reflects the fact that, while deflation has occurred occasionally,

Figure 4

Real and Nominal Interest Rates

This figure shows nominal and real interest rates using annual data since 1965. The nominal interest rate is the rate on a three-month Treasury bill. The real interest rate is the nominal interest rate minus the inflation rate as measured by the CPI. Notice that nominal and real interest rates often do not move together.

Source: U.S. Department of Labor; U.S. Department of Treasury.



the U.S. economy has experienced rising consumer prices in most years during this period. By contrast, if you look at data for the U.S. economy during the late 19th century or for the Japanese economy in some recent years, you will find significant periods of deflation. During deflation, the real interest rate exceeds the nominal interest rate.

The figure also shows that because inflation is variable, real and nominal interest rates do not always move together. For example, in the late 1970s, nominal interest rates were high. But because inflation was very high, real interest rates were low. Indeed, during much of the 1970s, real interest rates were negative: Inflation eroded people's savings more quickly than nominal interest payments increased them. By contrast, in the late 1990s, nominal interest rates were lower than they had been two decades earlier, but real interest rates were higher because inflation was much lower. Yet nominal and real interest rates often move in the same direction: During the coronavirus recession of 2020, for example, nominal interest rates fell to about zero, and real interest rates again turned negative. The coming chapters will examine the economic forces that determine both real and nominal interest rates. ●

QuickQuiz

5. If the CPI is 200 for the year 2010 and 300 today, then $600 in 2010 has the same purchasing power as _____ has today.
   a. $400
   b. $500
   c. $700
   d. $900

6. The main reason the cost of living varies across regions of the country is differences in the price of
   a. food.
   b. clothing.
   c. housing.
   d. medical care.

7. You deposit $2,000 in a savings account, and a year later, you have $2,100. Meanwhile, the CPI rises from 200 to 204. In this case, the nominal interest rate is _____ percent, and the real interest rate is _____ percent.
   a. 1; 5
   b. 3; 5
   c. 5; 1
   d. 5; 3

Answers are at the end of the chapter.

25-3 Conclusion

"A nickel ain't worth a dime anymore," the late, great baseball player Yogi Berra once observed. A nickel was never worth a dime, of course, but it's true that the real values behind the nickel, dime, and dollar have not been stable. Persistent increases in the overall level of prices have been the norm. Such inflation reduces the purchasing power of each unit of money. When comparing dollar figures from different times, it is important to keep in mind that a dollar today is not worth the same as a dollar 20 years ago or, most likely, 20 years from now.

This chapter has discussed how economists measure the overall level of prices and how they use price indexes to correct economic variables for the effects of inflation. Price indexes allow us to compare dollar figures from different points in time and, therefore, get a better sense of how the economy is changing.

The discussion of price indexes in this chapter, together with the preceding chapter's discussion of GDP, is the first step in the study of macroeconomics. Soon we will examine what determines a nation's GDP and the causes and effects of inflation. Having explained how economists measure macroeconomic quantities and prices in the past two chapters, we are now ready to develop the models that explain movements in these variables.

Here is our strategy in the upcoming chapters. First, we look at the long-run determinants of real GDP and related variables, such as saving, investment, real interest rates, and unemployment. Second, we look at the long-run determinants of the price level and related variables, such as the money supply, inflation, and nominal interest rates. Last of all, having seen how these variables are determined in the long run, we examine the more complex question of what causes short-run fluctuations in real GDP and the price level. In all of these chapters, the measurement issues we have just discussed will provide the foundation for the analysis.

Chapter in a Nutshell

- The consumer price index (CPI) shows the cost of a basket of goods and services relative to the cost of the same basket in the base year. The index is used to measure the overall level of prices in the economy. The percentage change in the CPI measures the inflation rate.
- The CPI is an imperfect measure of the cost of living for three reasons. First, it does not account for consumers' ability to substitute toward goods that become relatively cheaper over time. Second, it does not allow for increases in the purchasing power of the dollar that result from the introduction of new goods. Third, it is distorted by unmeasured changes in the quality of goods and services. Because of these measurement problems, the CPI overstates true inflation.
- Like the CPI, the GDP deflator measures the overall level of prices in the economy. The two price indexes usually move together, but there are important differences. Unlike the CPI, the GDP deflator reflects the prices of goods and services produced domestically rather than those bought by consumers. As a result, imported goods affect the CPI but not the GDP deflator. In addition, while the CPI uses a fixed basket of goods, the group of goods and services reflected in the GDP deflator automatically changes over time as the composition of GDP changes.
- Dollar figures from different times do not represent a valid comparison of purchasing power. To compare a dollar figure from the past with a dollar figure today, the older figure needs to be inflated using a price index.
- Various laws and private contracts use price indexes to correct for the effects of inflation. Tax laws, however, are only partially indexed for inflation.
- Correcting for inflation is especially important when looking at interest rates. The nominal interest rate—the interest rate usually reported—is the rate at which the number of dollars in a savings account increases over time. By contrast, the real interest rate is the rate at which the purchasing power of a savings account increases (or decreases) over time. The real interest rate equals the nominal interest rate minus the rate of inflation.

Key Concepts

consumer price index (CPI), p. 512
inflation rate, p. 514
core CPI, p. 514
producer price index (PPI), p. 515
indexation, p. 521
nominal interest rate, p. 522
real interest rate, p. 522

Questions for Review

1. Which do you think has a greater effect on the CPI: a 10 percent increase in the price of chicken or a 10 percent increase in the price of caviar? Why?
2. Describe the three problems that make the CPI an imperfect measure of the cost of living.
3. Does an increase in the price of imported French wine affect the CPI or the GDP deflator more? Why?
4. Over a long period of time, the price of a candy bar rose from $0.20 to $1.20. Over the same period, the CPI rose from 150 to 300. Adjusted for overall inflation, how much did the price of the candy bar change?
5. Explain the meanings of **nominal interest rate** and **real interest rate**. How are they related?

Problems and Applications

1. Suppose that the year you were born, someone bought $100 of goods and services for your baby shower. How much would you guess it would cost today to buy a similar amount of goods and services? Now find data on the CPI and compute the answer based on it. (You can find the BLS's inflation calculator here: http://www.bls.gov/data/inflation_calculator.htm.)
2. The residents of Vegopia spend all of their income on cauliflower, broccoli, and carrots. In 2023, they spend a total of $200 for 100 heads of cauliflower, $75 for 50 bunches of broccoli, and $50 for 500 carrots. In 2024, they spend a total of $225 for 75 heads of cauliflower, $120 for 80 bunches of broccoli, and $100 for 500 carrots.
   a. Calculate the price of one unit of each vegetable in each year.
   b. Using 2023 as the base year, calculate the CPI for each year.
   c. What is the inflation rate in 2024?
3. Suppose that people consume only three goods, as shown in this table:

	Tennis Balls	Golf Balls	Bottles of Gatorade
2023 price	$2	$4	$1
2023 quantity	100	100	200
2024 price	$2	$6	$2
2024 quantity	100	100	200

   a. What is the percentage change in the price of each of the three goods?
   b. Using a method similar to the CPI, compute the percentage change in the overall price level.
   c. If you were to learn that a bottle of Gatorade increased in size from 2023 to 2024, should that information affect your calculation of the inflation rate? If so, how?
   d. If you were to learn that Gatorade introduced new flavors in 2024, should that information affect your calculation of the inflation rate? If so, how?
4. Go to the website of the Bureau of Labor Statistics (http://www.bls.gov) and find data on the CPI. By how much has the index, including all items, risen over the past year? For which categories of spending have prices risen the most? The least? Have any categories experienced price declines? Can you explain any of these facts?
5. A small nation idolizes the TV show *The Voice*. All they produce and consume are karaoke machines and voice lessons, in the following amounts:

	Karaoke Machines		Voice Lessons	
	Quantity	Price	Quantity	Price
2023	10	$40	6	$50
2024	12	60	10	60

   a. Using a method similar to the CPI, compute the percentage change in the overall price level. Use 2023 as the base year and fix the basket at 5 karaoke machines and 3 voice lessons.
   b. Using a method similar to the GDP deflator, compute the percentage change in the overall price level. Again, use 2023 as the base year.
   c. Is the inflation rate in 2024 the same using the two methods? Explain why or why not.
6. Which of the problems in the construction of the CPI might be illustrated by each of the following situations? Explain.
   a. the invention of cell phones
   b. the introduction of airbags in cars
   c. increased personal computer purchases in response to a decline in their price
   d. more scoops of raisins in each package of Raisin Bran
   e. greater use of fuel-efficient cars after gasoline prices increase
7. A dozen eggs cost $0.88 in January 1980 and $1.47 in January 2021. The average hourly wage for production and nonsupervisory workers was $6.57 in January 1980 and $25.86 in January 2021.
   a. By what percentage did the price of eggs rise?
   b. By what percentage did the wage rise?
   c. In each year, how many minutes did a worker have to work to earn enough to buy a dozen eggs?
   d. Did workers' purchasing power in terms of eggs rise or fall?
8. The chapter explains that Social Security benefits are increased each year in proportion to the increase in the CPI, even though most economists believe that the CPI overstates actual inflation.
   a. If older people consume the same market basket as other people, does Social Security provide senior citizens with an improvement in their standard of living each year? Explain.
   b. In fact, older people consume more healthcare compared with younger people, and healthcare costs have risen faster than overall inflation. What would you do to determine whether older people are actually better off from year to year?

9. Suppose that a borrower and a lender agree on the nominal interest rate to be paid on a loan. Then inflation turns out to be higher than they both expected.
   a. Is the real interest rate on this loan higher or lower than expected?
   b. Does the lender gain or lose from this unexpectedly high inflation? Does the borrower gain or lose?
   c. Inflation during the 1970s was much higher than most people had expected when the decade began. How did this unexpectedly high inflation affect homeowners who obtained fixed-rate mortgages during the 1960s? How did it affect the banks that lent the money?

QuickQuiz Answers

1. **c** 2. **b** 3. **d** 4. **a** 5. **d** 6. **c** 7. **d**

Chapter

26

Production and Growth

When you travel around the world, you see tremendous variation in the standard of living. Adjusted for international differences in prices, the average income in a rich country, such as the United States, Japan, or Germany, is about 10 times the average income in a poor country, such as India, Nigeria, or Nicaragua. These income differences are reflected in many differences in the quality of life. People in richer countries typically have better nutrition, safer housing, better healthcare, and longer life expectancy, as well as more automobiles, more telephones, and more computers.

Within individual countries, there are also large changes in the standard of living over time. In the United States over the past century, average income as measured by real gross domestic product (GDP) per person has grown by about 2 percent per year. This rate of growth may seem small, but it implies that average income has roughly doubled every 35 years. That's why you are likely enjoying a higher standard of living than your grandparents did when they were your age.

ROMAN SAMBORSKYI/SHUTTERSTOCK.COM; ARNO VAN ENGELAND/EYEEM/GETTY IMAGES

Growth rates vary substantially from country to country. From 1990 to 2020, GDP per person in China grew at an average rate of 8.5 percent per year, resulting in a nearly twelve-fold increase in average income. This growth moved China from being one of the poorest countries in the world to being a middle-income country in roughly one generation. By contrast, over the same time span, income per person in Zimbabwe fell by a total of 24 percent, leaving the typical person in that nation mired in poverty.

What explains these diverse experiences? How can rich countries maintain their high standard of living? What policies can poor countries pursue to promote more rapid growth and join the developed world? These questions are among the most important in macroeconomics. As the economist Robert Lucas put it, "The consequences for human welfare in questions like these are simply staggering: Once one starts to think about them, it is hard to think about anything else."

The previous two chapters discussed how economists measure macroeconomic quantities and prices. We can now begin to study the forces that determine these variables. Recall that an economy's GDP measures both the total income earned in the economy and the total expenditure on the economy's output of goods and services. The level of real GDP is a gauge of economic prosperity, and the growth of real GDP is a gauge of economic progress. This chapter focuses on the long-run determinants of the level and growth of real GDP. Later, we will study the short-run fluctuations of real GDP around its long-run trend.

We proceed here in three steps. First, we examine international data on real GDP per person. These data give a sense of how much the level and growth of living standards vary around the world. Second, we discuss **productivity**—the amount of goods and services produced for each hour of work. Productivity is the main determinant of a nation's standard of living, so we carefully consider the numerous factors that influence productivity. Third, we consider the link between productivity and the policies that a nation pursues.

26-1 Economic Growth around the World

As a starting point for our study of long-run growth, let's look at the experiences of some of the world's economies. Table 1 shows data on real GDP per person for 13 countries. For each country, the data span more than a century of history. The first and second columns of the table present the countries and time periods. (The time periods differ somewhat from country to country because of differences in data availability.) The third and fourth columns show estimates of real GDP per person more than a century ago and for a recent year.

The data on real GDP per person show that living standards vary widely from country to country. Income per person in the United States, for instance, is now almost four times that in China and about ten times that in India. The poorest countries have average levels of income not seen in the developed world for many decades. The typical person in Bangladesh or Pakistan in 2020 had about the same real income as the typical American in 1870.

The last column of the table shows each country's growth rate. The growth rate measures how rapidly real income per person increased in the typical year. In the United States, for example, where real income per person was $4,668 in 1870 and $63,544 in 2020, the growth rate was 1.76 percent per year. This means that if real

Table 1

The Variety of Growth Experiences

Source: Robert J. Barro and Xavier Sala-i-Martin, *Economic Growth* (New York: McGraw-Hill, 1995), Tables 10.2 and 10.3; *World Bank* online data; and author's calculations. To account for international price differences, data are PPP-adjusted when available.

		Real GDP per Person (in 2020 dollars)		
Country	Period	At Beginning of Period	At End of Period	Growth Rate (per year)
China	1900–2020	$834	$17,312	2.56%
Japan	1890–2020	1,751	42,197	2.48
Brazil	1900–2020	907	14,836	2.36
Mexico	1900–2020	1,350	18,833	2.22
Indonesia	1900–2020	1,038	12,074	2.07
Germany	1870–2020	2,544	53,694	2.05
Canada	1870–2020	2,766	48,073	1.92
India	1900–2020	786	6,454	1.77
United States	1870–2020	4,668	63,544	1.76
Argentina	1900–2020	2,671	20,768	1.72
Bangladesh	1900–2020	726	5,083	1.64
Pakistan	1900–2020	859	4,877	1.46
United Kingdom	1870–2020	5,601	44,916	1.40

income per person were to increase by 1.76 percent for each of 150 years beginning at $4,668, it would end up at $63,544. Of course, income did not rise exactly 1.76 percent every year: Some years it rose by more, other years it rose by less, and in still other years it fell. The growth rate of 1.76 percent per year ignores short-run fluctuations around the long-run trend and represents an average rate of growth for real income per person over many years.

The countries in Table 1 are ordered by growth rate from the most to least rapid. High on the list are Brazil and China, which had been two of the poorest nations in the world and rose to middle-income status. Also high on the list is Japan, formerly a middle-income nation and now one of the richest.

Near the bottom of the list are Pakistan and Bangladesh, which were among the poorest nations at the end of the nineteenth century and remain so today. At the bottom of the list is the United Kingdom. In 1870, it was the richest country in the world, with an average income about 20 percent higher than that of the United States and about twice Canada's. Today, average income in the United Kingdom is 29 percent below that of the United States and 7 percent below Canada's.

These data show that the world's richest countries are not guaranteed to remain the richest and the world's poorest countries are not doomed to endless poverty. But what explains these changes over time? Why do some countries zoom ahead while others are left behind? These are the questions we take up next.

FYI

Are You Richer Than the Richest American?

BETTMANN/GETTY IMAGES

John D. Rockefeller

American Heritage magazine once published a list of the richest Americans of all time. The number 1 spot went to John D. Rockefeller, the oil entrepreneur who lived from 1839 to 1937. According to the magazine's calculations, his wealth would be the equivalent of about $250 billion today, which roughly equals the wealth of Elon Musk, the technology entrepreneur who is today's richest American.

Despite his great wealth, Rockefeller did not enjoy many of the conveniences that we now take for granted. He couldn't watch television, play video games, surf the Internet, or send e-mail. During the heat of summer, he couldn't cool his home with air-conditioning. For much of his life, he couldn't travel by car or plane, and he couldn't use a phone to call friends or family. He couldn't take advantage of many vaccines and medicines that doctors today routinely use to protect, prolong, and enhance life.

Now consider: How much money would someone have to pay you to give up, for the rest of your life, all the modern conveniences that Rockefeller lived without? Would you do it for $250 billion? Perhaps not. And if you wouldn't, is it fair to say that you are better off than John D. Rockefeller, whose wealth was legendary?

As the preceding chapter discussed, using standard price indexes to compare sums of money from different times doesn't fully reflect the introduction of new goods. As a result, the rate of inflation is overestimated. The flip side of this observation is that the rate of real economic growth is underestimated. Pondering Rockefeller's life shows how significant this problem might be. Because of tremendous technological advances, the average American today is arguably "richer" than the richest American a century ago, even if that reality is lost in standard economic statistics. ■

QuickQuiz

1. Over the past century, real GDP per person in the United States has grown about _____ percent per year, meaning it has roughly doubled every _____ years.
   a. 2; 14
   b. 2; 35
   c. 5; 14
   d. 5; 35
2. The world's rich countries, such as the United States and Germany, have income per person that is about _____ times income per person in the world's poor countries, such as Pakistan and India.
   a. 2
   b. 4
   c. 10
   d. 30
3. Over the past century, _____ has experienced particularly strong growth, and _____ has experienced particularly weak growth.
   a. China; the United Kingdom
   b. China; Canada
   c. the United Kingdom; Canada
   d. Canada; China

Answers are at the end of the chapter.

26-2 Productivity: Its Role and Determinants

Explaining why living standards vary so much from country to country and over time is, in one sense, easy. The answer can be summarized in a single word—**productivity**. But in another sense, the variation in living standards is deeply puzzling because many factors influence a nation's productivity.

26-2a Why Productivity Is So Important

Consider a simple model based loosely on Daniel Defoe's novel *Robinson Crusoe* about a sailor stranded on a desert island. Because Crusoe lives alone, he catches his own fish, grows his own vegetables, and makes his own clothes. Crusoe's activities—his production and consumption of fish, vegetables, and clothing—are a simple economy. By examining Crusoe's economy, we can learn some lessons that also apply to more complex and realistic economies.

What determines Crusoe's standard of living? In a word, **productivity**, the quantity of goods and services produced from each unit of labor. If Crusoe is good at catching fish, growing vegetables, and making clothes, he lives well—or at least as well as possible on that island. If he is bad at doing these things, he lives poorly. Because Crusoe can consume only what he produces, his living standard is tied to his productivity.

productivity
the quantity of goods and services produced from each unit of labor

In Crusoe's economy, it is easy to see that productivity is the key determinant of living standards and that growth in productivity is the key determinant of growth in living standards. The more fish Crusoe catches per hour, the more he can eat at dinner. If he finds a better place to catch fish, his productivity rises. This increase in productivity makes Crusoe better off: He can eat the extra fish, or he can spend less time fishing and devote more time to other things.

Productivity's key role in determining living standards is as true for nations as it is for stranded sailors. Recall that GDP measures two things at once about an economy: the total income earned by everyone and the total expenditure on the output of goods and services. GDP can measure these two things simultaneously because, for the overall economy, they must be equal. Put simply, an economy's income is the economy's output.

That means that a nation can enjoy a high standard of living only if it can produce a large quantity of goods and services. Americans live better than Nigerians because American workers are more productive than Nigerian workers. The Japanese have enjoyed more rapid growth in living standards than Argentineans because Japanese workers have experienced more rapid growth in productivity. Recall that one of the **Ten Principles of Economics** in Chapter 1 is that a country's standard of living depends on its ability to produce goods and services. That's just as true for today's large and complex economies as it is for Crusoe's small and simple one.

But seeing the link between living standards and productivity is only the first step. It leads naturally to the next question: Why are some economies so much better at producing goods and services than others?

26-2b How Productivity Is Determined

Productivity is uniquely important in determining Robinson Crusoe's standard of living, but many factors determine Crusoe's productivity. He will be better at catching fish, for instance, if he has more fishing poles, if he has been trained in the best fishing techniques, if his island has a plentiful fish supply, or if he invents a better fishing lure. Each of these determinants of Crusoe's productivity—which we can call **physical capital**, **human capital**, **natural resources**, and **technological knowledge**—has a counterpart in more realistic economies. Let's consider each factor in turn.

physical capital
the stock of equipment and structures that are used to produce goods and services

Physical Capital per Worker Workers are more productive if they have tools with which to work. The stock of equipment and structures used to produce goods and services is called **physical capital**, or just **capital**. For example, when woodworkers

make furniture, they use saws, lathes, and drill presses. More tools allow the woodworkers to produce their output more quickly and accurately: A worker with only basic hand tools can make less furniture each week than one with sophisticated and specialized equipment.

Recall that the inputs used to produce goods and services—labor, capital, and so on—are called the **factors of production**. An important feature of capital is that it is a **produced** factor of production. That is, capital is an input into the production process that in the past was an output from the production process. The woodworker uses a lathe to make the leg of a table. Earlier, the lathe itself was the output of a firm that manufactures lathes. The lathe manufacturer, in turn, used other equipment to make its product. Thus, capital is a factor of production used to produce all kinds of goods and services, including more capital.

human capital
the knowledge and skills that workers acquire through education, training, and experience

Human Capital per Worker A second determinant of productivity is **human capital**. This is the economist's term for the knowledge and skills that workers acquire through education, training, and experience. Human capital includes the skills accumulated in early childhood programs, grade school, high school, college, and on-the-job training for adults in the labor force.

Education, training, and experience are less tangible than lathes, bulldozers, and buildings, but human capital is like physical capital in many ways. Both raise a nation's ability to produce goods and services, and both are produced factors of production. Producing human capital requires inputs in the form of teachers, books, and student time. In a sense, students are "workers" who have the important job of producing the human capital they will use in future production.

natural resources
the inputs into the production of goods and services that are provided by nature, such as land, rivers, and mineral deposits

Natural Resources per Worker **Natural resources** are a third determinant of productivity. These are the land, rivers, mineral deposits, and other resources provided by nature and used as inputs into production. Natural resources take two forms: renewable and nonrenewable. A forest, if properly managed, is a renewable resource. When one tree is cut down, a seedling can be planted in its place to be harvested in the future. Oil is a nonrenewable resource. Because oil is produced by nature over many millions of years, there is a limited supply. Once the supply is depleted, it is impossible to create more.

Differences in natural resources are responsible for some of the differences in standards of living around the world. The historical success of the United States was driven in part by the large supply of land well suited for agriculture. Today, some countries in the Middle East, such as Kuwait and Saudi Arabia, are rich simply because they happen to be on top of some of the largest pools of oil in the world.

Important as natural resources may be, they are not necessary for an economy to be highly productive. Japan has few natural resources, yet it is one of the richest countries. International trade makes Japan's success possible. Japan imports oil and other natural resources and exports its manufactured goods. Countries rich in natural resources also benefit from the exchange.

technological knowledge
society's understanding of the best ways to produce goods and services

Technological Knowledge **Technological knowledge**, the understanding of the best ways to produce goods and services, is the fourth determinant of productivity. Two hundred years ago, farm technology required a high input of labor to feed an entire population, so most Americans worked on farms. Today, thanks to advances

in fertilizers, pesticides, agricultural machinery, and plant hybrids, a small fraction of the population can produce enough food to feed the entire country. This technological change freed up labor, which could then be used to produce other goods and services.

Technological knowledge takes many forms. Some technology is common knowledge—after one person uses it, everyone becomes aware of it. Once Henry Ford introduced assembly-line production, rival carmakers quickly followed suit. Other technology is proprietary, known only by the company that discovers it. Only the Coca-Cola Company, for instance, knows the recipe for making its soft drink. Still other technology is proprietary for a limited time. When a pharmaceutical company discovers a new drug, the patent system gives that company a temporary right to be its exclusive manufacturer. When the patent expires, however, other companies can make generic versions. All these forms of technological knowledge are important for the economy's production of goods and services.

It is worthwhile to distinguish between technological knowledge and human capital. Technological knowledge refers to society's understanding of how the world works. Human capital refers to the resources expended transmitting this understanding to the people who make use of it. To use an apt metaphor, technological knowledge is the quality of society's textbooks, while human capital is the amount of time that the population has spent reading them. Productivity depends on both.

The Production Function

Economists often use a **production function** to describe the relationship between the quantity of inputs used in production and the quantity of output from production. For example, suppose Y denotes the quantity of output, L the quantity of labor, K the quantity of physical capital, H the quantity of human capital, and N the quantity of natural resources. Then we might write

$$Y = AF(L, K, H, N),$$

where $F(\,)$ is a function that shows how the inputs are combined to produce output. A is a variable that reflects the available production technology. As technology improves, A rises, so the economy produces more output from any combination of inputs.

Many production functions have a property called **constant returns to scale**. If a production function has constant returns to scale, then doubling all inputs causes the amount of output to double as well. Mathematically, we write that a production function has constant returns to scale if, for any positive number x,

$$xY = AF(xL, xK, xH, xN).$$

A doubling of all inputs would be represented in this equation by $x = 2$. The right side shows the inputs doubling, and the left side shows output doubling.

Production functions with constant returns to scale have an interesting and useful implication. To see it, set $x = 1/L$ so the preceding equation becomes

$$Y/L = AF(1, K/L, H/L, N/L).$$

Notice that Y/L is output per worker, which is a measure of productivity. This equation says that labor productivity depends on the amounts of physical capital per worker (K/L), human capital per worker (H/L), and natural resources per worker (N/L) and on the state of technology, as represented by the variable A. The equation summarizes the four determinants of productivity just discussed. ■

Case Study

Are Natural Resources a Limit to Growth?

Today, the world's population is almost 8 billion, about four times what it was a century ago. Over the same period, the average standard of living has risen, too. A perennial debate concerns whether this growth in population and living standards can continue in the future.

Some commentators have argued that natural resources will eventually limit how much the world's economies can grow. The world has only fixed amounts of nonrenewable natural resources such as oil and minerals, and these are gradually depleted over time. As the supplies of these resources dwindle, economic growth will slow, and living standards might even be forced to decline. Or so the argument goes.

Despite the apparent appeal of this logic, most economists are less concerned about such limits than one might expect. They argue that technological progress often yields ways to avoid these limits. Compare the economy past and present and you will immediately see ways in which the use of natural resources has improved. Modern cars have better gas mileage, and some no longer use gas at all. New houses have better insulation and require less energy to heat and cool. More efficient oil rigs waste less oil in the process of extraction. Recycling allows some nonrenewable resources to be reused. The development of alternative sources of power allows the substitution of renewable for nonrenewable resources. The wind and the sun, for instance, are gradually replacing coal and oil in the generation of electricity.

In the middle of the twentieth century, some conservationists were concerned about the excessive use of tin and copper. At the time, these were crucial commodities: Tin was used to make many food containers, and copper was needed to make telephone wire. Some people advocated mandatory recycling and rationing of tin and copper so supplies would be available for future generations. Today, however, plastic has replaced tin as a material for many food containers, and phone calls often travel over fiber-optic cables, which are made from sand, or over airwaves, needing no cables at all. Technological progress has made once crucial natural resources less necessary.

But are these efforts enough to sustain economic growth? One way to answer this question is to look at the prices of natural resources. In a market economy, scarcity is reflected in market prices. If the world were running out of crucial natural resources, then the prices of those resources would be rising over time. But in fact, the opposite is more often true. Natural resource prices exhibit substantial short-run fluctuations, but over long spans, the prices of most natural resources (adjusted for overall inflation) are stable or falling. It appears that our ability to conserve these resources is growing more rapidly than their supplies are dwindling. Market prices give no reason to believe that natural resources are a limit to economic growth. ●

QuickQuiz

4. Increases in the amount of human capital in the economy tend to _____ real incomes because they increase the _____ of labor.
   a. increase; bargaining power
   b. increase; productivity
   c. decrease; bargaining power
   d. decrease; productivity

5. Most economists are _____ that natural resources will eventually limit economic growth. As evidence, they note that the prices of most natural resources, adjusted for overall inflation, have tended to _____ over time.
   a. concerned; rise
   b. concerned; fall
   c. not concerned; rise
   d. not concerned; fall

Answers are at the end of the chapter.

26-3 Economic Growth and Public Policy

A society's standard of living depends on its ability to produce goods and services, and its productivity depends on the amounts of physical capital, human capital, and natural resources available per worker, as well as on technological knowledge. Given these relationships, policymakers around the world face a central question: What can government policy do to raise productivity and living standards?

26-3a Saving and Investment

Because capital is a produced factor of production, a society can change the amount of capital it has. If the economy produces a large quantity of capital goods today, then tomorrow it will have a larger capital stock and a greater capacity to produce goods and services. One way to raise future productivity, then, is to devote more current resources to the production of capital.

One of the **Ten Principles of Economics** in Chapter 1 is that people face trade-offs. This is especially important when considering the accumulation of capital. Because resources are scarce, devoting more resources to producing capital requires devoting fewer resources to producing goods and services for current consumption. That is, for society to invest more in capital, it must consume less and save more of its current income. The growth that arises from capital accumulation is not a free lunch: It requires that society sacrifice consumption in the present to enjoy higher consumption in the future.

The next chapter examines in more detail how an economy's financial markets coordinate saving and investment. It also discusses how government policies influence the amount of saving and investment that take place. At this point, note that encouraging saving and investment is one way a government can promote growth and, in the long run, raise an economy's standard of living.

26-3b Diminishing Returns and the Catch-Up Effect

Suppose that a government pursues policies that raise a country's saving rate—the percentage of GDP devoted to saving rather than consumption. What happens? With greater saving, fewer resources are needed to make consumption goods, and more are available to make capital goods. As a result, the capital stock increases, leading to rising productivity and more rapid growth in GDP. But how long does this higher rate of growth last? Assuming that the saving rate remains at its new, higher level, does GDP growth stay high indefinitely or only for a period of time?

The traditional view of the production process is that capital is subject to **diminishing returns**: As the stock of capital rises, the extra output produced from an additional unit of capital falls. In other words, when workers already have a large quantity of capital to use in producing goods and services, giving them an additional unit of capital increases productivity only slightly. This phenomenon is illustrated in Figure 1, which shows how the amount of capital per worker determines the amount of output per worker, holding constant other determinants of output (such as natural resources and technological knowledge). Capital's diminishing returns is sometimes called the **diminishing marginal product of capital**.

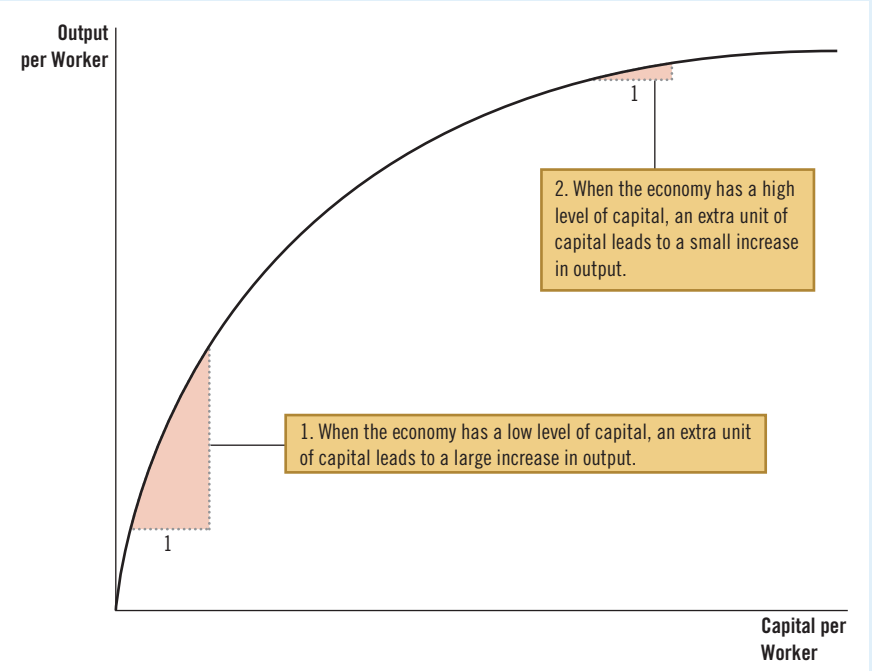
diminishing returns the property whereby the benefit from an extra unit of an input declines as the quantity of the input increases

Because of diminishing returns, an increase in the saving rate leads to higher growth only for a while. As the higher saving rate allows more capital to be accumulated, the benefits from additional capital become smaller over time, and growth slows. **In the long run, the higher saving rate leads to a higher level of productivity and income but not to higher growth in these variables.** Reaching this long run, however, can take quite a while. Studies of international data on economic growth show that increasing the saving rate can lead to substantially higher growth for a period of several decades.

Figure 1

Illustrating the Production Function

This figure shows how the amount of capital per worker influences the amount of output per worker. Other determinants of output, including human capital, natural resources, and technology, are held constant. The curve becomes flatter as the amount of capital increases because of diminishing returns to capital.



catch-up effect
the property whereby countries that start off poor tend to grow more rapidly than countries that start off rich

The property of diminishing returns to capital has another important implication: Other things being equal, it is easier for a country to grow quickly if it starts out relatively poor. This effect of initial conditions on subsequent growth is sometimes called the **catch-up effect**. In poor countries, workers lack even the most rudimentary tools and, as a result, have low productivity. In such conditions, small amounts of capital investment can substantially raise productivity. By contrast, workers in rich countries have high productivity partly because they already have large amounts of capital with which to work. In this case, additional capital investment has a relatively small effect on productivity. Studies of international data on economic growth confirm this catch-up effect: Controlling for other variables, such as the percentage of GDP devoted to investment, poor countries tend to grow at faster rates than rich countries.

This catch-up effect can help explain some otherwise puzzling facts. From 1960 to 1990, the United States and South Korea devoted a similar share of GDP to investment. Yet over this time, the United States experienced only moderate growth of about 2 percent, while South Korea grew at a rate of more than 6 percent. The catch-up effect explains the difference. In 1960, South Korea had GDP per person less than one-tenth of the U.S. level, in part because previous investment had been so low. With a small initial capital stock, South Korea realized greater benefits to capital accumulation and thus had a higher subsequent growth rate.

This catch-up effect shows up in other aspects of life. When a school gives an end-of-year award to the "Most Improved" student, that person usually began the year with relatively poor performance. Students who started out not studying find improvement easier than those who always worked hard. Note that it is good to be

the "Most Improved," given the starting point, but it is even better to be the "Best Student." Similarly, economic growth between 1960 and 1990 was much more rapid in South Korea than in the United States, but GDP per person was still higher in the United States.

26-3c Investment from Abroad

Policies aimed at increasing a country's saving rate can increase investment and long-term economic growth, but saving by domestic residents isn't the only way for a country to increase its capital stock. Investment by foreigners can do it as well.

Investment from abroad takes several forms. Ford Motor Company might build a car factory in Mexico. A capital investment that is owned and operated by a foreign entity is called **foreign direct investment**. Alternatively, an American might buy stock in a Mexican corporation (that is, buy a share in the ownership of the corporation), and the corporation can use the proceeds from the stock sale to build a new factory. An investment financed with foreign money but operated by domestic residents is called **foreign portfolio investment**. In both cases, Americans provide the resources necessary to increase the stock of capital in Mexico. That is, American saving is being used to finance Mexican investment.

When foreigners invest in a country, they do so because they expect to earn a return on their investment. Ford's car factory increases the Mexican capital stock and, therefore, increases Mexican productivity and Mexican GDP. Yet Ford takes some of this additional income back to the United States in the form of profit. Similarly, when an American investor buys Mexican stock, the investor has a right to a portion of the profit the Mexican corporation earns.

Investment from abroad, therefore, does not have the same effect on all measures of economic prosperity. Recall that a country's gross domestic product (GDP) is the income earned within the country by both residents and nonresidents, while a country's gross national product (GNP) is the income earned by residents of the country both at home and abroad. When Ford opens its car factory in Mexico, some of the income the factory generates accrues to people who do not live in Mexico. As a result, foreign investment in Mexico raises the income of Mexicans (measured by GNP) less than it raises the production in Mexico (measured by GDP).

Nonetheless, investment from abroad is one way for a country to grow. Even though some of the benefits flow back to the foreign owners, this investment does increase the economy's stock of capital, leading to higher productivity and higher wages. Moreover, investment from abroad is one way for poor countries to learn technologies developed and used in richer countries. For these reasons, many economists recommend that less developed economies encourage investment from abroad. Often, this means removing restrictions on foreign ownership of domestic capital.

An international organization that tries to promote the flow of capital to poor countries is the World Bank. It obtains funds from countries with abundant capital: Its largest shareholders are the United States, Japan, China, Germany, France, and the United Kingdom. The bank lends to less developed countries so that they can invest in roads, sewer systems, schools, and other types of capital. It also offers the countries advice about how the funds might best be used. The World Bank and its sister organization, the International Monetary Fund, were set up after World War II. One lesson from the war was that economic distress often leads to political turmoil, international tensions, and military conflict. Every country, therefore, has an interest in promoting economic prosperity around the world. The World Bank and the International Monetary Fund were established to achieve that common goal.

26-3d Education

Education—a form of investment in human capital—is at least as important as investment in physical capital for a country's long-run economic success. In the United States, each year of schooling has historically raised a person's wage by an average of about 10 percent. In less developed countries, where human capital is especially scarce, the gap between the wages of educated and uneducated workers is even larger. By providing good schools and encouraging people to take advantage of them, government policy can enhance the standard of living.

Investment in human capital, like investment in physical capital, has an opportunity cost. When students are in school, they forgo the wages they could have earned as members of the labor force. In less developed countries, children often drop out of school at an early age, even though the benefit of additional schooling is high, simply because their labor is needed to help support the family.

Some economists argue that human capital is particularly important for economic growth because it confers positive externalities. An **externality** is the effect of one person's actions on the well-being of a bystander. An educated person, for instance, might generate new ideas about how best to produce goods and services. If these ideas spread so everyone can use them, the ideas are an external benefit of education. In this case, the return from schooling for society is even greater than the return for the individual. This argument justifies the large subsidies to human-capital investment that we observe in the form of public education.

One problem facing some poor countries is a **brain drain**—the emigration of highly educated workers to rich countries, where these workers can earn more. If human capital has positive externalities, this brain drain makes those people left behind even poorer. Policymakers face a dilemma. The United States and other rich countries have the best systems of higher education, and it would seem natural for poor countries to send their best students abroad to earn higher degrees. Yet students who have spent time abroad may choose not to return home, and this brain drain will reduce the poor nation's stock of human capital even further.

26-3e Health and Nutrition

The term **human capital** usually refers to education, but it can describe another type of investment in people: expenditures that lead to a healthier population. The right investments in the health of the population provide one way for a nation to increase productivity and raise living standards.

The late economic historian Robert Fogel found that improved health from better nutrition has been a significant factor in long-run economic growth. Fogel estimated that in Great Britain in 1780, about one in five people were so malnourished they were incapable of manual labor. Among those who could work, insufficient caloric intake substantially reduced the work effort they could put forth. As nutrition improved, so did workers' productivity.

Fogel studied these trends in part by looking at measurements of people's height. Short stature can be an indicator of malnutrition, especially during gestation and the early years of life. Fogel found that as nations develop economically, people eat more, and the population gets taller. From 1775 to 1975, the average caloric intake in Great Britain rose by 26 percent, and the height of the average man rose by 3.6 inches. Similarly, during the spectacular economic growth in South Korea from 1962 to 1995, caloric consumption rose by 44 percent, and average male height rose by 2 inches. Of course, height is determined by a combination of genetics and environment. But because the population's genetic makeup is slow to change, such increases in average height are most likely due to changes in the environment—nutrition being the obvious explanation.

Fogel won the Nobel Prize in Economics in 1993 for his work in economic history, which includes not only his studies of nutrition but also his work on American slavery and the role of railroads in the development of the American economy. In his Nobel lecture, he surveyed the evidence on health and economic growth. Fogel concluded that "improved gross nutrition accounts for roughly 30 percent of the growth of per capita income in Britain between 1790 and 1980."

Today, inadequate nutrition is rare in developed nations such as Great Britain and the United States. (Obesity is a more common problem.) But for people in developing nations, poor health and nutrition remain grave problems. The United Nations estimates that about a quarter of the population in sub-Saharan Africa is undernourished.

The causal link between health and wealth runs in both directions. Poor countries are poor in part because their populations are not healthy, and their populations are not healthy in part because they are poor and cannot afford adequate healthcare and nutrition. Too often, this is a vicious circle. But there is a possibility of a virtuous circle, too: Policies that lead to more rapid growth would improve health outcomes, which in turn would further promote growth.

26-3f Property Rights and Political Stability

Another way policymakers can foster economic growth is by protecting property rights and promoting political stability. This issue goes to the very heart of how market economies work.

Production in market economies arises from the interactions of millions of individuals and firms. When you buy a car, you are buying the output of a car dealer, a car manufacturer, a steel company, an iron ore mining company, and so on. This division of production allows the economy's factors of production to be used effectively. To achieve this outcome, the economy has to coordinate transactions among these firms and between firms and consumers. Market economies achieve this coordination through market prices. That is, prices are the instrument with which the invisible hand of the marketplace brings supply and demand into balance in each of the thousands of markets that make up the economy.

Property rights, the ability of people to exercise authority over the resources they own, need to be respected for this process to work. A mining company will not make an effort to mine iron ore if it expects the ore to be stolen. A steel company will not pay for the ore if they expect the mining company to renege on its promise of delivery. For this reason, courts serve an important role in a market economy: They enforce property rights. Through the criminal justice system, the courts discourage theft and fraud. Through the civil justice system, the courts ensure that buyers and sellers live up to their contracts.

In less developed countries, where the systems of justice often do not work well, a lack of property rights can be a major problem. Contracts are hard to enforce, and fraud often goes unpunished. In some cases, a government not only fails to enforce property rights but actually infringes upon them. To do business in some places, firms are expected to bribe government officials. Such corruption degrades life in many ways. In economic terms, it impedes the coordinating power of markets. It also discourages domestic saving and investment from abroad.

One threat to property rights is political instability. Frequent revolutions and coups cause doubt about whether property rights will be respected in the future. If a revolutionary government might confiscate the capital of some businesses, as was often true after communist revolutions, domestic residents have less incentive to save, invest, and start new businesses. At the same time, foreigners have

less incentive to invest in the country. Even the threat of revolution can depress a nation's standard of living.

In short, economic prosperity depends in part on favorable political institutions. A country with an efficient court system, honest government officials, and a stable constitution will enjoy a higher standard of living than a country with a poor court system, corrupt officials, and frequent regime changes.

26-3g Free Trade

Some of the world's poorest countries have tried to achieve more rapid economic growth by pursuing **inward-oriented policies**, which aim to increase productivity and living standards by avoiding interaction with the rest of the world. Domestic firms often advance the infant-industry argument, claiming that they need protection from foreign competition to thrive and grow. Together with a general distrust of foreigners, this argument has at times led policymakers in less developed countries to impose tariffs and other trade restrictions.

Most economists today believe that poor countries are better off pursuing **outward-oriented policies** that integrate these countries into the world economy. International trade in goods and services can improve the economic well-being of a country's citizens. Trade is, in some ways, a type of technology. When a country exports wheat and imports textiles, the country benefits as if it had invented a technology for turning wheat into textiles. A country that eliminates trade restrictions will, therefore, experience the same kind of economic growth that would occur after a major technological advance.

The adverse impact of inward orientation becomes clear when one considers the small size of many less developed economies. The total GDP of Argentina, for instance, is roughly equal to that of Michigan. Imagine what would happen if Michigan residents were prohibited from trading with people living in other states and other countries. Without being able to take advantage of the gains from trade, Michigan would need to produce all the goods it consumes. It would also have to produce all its own capital goods, rather than importing state-of-the-art equipment from elsewhere. Living standards would fall immediately, and the problem would get worse over time. This is what happened when Argentina pursued inward-oriented policies throughout much of the 20th century. In contrast, countries that pursued outward-oriented policies, such as South Korea, Singapore, and Taiwan, enjoyed high rates of economic growth.

The amount that a nation trades with others is determined not only by government policy but also by geography. Countries with natural seaports find trade easier than those without this resource. It is not a coincidence that many of the world's major cities, such as New York, San Francisco, and Hong Kong, are located next to oceans. Similarly, because many landlocked countries find international trade more difficult, they tend to have lower levels of income than countries with easy access to waterways.

26-3h Research and Development

The primary reason that living standards are higher today than a century ago is that technological knowledge has advanced. The phone, the transistor, the computer, and the electric motor are among the innovations that have expanded the economy's capacity to produce goods and services.

Most technological advances come from private research by firms and individual inventors, but there is a public interest in promoting these efforts. To a large extent, knowledge is a **public good**: Once one person discovers an idea, it enters

society's pool of knowledge, and other people can freely use it. Just as government has a role in providing a public good such as national defense, it also has a role in encouraging the research and development of new technologies.

The U.S. government has long participated in the creation and dissemination of technological knowledge. A century ago, the government sponsored research about agricultural methods and advised farmers how best to use their land. More recently, the U.S. government, through the Air Force and NASA, has supported aerospace research; as a result, the United States is a leading maker of rockets and planes. The government continues to foster advances in knowledge with research grants from the National Science Foundation and the National Institutes of Health and with tax breaks for firms engaging in research and development.

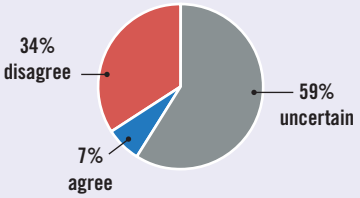
Another way in which government policy encourages research is through the patent system. When a person or firm creates an innovative product, such as a new drug, the inventor can apply for a patent. If the product is deemed original, the government awards the patent, which gives the inventor the exclusive right to make the product for a specified number of years. In essence, patents give inventors property rights over their inventions, turning their new ideas from public goods into private ones. This may slow the dissemination of innovations. But by allowing inventors to profit from their inventions—even if only temporarily—the patent system increases the incentive for individuals and firms to engage in research.

Ask the Experts

Innovation and Growth

"Future innovations worldwide will not be transformational enough to promote sustained per-capita economic growth rates in the United States and western Europe over the next century as high as those over the past 150 years."

What do economists say?



Source: IGM Economic Experts Panel, February 11, 2014.

26-3i Population Growth

Economists and other social scientists have long debated how the size and growth of the population affect a society. The most direct effect is on the size of the labor force: A large population means more workers are available to produce goods and services. The tremendous size of the Chinese population is one reason China is such an important player in the world economy.

At the same time, however, a large population means there are more people to consume those goods and services. So while a large population leads to a larger total output of goods and services, it need not translate into a higher standard of living for the typical citizen. Indeed, large and small nations are found at all levels of economic development.

Beyond these obvious effects of population size, population growth interacts with the other factors of production in ways that are more subtle and open to debate.

Stretching Natural Resources Thomas Robert Malthus (1766–1834), an English minister and early economic thinker, is famous for the book *An Essay on the Principle of Population as It Affects the Future Improvement of Society.* In it, he offered a chilling forecast. Malthus argued that an ever-increasing population would continually strain society's ability to provide for itself. As a result, mankind was doomed to live forever in poverty.

Malthus's logic was simple. He began by noting that "food is necessary to the existence of man" and that "the passion between the sexes is necessary and will remain nearly in its present state." He concluded that "the power of population is infinitely greater than the power in the earth to produce subsistence for

2002 ARPL/TOPFOTO/THE IMAGE WORKS

Thomas Robert Malthus

man." According to Malthus, the only check on population growth was "misery and vice." Attempts by charities or governments to alleviate poverty were counterproductive, he argued, because they allowed the poor to have more children, placing even greater strains on society's productive capabilities.

Malthus may have correctly described the world when he lived, but his dire forecast was far off the mark. The world population has increased about sixfold over the past two centuries, but living standards around the world have increased as well. Economic growth has made chronic hunger and malnutrition less common than in Malthus's day. Modern famines occur from time to time but more often result from income inequality or political instability than from inadequate food production.

Where did Malthus go wrong? As we discussed earlier in this chapter, growth in human ingenuity has offset the effects of a larger population. Pesticides, fertilizers, mechanized farm equipment, new crop varieties, and other technological advances that Malthus never imagined have allowed each farmer to feed ever greater numbers of people. Even with more mouths to feed, fewer farmers are necessary because each farmer is much more productive.

Diluting the Capital Stock While Malthus worried about the effects of population on the use of natural resources, some modern theories of economic growth emphasize its effects on capital accumulation. According to these theories, high population growth reduces GDP per worker because rapid growth in the number of workers forces the capital stock to be spread more thinly. In other words, when population growth is rapid, each worker is equipped with less capital. A smaller quantity of capital per worker leads to lower productivity and lower GDP per worker.

This problem is most apparent in the case of human capital. Countries with high population growth have many school-age children, placing a larger burden on the educational system. That's one reason that educational attainment tends to be low in countries with high population growth.

The differences in population growth around the world are large. In developed countries, such as the United States and those in Western Europe, the population has risen less than 1 percent per year in recent decades and is expected to rise even more slowly in the future. By contrast, in some poor African countries, the population has grown at about 3 percent per year. At this rate, the population doubles every 23 years. This rapid population growth makes it harder to provide workers with the tools and skills needed to achieve high levels of productivity.

Rapid population growth is not the main reason that less developed countries are poor, but some analysts believe that reducing the rate of population growth would help these countries raise their standards of living. In some countries, this goal is accomplished directly with laws that regulate the number of children families may have. For example, from 1980 to 2015, China allowed only one child per family; couples who violated this rule were subject to substantial fines. In less authoritarian countries, the goal of reduced population growth is accomplished less directly by increasing awareness of birth control techniques.

Another way in which a country can influence population growth is to apply one of the **Ten Principles of Economics**: People respond to incentives. Bearing a child, like any decision, has an opportunity cost. When the cost rises, people choose to have smaller families. In particular, women with good educations and employment prospects tend to want fewer children than those with fewer opportunities outside the home. Policies that foster the equal treatment of women are one way

for less developed economies to reduce their rates of population growth and raise their standards of living.

Promoting Technological Progress Rapid population growth may depress economic prosperity by reducing the amount of capital each worker has, but it may also have an upside. Some economists have suggested that world population growth has been an engine of technological progress. The mechanism is simple: If there are more people, then there are more scientists, inventors, and engineers helping to advance technology, benefiting everyone.

The economist Michael Kremer provided some support for this hypothesis in a 1993 article titled "Population Growth and Technological Change: One Million B.C. to 1990." Kremer began by noting that over the broad span of human history, world growth rates have increased with the world population. For example, world growth was more rapid when the population was 1 billion (around the year 1800) than when the population was only 100 million (around 500 B.C.). This is consistent with the hypothesis that a larger population induces more technological progress.

Kremer's second piece of evidence comes from comparing regions of the world. The melting of the polar icecaps at the end of the Ice Age around 10,000 B.C. flooded the land bridges and separated the world into distinct regions that could not communicate with one another for thousands of years. If technological progress is more rapid when there are more people to discover things, then the more populous regions should have experienced more rapid growth.

According to Kremer, that is what happened. The most successful region of the world in 1500 (when European navigators reestablished contact) comprised the "Old World" civilizations of the large Eurasia-Africa region. Next in technological development were the Aztec and Mayan civilizations in the Americas, followed by the hunter-gatherers of Australia, and then the primitive people of Tasmania, who lacked even fire-making and most stone and bone tools.

The smallest isolated region was Flinders Island, a tiny island between Tasmania and Australia. With the smallest population, Flinders Island had the fewest opportunities for technological advance and seemed to have regressed. Around 3000 B.C., human society on Flinders Island died out completely.

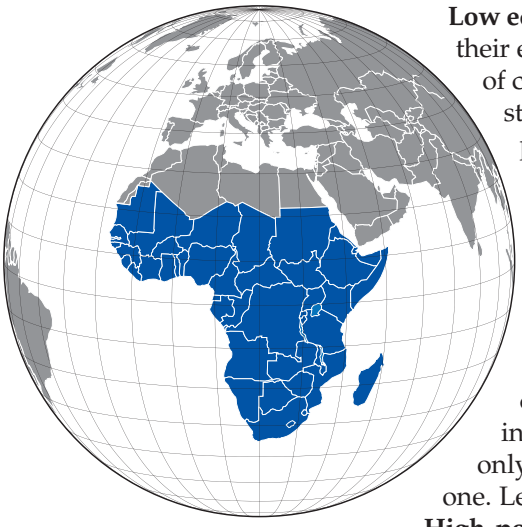
Kremer concluded from this evidence that larger populations increase the potential for technological advance.

Why Is So Much of Africa Poor?

Many of the poorest people on the planet live in sub-Saharan Africa. In 2020, GDP per person there was only $3,821, just 22 percent of the world average. Extreme poverty is endemic in the region: 40 percent of its population lives on less than $1.90 per day, compared with 9 percent of the population worldwide.

There is no easy explanation for this low level of economic development. Many interrelated forces are at work, and sometimes it is hard to distinguish the causes of poverty from the effects. But here are some clues.

Low capital investment. Because sub-Saharan Africa has low levels of income and capital per worker, one might expect the returns to capital to be high, making the region an attractive place to invest for both domestic savers and investors abroad. But, in fact, as a percentage of GDP, capital formation in sub-Saharan Africa is several percentage points below the world average. The low level of investment may be driven by some of the following factors.



NIKKULCH/SHUTTERSTOCK.COM

Sub-Saharan Africa is the poorest region of the world.

Low educational attainment. For children living in the region, 31 percent stop their education before completing primary school, compared with 10 percent of children worldwide. And their quality of schooling is lower as well: The student-teacher ratio in primary schools is 37 in sub-Saharan Africa, compared with a world average of 23. Only 65 percent of adults are literate, compared with 86 percent worldwide. Less educated workers are less productive.

Poor health. Among one-year-olds in sub-Saharan Africa, 27 percent have not been immunized for DPT (diphtheria, pertussis, and tetanus), and 30 percent have not been immunized for measles—in both cases, about twice the world average. Among children under age 5, 33 percent are malnourished enough to have stunted growth, compared with 22 percent worldwide. Among adults, 1.6 percent are infected with HIV, four times the world average. These statistics not only reflect extreme personal tragedy but also help explain an economic one. Less healthy workers are less productive.

High population growth. The population in sub-Saharan Africa has recently grown about 2.7 percent per year, so it doubles every 26 years. By contrast, the world population has grown 1.1 percent per year, doubling every 64 years. Rapid population growth makes it hard to equip workers with the physical and human capital needed to achieve high productivity.

Geographic disadvantages. More than 25 percent of the people in sub-Saharan Africa live in a landlocked nation, such as Ethiopia, Uganda, Chad, Niger, and Mali, compared with 7 percent of the world population. Landlocked nations tend to be poor. Without easy access to the oceans for purposes of transport, it is hard for them to take advantage of the gains from trade.

Restricted freedom. Social scientists have developed indexes to gauge the degree of freedom available to a nation's citizens. These indexes measure characteristics such as the reliability of the justice system, personal security and safety, freedom of expression, the right to engage in international trade, and so on. Nations in sub-Saharan Africa tend to rank low on these measures, as do those in South Asia, Eastern Europe, and the Middle East. The freest nations tend to be those in Western Europe, Northern Europe, and North America. (Other regions of the world, such as South America, fall between these extreme cases.) These freedom indexes are positively correlated with economic prosperity: Greater freedom is generally associated with higher incomes, perhaps because certain restrictions impede the invisible hand's ability to allocate resources efficiently.

Rampant corruption. The governments in many African nations exhibit high levels of corruption. According to Transparency International, a nonprofit organization that monitors corruption, Somalia and South Sudan were the most corrupt countries in the world in 2020, and sub-Saharan Africa overall was the most corrupt region. (The least corrupt countries were New Zealand and Denmark.) High levels of corruption discourage domestic residents from saving and investing and deter investment from abroad.

The legacy of colonization. The economists Daron Acemoglu and James Robinson attribute the low level of economic development in much of Africa to institutions originally put in place by European colonizers. In the 17th and 18th centuries, Europeans preferred to settle in colonies with moderate climates, such as the United States, Canada, and New Zealand. Because they planned to stay there, Europeans brought **inclusive institutions** like those in their homelands. Inclusive institutions spread political power widely, respect

property rights and the rule of law, and thereby foster economic prosperity. In places with tropical climates, including much of Africa, the colonizers had little interest in permanent settlement. So they established **extractive institutions**, such as authoritarian governments, designed to exploit the region's population and natural resources. Even after the colonizers left, the extractive institutions remained and were taken over by new ruling elites, impeding economic development.

The aftermath of the slave trade. The economist Nathan Nunn has proposed that the low level of African development is partly the result of the international slave trade, which began around 1400 and lasted about 500 years. During this period, about 20 million Africans were enslaved through warfare, kidnappings, and betrayal by friends and relatives. Nunn provides evidence for his hypothesis by showing that the poorest areas of Africa today tend to be those from which the largest number of slaves were taken (even though these areas were not originally the poorest). To explain this finding, he suggests that the slave trade impeded economic development by undermining political and legal institutions and leaving a culture of distrust.

None of these causes suggests an easy solution to Africa's problems. But neither is poverty a foregone conclusion. Through a combination of good policy and good luck, the African nation of Botswana has become a middle-income country, with GDP per person about equal to the world average and a rate of extreme poverty less than half of that in the rest of sub-Saharan Africa. Botswana has the disadvantage of being landlocked, and it is plagued by widespread HIV. But compared with most of its neighbors, Botswana has higher investment, better education, lower population growth, higher vaccination rates, lower malnutrition, greater freedoms, and less corruption, as well as being more fortunate for not having an extensive history of enslavement. Once a colonized country, it has become Africa's oldest continuous democracy. In some ways, Botswana is a model for what a nation can accomplish by focusing on the forces that shape economic growth. ●

QuickQuiz

6. Because capital is subject to diminishing returns, higher saving and investment do not lead to higher
   a. income in the long run.
   b. income in the short run.
   c. growth in the long run.
   d. growth in the short run.
7. When the Japanese car maker Toyota expands one of its car factories in the United States, what is the likely impact of this event on the gross domestic product and gross national product of the United States?
   a. GDP rises, and GNP falls.
   b. GNP rises, and GDP falls.
   c. GDP and GNP both rise, but GDP rises by more.
   d. GDP and GNP both rise, but GNP rises by more.
8. Thomas Robert Malthus believed that population growth would
   a. put stress on the economy's ability to produce food, dooming humans to remain in poverty.
   b. spread the capital stock too thinly across the labor force, lowering each worker's productivity.
   c. promote technological progress because there would be more scientists and inventors.
   d. eventually decline to sustainable levels, as birth control improved and people had smaller families.

Answers are at the end of the chapter.

In the News

The Secret Sauce of American Prosperity

Among large countries, the United States has long had the highest average income. In this opinion piece, an economist ponders the reasons for that success.

Why the U.S. Is Still Richer Than Every Other Large Country

By Martin Feldstein

Each year, the United States produces more per person than most other advanced economies. In 2015 real GDP per capita was $56,000 in the United States. The real GDP per capita in that same year was only $47,000 in Germany, $41,000 in France and the United Kingdom, and just $36,000 in Italy, adjusting for purchasing power.

In short, the U.S. remains richer than its peers. But why? I can think of 10 features that distinguish America from other industrial economies. . . .

An entrepreneurial culture. Individuals in the U.S. demonstrate a desire to start businesses and grow them, as well as a willingness to take risks. There is less penalty in U.S. culture for failing and starting again. Even students who have gone to college or a business school show this entrepreneurial desire, and it is self-reinforcing: Silicon Valley successes like Facebook inspire further entrepreneurship.

A financial system that supports entrepreneurship. The U.S. has a more developed system of equity finance than the countries of Europe, including angel investors willing to finance startups and a very active venture capital market that helps finance the growth of those firms. We also have a decentralized banking system, including more than 7,000 small banks, that provides loans to entrepreneurs.

World-class research universities. U.S. universities produce much of the basic research that drives high-tech entrepreneurship. Faculty members and doctoral graduates often spend time with nearby startups, and the culture of both the universities and the businesses encourage this overlap. Top research universities attract talented students from around the world, many of whom end up remaining in the United States.

Labor markets that generally link workers and jobs unimpeded by large trade unions, state-owned enterprises, or excessively restrictive labor regulations. Less than 7% of the private sector U.S. labor force is unionized, and there are virtually no state-owned enterprises. While the U.S. does regulate working conditions and hiring, the rules are much less onerous than in Europe. As a result, workers have a better chance of finding the right job, firms find it easier to innovate, and new firms find it easier to get started.

A growing population, including from immigration. America's growing population means a younger and therefore more flexible and trainable workforce. Although there are restrictions on immigration to the United States, there are also special rules that provide access to the U.S. economy and a path for citizenship (green cards), based on individual talent and industrial sponsorship. A separate "green card lottery" provides a way for eager people to come to the United States. The country's ability to attract immigrants has been an important reason for its prosperity.

A culture (and a tax system) that encourages hard work and long hours. The average employee in the United States works 1,800 hours per year, substantially more

26-4 Conclusion: The Importance of Long-Run Growth

This chapter discussed what determines the standard of living in a nation and how policymakers can try to raise it through policies that promote economic growth. Most of the analysis is summarized in one of the **Ten Principles of Economics**: A country's standard of living depends on its ability to produce goods and services. Policymakers who want to foster growth in living standards must aim to increase productivity by encouraging the rapid accumulation of the factors of production and ensuring these factors are employed as effectively as possible.

Economists differ on the role of government in promoting economic growth. At the very least, government can lend support to the invisible hand by maintaining property rights and political stability. More controversial is whether government should target and subsidize specific industries that might be especially important for technological progress. These issues are among the most important in economics. The success of one generation's policymakers in learning and heeding the fundamental lessons about economic growth determines what kind of world the next generation will inherit.

than the 1,500 hours worked in France and the 1,400 hours worked in Germany (though not as much as the 2,200+ in Hong Kong, Singapore, and South Korea). In general, working longer means producing more, which means higher real incomes.

A supply of energy that makes North America energy independent. Natural gas fracking in particular has provided U.S. businesses with plentiful and relatively inexpensive energy.

A favorable regulatory environment. Although U.S. regulations are far from perfect, they are less burdensome on businesses than the regulations imposed by European countries and the European Union.

A smaller size of government than in other industrial countries. According to the OECD, outlays of the U.S. government at the federal, state, and local levels totaled 38% of GDP, while the corresponding figure was 44% in Germany, 51% in Italy, and 57% in France. The higher level of government spending in other countries implies not only a higher share of income taken in taxes but also higher transfer payments that reduce incentives to work. It's no surprise that Americans work a lot; they have extra incentive to do so.

A decentralized political system in which states compete. Competition among states encourages entrepreneurship and work, and states compete for businesses and for individual residents with their legal rules and tax regimes. Some states have no income taxes and have labor laws that limit unionization. States provide high-quality universities with low tuition for in-state students. They compete in their legal liability rules, too. The legal systems attract both new entrepreneurs and large corporations. The United States is perhaps unique among high-income nations in its degree of political decentralization.

Will America maintain these advantages? In his 1942 book, *Socialism, Capitalism, and Democracy*, Joseph Schumpeter warned that capitalism would decline and fail because the political and intellectual environment needed for capitalism to flourish would be undermined by the success of capitalism and by the critique of intellectuals. He argued that popularly elected social democratic parties would create a welfare state that would restrict entrepreneurship.

Although Schumpeter's book was published more than 20 years after he had moved from Europe to the United States, his warning seems more appropriate to Europe today than to the United States. The welfare state has grown in the United States, but much less than it has grown in Europe. And the intellectual climate in the United States is much more supportive of capitalism.

If Schumpeter were with us today, he might point to the growth of the social democratic parties in Europe and the resulting expansion of the welfare state as reasons why the industrial countries of Europe have not enjoyed the same robust economic growth that has prevailed in the United States. ■

Source: *Harvard Business Review*, April 20, 2017.

Questions to Discuss

1. Which attributes of the United States listed in this article do you think best explain U.S. prosperity? Why?
2. Which of the attributes listed in this article do you think are most at risk of being undermined by poor policy choices? Why?
3. Do you agree with the author's point of view? On what items do you disagree, and why?

Mr. Feldstein was a professor of economics at Harvard University.

Chapter in a Nutshell

- Economic prosperity, as measured by GDP per person, varies substantially around the world. The average income in the world's richest countries is more than 10 times that in the poorest. Because growth rates of real GDP also vary substantially, the relative positions of countries can change dramatically over time.
- The standard of living depends on an economy's ability to produce goods and services. Productivity, in turn, depends on the physical capital, human capital, natural resources, and technological knowledge available to workers.
- Government policies can try to influence the economy's growth rate in many ways: by encouraging saving and investment, facilitating investment from abroad, fostering education, promoting good health, maintaining property rights and political stability, allowing free trade, and supporting the research and development of new technologies.
- The accumulation of capital is subject to diminishing returns: The more capital an economy has, the less additional output it gets from an extra unit of capital. Although higher saving and investment lead to

higher growth for a while, growth eventually slows down as capital, productivity, and income rise. Also because of diminishing returns, the return to capital is often high in poor countries. Other things being equal, these countries can grow faster because of the catch-up effect.

- Population growth has various effects on economic growth. More rapid population growth may lower productivity by stretching the supply of natural resources and by reducing the amount of capital available to each worker. But a larger population may enhance the rate of technological progress because there are more scientists and engineers.

Key Concepts

productivity, p. 533
physical capital, p. 533
human capital, p. 534
natural resources, p. 534
technological knowledge, p. 534
diminishing returns, p. 537
catch-up effect, p. 538

Questions for Review

1. What does the level of a nation's GDP measure? What does the growth rate of GDP measure? Would you rather live in a nation with a high level of GDP and a low growth rate or in a nation with a low level of GDP and a high growth rate?
2. List and describe four determinants of productivity.
3. In what way is a college degree a form of capital?
4. Explain how higher saving leads to a higher standard of living. What might deter a policymaker from trying to raise the rate of saving?
5. Does a higher rate of saving lead to higher growth temporarily or indefinitely?
6. Why would removing a trade restriction, such as a tariff, lead to more rapid economic growth?
7. How does the rate of population growth influence the level of GDP per person?
8. Describe two ways the U.S. government tries to encourage advances in technological knowledge.

Problems and Applications

1. Most countries, including the United States, import substantial amounts of goods and services from other countries. Yet the chapter says that a nation can enjoy a high standard of living only if it can produce a large quantity of goods and services itself. Can you reconcile these two facts?
2. Suppose that society decided to reduce consumption and increase investment.
   a. How would this change affect economic growth?
   b. What groups in society would benefit from this change? What groups might be hurt?
3. Societies choose what share of their resources to devote to consumption and what share to devote to investment. Some of these decisions involve private spending; others involve government spending.
   a. Describe some forms of private spending that represent consumption and some forms that represent investment. The national income accounts include tuition as a part of consumer spending. In your opinion, are the resources you devote to your education a form of consumption or a form of investment?
   b. Describe some forms of government spending that represent consumption and some forms that represent investment. In your opinion, should we view government spending on health programs as a form of consumption or investment? Would

you distinguish between health programs for the young and health programs for the elderly?

4. What is the opportunity cost of investing in capital? Do you think a country can overinvest in capital? What is the opportunity cost of investing in human capital? Do you think a country can overinvest in human capital? Explain.
5. In the 1990s and the first two decades of the 2000s, investors from the Asian economies of Japan and China made significant direct and portfolio investments in the United States. At the time, many Americans were unhappy that this investment was occurring.
   a. In what way was it better for the United States to receive this foreign investment than not to receive it?
   b. In what way would it have been even better for Americans to have made this investment themselves?
6. In many developing nations, young women have lower enrollment rates in secondary school than young men. Describe several ways in which greater educational opportunities for young women could lead to faster economic growth in these countries.
7. The International Property Right Index scores countries based on their legal and political environments and the extent to which they protect property rights. Go online and find a recent ranking. Choose three countries with high scores and three countries with low scores. Then find estimates of GDP per person in each of these six countries. What pattern do you find? Give two possible interpretations of the pattern.
8. International data show a positive correlation between income per person and the health of the population.
   a. Explain how higher income might cause better health outcomes.
   b. Explain how better health outcomes might cause higher income.
   c. How might the relative importance of your two hypotheses be relevant for public policy?
9. The great 18th-century economist Adam Smith wrote, "Little else is requisite to carry a state to the highest degree of opulence from the lowest barbarism but peace, easy taxes, and a tolerable administration of justice: all the rest being brought about by the natural course of things." Explain how each of the three conditions Smith describes promotes economic growth.

QuickQuiz Answers

1. **b** 2. **c** 3. **a** 4. **b** 5. **d** 6. **c** 7. **c** 8. **a**

Chapter

27

Saving, Investment, and the Financial System

Imagine that after graduating from college (with a degree in economics, of course), you decide to start your own business—an economic forecasting firm. Before selling any forecasts, you have to incur substantial set-up costs. You need computers, printers, software, desks, chairs, and filing cabinets. These are capital goods that your firm will use to produce and sell its services.

How do you obtain the funds to invest in this capital? Perhaps you can draw down your own savings. More likely, though, you are like most entrepreneurs, who don't have enough money on their own to start a new business. You must get the money you need from somewhere else.

There are various ways to finance these capital investments. You could borrow from a bank, friend, or relative, promising to return the money later and pay interest for the use of it. Or you could convince someone to provide the money in exchange for a share of your future profits. In either case, your investment in computers and office equipment would be financed by someone else's saving.

GAUDILAB/SHUTTERSTOCK.COM; LECHATNOIR/E+/GETTY IMAGES

financial system
the group of institutions in the economy that help to match one person's saving with another person's investment

The **financial system** consists of the institutions that help match one person's saving with another person's investment. As the previous chapter discussed, saving and investment are key ingredients to long-run economic growth: When a country saves a large portion of its GDP, more resources are available for investment in capital, and greater capital increases a country's productivity and living standard. The coordination of saving and investment isn't a simple matter, however. At any time, some people want to save some of their income for the future, and others want to borrow to invest in new and growing businesses. What brings these two groups of people together? What ensures that the supply of funds from those who want to save balances the demand for funds from those who want to invest?

This chapter examines how the financial system works. First, we discuss the large variety of institutions that make up the economy's financial system. Second, we examine the relationship between the financial system and some key macroeconomic variables—notably saving and investment. Third, we develop a model of the supply and demand for funds in financial markets. In this simple model, there is only one interest rate, and it functions as the price that adjusts to balance supply and demand. The model shows how government policies affect the interest rate and, in turn, society's allocation of scarce resources.

27-1 Financial Institutions in the U.S. Economy

At the broadest level, the financial system moves scarce resources from savers (people who spend less than they earn) to borrowers (people who spend more than they earn). Savers have various goals—from putting aside enough money for a vacation in a few months to putting a child through college in several years to retiring comfortably decades from now. Similarly, borrowers need money for various reasons—from buying a car to buying a house to starting a business. Savers supply their money to the financial system with the expectation that they will get it back with interest later. Borrowers demand money from the financial system with the knowledge that they will be required to pay it back with interest.

The financial system is made up of the institutions that help coordinate the actions of savers and borrowers. As a prelude to analyzing the economic forces that drive the financial system, let's consider the most important of these institutions. They can be grouped into two categories: financial markets and financial intermediaries.

27-1a Financial Markets

financial markets
financial institutions through which savers can directly provide funds to borrowers

Financial markets are the institutions through which a person who wants to save can directly supply funds to a person who wants to borrow. The two most important financial markets are the bond market and the stock market.

bond
a certificate of indebtedness

The Bond Market When Intel, the maker of computer chips, wants to borrow to finance the construction of a new factory, it can borrow directly from the public. It does so by selling bonds. A **bond** is a certificate of indebtedness that specifies the obligations of the borrower to the buyer of the bond. Put simply, a bond buyer is a lender, and a bond is an IOU. The bond identifies the time at which the loan will be repaid, called the **date of maturity**, and the rate of interest the borrower will pay periodically until the loan matures. Buyers of the bond give their money to Intel in exchange for this promise of interest and eventual repayment of the amount borrowed, called the **principal**. The buyers can hold the bond until maturity, or they can sell the bond at an earlier date to someone else.

There are millions of different bonds in the U.S. economy. When large corporations, the federal government, or state and local governments need to borrow to finance the purchase of a new factory, a new jet fighter, or a new school, they usually do so by issuing bonds. If you look at *The Wall Street Journal* or the business section of your news service, you will find a listing of the prices and interest rates on some of the most important bond issues. These bonds differ according to four significant characteristics.

The first is its **term**—the length of time until the bond matures. Some bonds have short terms, such as a few months, while others have terms for thirty or more years. (The British government has at times even issued bonds that never mature, called **perpetuities**. These bonds pay interest forever, but the principal never needs to be repaid.) The interest rate on a bond depends, in part, on its term. Long-term bonds are riskier than short-term ones because holders of long-term bonds have to wait longer for repayment of principal. If the holders of a long-term bond need their money earlier than the distant date of maturity, they have no choice but to sell the bond to someone else, perhaps at a reduced price. To compensate for this risk, long-term bonds tend to pay higher interest rates than short-term bonds.

The second important characteristic of a bond is its **credit risk**—the probability that the borrower will fail to pay some of the interest or principal. Such a failure to pay is called a **default**. Borrowers can (and sometimes do) default on their loans by declaring bankruptcy. When bond buyers perceive that the probability of default is high, they demand a higher interest rate as compensation. Because the U.S. government is considered to have low credit risk, U.S. government bonds tend to pay low interest rates. By contrast, financially shaky corporations raise money by issuing **high-yield bonds**, commonly known as junk bonds, which pay high interest rates. Buyers of bonds can judge credit risk by checking with private agencies that evaluate the financial condition of bond issuers. For example, Standard & Poor's rates bonds from AAA (the safest) to D (those already in default).

The third important characteristic of a bond is its **tax treatment**—the way the tax laws treat the interest earned on it. For most bonds, the interest is taxable income; that is, bond owners have to pay a portion of the interest they earn in income taxes. But when state and local governments issue bonds, called **municipal bonds**, the bond owners are not required to pay federal income tax on the interest income and, in some cases, may not need to pay state and local taxes either. Because of this tax advantage, bonds issued by state and local governments typically pay a lower interest rate than bonds issued by corporations or the federal government.

The fourth important characteristic of a bond is whether it offers **inflation protection**. Most bonds are written in nominal terms—that is, they promise to pay interest and principal in a specific number of dollars (or perhaps another currency). If prices rise and dollars have less purchasing power, the bondholder is worse off. Some bonds, however, index the payments of interest and principal to a measure of inflation so that when prices rise, the payments rise proportionately. Since 1997, the U.S. government has issued such bonds, called Treasury Inflation-Protected Securities (TIPS). Because TIPS offer inflation protection, they generally pay a lower interest rate than similar bonds without this feature.

The Stock Market Another way for Intel to raise funds to build a new semiconductor factory is to sell stock in the company. A share of **stock** represents partial ownership in a firm and is a claim to some of the profits the firm makes. For example, if Intel sells a total of 1,000,000 shares of stock, then each share represents ownership of 1/1,000,000 of the business.

stock
a claim to partial ownership in a firm

The sale of stock to raise money is called **equity finance**, while the sale of bonds is called **debt finance**. Although corporations use both equity and debt finance to raise money for new investments, stocks and bonds are very different. The owner of shares of Intel stock is a part owner of Intel, while the owner of an Intel bond is a creditor of the corporation. If Intel is very profitable, the stockholders enjoy the benefits of these profits, but the bondholders get only the stated interest on their bonds. And if Intel runs into financial trouble, the bondholders are paid what they are due before stockholders receive anything at all. Compared with bonds, stocks carry greater risk but offer potentially higher returns.

After a corporation issues stock by selling shares to the public, these shares trade on organized stock exchanges. In these transactions, the corporation itself receives no money when its stock changes hands. The most important stock exchanges in the U.S. economy are the New York Stock Exchange and the Nasdaq (National Association of Securities Dealers Automated Quotations). Most of the world's countries have their own stock exchanges on which the shares of local companies trade, the most important being those in Tokyo, Shanghai, Hong Kong, and London.

The prices at which shares trade on stock exchanges are determined by supply and demand. Because stock represents ownership in a corporation, the demand for a stock (and thus its price) is in part based on people's perception of the corporation's future profitability. When people become optimistic about a company, they raise their demand for its stock and bid up its share price. Conversely, when people's expectations of a company's prospects decline, the price of a share falls.

Stock indexes track the overall level of stock prices. A **stock index** is computed as an average of a group of stock prices. The most famous stock index is the Dow Jones Industrial Average, which was introduced in 1896. It is calculated from the stock prices of thirty major U.S. companies, such as Disney, Microsoft, Coca-Cola, Boeing, Apple, and Walmart. Another well-known stock index is the Standard & Poor's 500 Index, which is based on the stock prices of 500 major companies. Because stock prices reflect expected profitability, these indexes are watched closely as possible indicators of future economic conditions.

27-1b Financial Intermediaries

financial intermediaries
financial institutions through which savers can indirectly provide funds to borrowers

Financial intermediaries are financial institutions through which savers can indirectly provide funds to borrowers. The term **intermediary** reflects the role of these institutions in standing between savers and borrowers. Banks and mutual funds are two of the most important financial intermediaries.

Banks Consider the Kim family, which owns a small convenience store. If the Kims want to finance a business expansion, they will proceed differently than Intel. Unlike a large corporation, a small, family-owned business would find it difficult to raise funds in the stock and bond markets. Most buyers of stocks and bonds prefer to buy those issued by larger, more familiar companies. The Kims will most likely finance their business expansion with a loan from a local bank.

Banks are the financial intermediaries with which people are most familiar. A primary job of banks is to take in deposits from people who want to save and use these deposits to make loans to people who want to borrow. Banks pay depositors interest on their deposits and charge borrowers slightly higher interest on their loans. The difference between these rates of interest covers the banks' costs and returns some profit to the banks' owners.

Besides being financial intermediaries, banks play another important role: They facilitate the exchange of goods and services. People can use bank deposits to buy

things by writing a check, making an electronic payment, or using a debit card. In other words, banks help create a special asset, called a **medium of exchange**, that people can use to engage in transactions. A bank's role in providing a medium of exchange distinguishes it from many other financial institutions. Stocks and bonds, like bank deposits, offer a **store of value** for the wealth that people have accumulated in past saving, but they do not offer the easy, cheap, and immediate access to wealth that a person gets by writing a check, tapping a "Pay now" button, or swiping a debit card. For now, we ignore this second role of banks but will return to it later when we discuss the monetary system.

Mutual Funds A financial intermediary of increasing importance in the U.S. economy is the mutual fund. A **mutual fund** is an institution that sells shares to the public and uses the proceeds to buy a selection, or **portfolio**, of various types of stocks, bonds, or both stocks and bonds. The shareholder of the mutual fund accepts all the risk and return associated with the portfolio. If the value of the portfolio rises, the shareholder benefits; if the value of the portfolio falls, the shareholder suffers the loss.

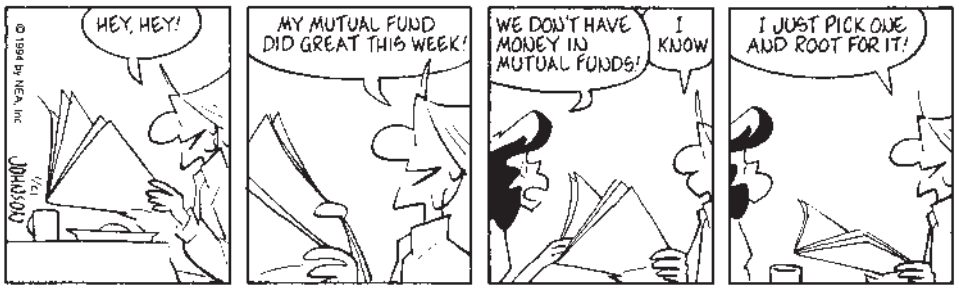
mutual fund
an institution that sells shares to the public and uses the proceeds to buy a portfolio of stocks and bonds

The primary advantage of mutual funds (and their close cousins, exchange-traded funds) is that they allow people with small amounts of money to diversify their holdings. Because the value of any single stock or bond is tied to the fortunes of one company, holding a single kind of stock or bond is very risky. By contrast, people who hold a diverse portfolio of stocks and bonds face less risk because they have only a small stake in each company. Mutual funds make this diversification easy. With only a few hundred dollars, a person can buy shares in a mutual fund and, indirectly, become the part owner or creditor of hundreds of major companies. For this service, the company operating the mutual fund typically charges shareholders a fee, usually between 0.1 and 1.5 percent of assets each year.

A second advantage claimed by mutual fund companies is that the funds can give ordinary people access to the skills of professional money managers. The managers of many mutual funds pay close attention to the developments and prospects of the companies in which they buy stock. These managers buy the stock of companies they view as having profitable futures and sell the stock of companies with less promising prospects. This professional management, it is argued, should increase the return that mutual fund depositors earn on their savings.

Financial economists, however, are often skeptical of this argument. Because thousands of money managers are paying close attention to each company's prospects, a company's stock usually trades at a price that reflects the company's true value. As a result, it is hard to "beat the market" by buying good stocks and selling bad ones. In fact, mutual funds called **index funds**, which buy all the stocks in a stock

ARLO AND JANIS by Jimmy Johnson



index, perform somewhat better on average than mutual funds that take advantage of active trading by professional money managers. The explanation for the superior performance of index funds is that they keep costs low by buying and selling very rarely and by not having to pay the salaries of professional money managers.

27-1c Summing Up

The U.S. economy contains a large variety of financial institutions. In addition to the bond market, the stock market, banks, and mutual funds, there are also pension funds, credit unions, insurance companies, and even the local loan shark. These institutions differ in many ways. When analyzing the macroeconomic role of the financial system, however, it is more important to keep in mind that, despite their differences, these financial institutions all serve the same goal: directing the resources of savers into the hands of borrowers.

QuickQuiz

1. Lois wants to publish a local newspaper but doesn't have the financial resources to start the business. She borrows $60,000 from her friend Clark, to whom she promises an interest rate of 7 percent, and gets another $40,000 from her friend Jimmy, to whom she promises 10 percent of her profits. What best describes this situation?
   a. Clark is a stockholder, and Lois is a bondholder.
   b. Clark is a stockholder, and Jimmy is a bondholder.
   c. Jimmy is a stockholder, and Lois is a bondholder.
   d. Jimmy is a stockholder, and Clark is a bondholder.
2. A bond tends to pay a high interest rate if it is
   a. a short-term bond rather than a long-term bond.
   b. a municipal bond exempt from federal taxation.
   c. issued by the federal government rather than a corporation.
   d. issued by a corporation of dubious credit quality.
3. The main advantage of mutual funds is that they provide
   a. a return insured by the government.
   b. an easy way to hold a diversified portfolio.
   c. an asset that is widely used as the medium of exchange.
   d. a way to avoid fluctuations in stock and bond prices.

Answers are at the end of the chapter.

27-2 Saving and Investment in the National Income Accounts

Events that occur within the financial system are central to developments in the overall economy. As we have just seen, the institutions that make up this system—the bond market, the stock market, banks, and mutual funds—coordinate the economy's saving and investment. As the previous chapter discussed, saving and investment are important determinants of long-run growth in GDP and living standards. As a result, macroeconomists need to understand how financial markets work and how events and policies affect them.

As a starting point, let's consider the key macroeconomic variables that measure activity in these markets. Our emphasis here is not on behavior but on accounting. **Accounting** refers to the way in which various numbers are defined and added up. A personal accountant might help a family add up its income and expenses. A national income accountant does the same thing for the overall economy. The national income accounts include, in particular, GDP and the many related statistics.

The rules of national income accounting include several important identities. Recall that an **identity** is an equation that must be true because of the way the variables in the equation are defined. Identities clarify how different variables are related to one another. Here, we consider some accounting identities that shed light on the macroeconomic role of financial markets.

27-2a Some Important Identities

Recall that gross domestic product (GDP) is both total income in an economy and the total expenditure on the economy's output of goods and services. GDP (denoted as Y) is divided into four components of expenditure: consumption (C), investment (I), government purchases (G), and net exports (NX):

$$Y = C + I + G + NX.$$

This equation is an identity because every dollar of expenditure that shows up on the left side also shows up in one of the four components on the right side. Because of the way each of the variables is defined and measured, this equation must always hold.

To keep things simple, this chapter assumes that the economy we are examining is closed. A **closed economy** is one that does not interact with other economies. In particular, a closed economy does not engage in international trade in goods and services, and it does not engage in international borrowing and lending. Actual economies are **open economies**—that is, they interact with other economies around the world. Nonetheless, assuming a closed economy is a useful simplification with which we can learn some lessons that apply to all economies. Moreover, this assumption applies perfectly to the world economy (interplanetary trade is not yet common!).

Because a closed economy does not engage in international trade, there are no imports and exports, making net exports (NX) exactly zero. We can simplify the identity as

$$Y = C + I + G.$$

This equation states that GDP is the sum of consumption, investment, and government purchases. Each unit of output sold in a closed economy is consumed, invested, or bought by the government.

To see what this identity says about financial markets, subtract C and G from both sides of this equation to obtain

$$Y - C - G = I.$$

The left side of this equation ($Y - C - G$) is the total income in the economy that remains after paying for consumption and government purchases: This amount is called **national saving**, or just **saving**, and is denoted S. Substituting S for $Y - C - G$, we can write the last equation as

national saving (saving)
the total income in the economy that remains after paying for consumption and government purchases

$$S = I.$$

This equation states that saving equals investment.

To understand the meaning of national saving, it is helpful to manipulate the definition a bit more. Let T denote the amount that the government collects from households in taxes minus the amount it pays back to households in the form of

transfer payments (such as Social Security and welfare). We can then write national saving in either of two ways:

$$S = Y - C - G$$

or

$$S = (Y - T - C) + (T - G).$$

These equations are the same because the two *T*s in the second equation cancel each other, but each reveals a different way of thinking about national saving. In particular, the second equation separates national saving into two pieces: private saving $(Y - T - C)$ and public saving $(T - G)$.

private saving
the income that households have left after paying for taxes and consumption

public saving
the tax revenue that the government has left after paying for its spending

budget surplus
an excess of tax revenue over government spending

budget deficit
a shortfall of tax revenue from government spending

Consider each of these pieces. **Private saving** is the amount of income that households have left after paying their taxes and paying for their consumption. In particular, because households receive income of *Y*, pay taxes of *T*, and spend *C* on consumption, private saving is $Y - T - C$. **Public saving** is the amount of tax revenue that the government has left after paying for its spending. The government receives *T* in tax revenue and spends *G* on goods and services. If *T* exceeds *G*, the government receives more money than it spends. In this case, public saving $(T - G)$ is positive, and the government is said to run a **budget surplus**. If *G* exceeds *T*, the government spends more than it receives in tax revenue. In this case, public saving $(T - G)$ is negative, and the government is said to run a **budget deficit**.

Now consider how these accounting identities are related to financial markets. The equation $S = I$ reveals an important fact: **For the economy as a whole, saving must equal investment.** Yet this fact raises some important questions: What mechanisms lie behind this identity? What coordinates those people who are deciding how much to save and those people who are deciding how much to invest? The answer is the financial system. The bond market, the stock market, banks, mutual funds, and other financial markets and intermediaries stand between the two sides of the $S = I$ equation. They take in the nation's saving and direct it to the nation's investment.

27-2b The Meaning of Saving and Investment

The terms **saving** and **investment** can sometimes be confusing. Most people use these terms casually and sometimes interchangeably. By contrast, the macroeconomists who construct the national income accounts use these terms carefully and distinctly.

Consider an example. Suppose that Larry earns more than he spends and deposits his unspent income in a bank or uses it to buy some stock or a bond from a corporation. Because Larry's income exceeds his consumption, he adds to the nation's saving. Larry might think of himself as "investing" his money, but a macroeconomist would call Larry's act saving rather than investment.

In the language of macroeconomics, investment refers to the purchase of new capital, such as equipment or buildings. When Moe borrows from the bank to build himself a new house, he adds to the nation's investment. (Remember, the purchase of a new house is the one form of household spending that is investment rather than consumption.) Similarly, when the Curly Corporation sells some stock and uses the proceeds to build a new factory, it also adds to the nation's investment.

Although the accounting identity $S = I$ shows that saving and investment are equal for the overall economy, it does not mean that saving and investment are

equal for every individual household or firm. Larry's saving can be greater than his investment, and he can deposit the excess in a bank. Moe's saving can be less than his investment, and he can borrow the shortfall from a bank. Banks and other financial institutions make these individual differences between saving and investment possible by allowing one person's saving to finance another person's investment.

QuickQuiz

4. If the government collects more in tax revenue than it spends, and households consume more than they get in after-tax income, then
   a. private saving and public saving are both positive.
   b. private saving and public saving are both negative.
   c. private saving is positive, but public saving is negative.
   d. private saving is negative, but public saving is positive.

5. A closed economy has income of $1,000, government spending of $200, taxes of $150, and investment of $250. What is private saving?
   a. $100
   b. $200
   c. $300
   d. $400

Answers are at the end of the chapter.

27-3 The Market for Loanable Funds

Having discussed some of the important financial institutions in our economy and the macroeconomic role of these institutions, we are ready to build a model of financial markets. Our goal is to explain how financial markets coordinate an economy's saving and investment. The model also provides a tool with which we can analyze government policies that influence saving and investment.

To keep things simple, we assume that the economy has only one financial market, called the **market for loanable funds**. All savers deposit their saving in this market, and all borrowers take out their loans there. The term **loanable funds** refers to the income that people have chosen to save and lend out rather than use for their own consumption and to the amount that investors have chosen to borrow to fund new investment projects. In the market for loanable funds, there is one interest rate, which is both the return to saving and the cost of borrowing.

market for loanable funds
the market in which those who want to save supply funds and those who want to borrow to invest demand funds

The assumption of a single financial market isn't realistic, as the economy has many types of financial institutions. But recall from Chapter 2 that the art in building an economic model is simplifying the world in order to explain it. For our purposes, we can ignore the diversity of financial institutions and assume that the economy has a single financial market.

27-3a Supply and Demand for Loanable Funds

Like most other markets, the economy's market for loanable funds is governed by supply and demand.

The supply of loanable funds comes from people who have extra income they want to save and lend out. This lending can occur directly, such as when a household buys a bond from a firm, or indirectly, such as when a household makes a deposit in a bank, which then uses the funds to make loans. **In both cases, saving is the source of the supply of loanable funds.**

The demand for loanable funds comes from households and firms that wish to borrow to make investments. This demand includes families taking out mortgages to buy new homes. It also includes firms borrowing to buy new equipment or build factories. **In both cases, investment is the source of the demand for loanable funds.**

The interest rate may be viewed as the price of a loan. It represents the amount that borrowers pay for loans and the amount that lenders receive on their saving. Because a high interest rate makes borrowing more expensive, the quantity of loanable funds demanded falls as the interest rate rises. Similarly, because a high interest rate makes saving more attractive, the quantity of loanable funds supplied rises as the interest rate rises. In other words, the demand curve for loanable funds slopes downward, and the supply curve for loanable funds slopes upward.

Figure 1 shows the interest rate that balances the supply and demand for loanable funds. In the equilibrium shown, the interest rate is 5 percent, and the quantity of loanable funds demanded and the quantity of loanable funds supplied both equal $1,200 billion.

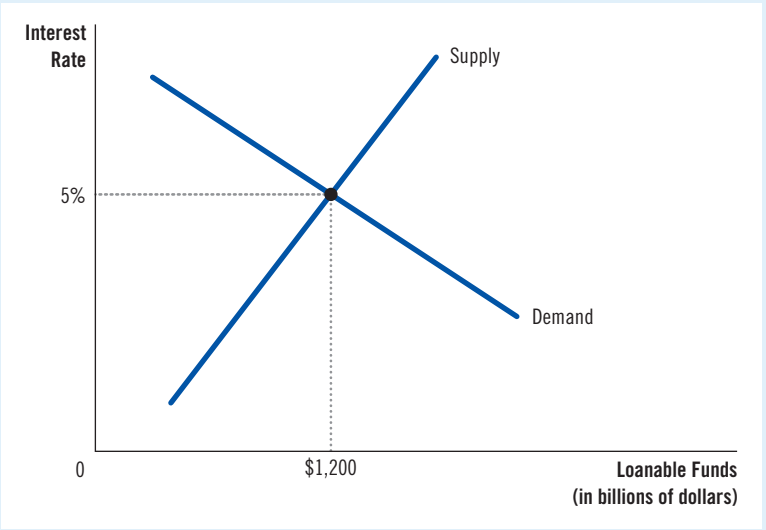
The adjustment of the interest rate to the equilibrium level occurs for the usual reasons. If the interest rate were lower than the equilibrium level, the quantity of loanable funds supplied would be less than the quantity demanded. The resulting shortage would encourage lenders to raise the interest rate they charge. A higher interest rate would encourage saving (increasing the quantity of loanable funds supplied) and discourage borrowing for investment (decreasing the quantity demanded). Conversely, if the interest rate were higher than the equilibrium level, the quantity of loanable funds supplied would exceed the quantity demanded. As lenders compete for scarce borrowers, interest rates would fall. In this way, the interest rate approaches the equilibrium level at which the supply and demand for loanable funds exactly balance.

Recall that economists distinguish between the real and nominal interest rate. The nominal interest rate is the interest rate as usually reported—the monetary

Figure 1

The Market for Loanable Funds

The interest rate in the economy adjusts to balance the supply and demand for loanable funds. The supply of loanable funds comes from national saving, both private and public. The demand for loanable funds comes from firms and households that want to borrow for purposes of investment. Here, the equilibrium interest rate is 5 percent, and $1,200 billion of loanable funds are supplied and demanded.



return to saving and the monetary cost of borrowing. The real interest rate is the nominal interest rate corrected for changes in the price level: It equals the nominal interest rate minus the inflation rate. Because inflation or deflation changes the value of money over time, the real interest rate more accurately reflects the real return to saving and the real cost of borrowing. Therefore, the supply and demand for loanable funds depend on the real (rather than nominal) interest rate, and the equilibrium in Figure 1 should be interpreted as determining the economy's real interest rate. For the rest of this chapter, when you see the term **interest rate**, remember that we are talking about the real interest rate.

This model of the supply and demand for loanable funds shows that financial markets work much like other markets. In the market for milk, for instance, the price of milk adjusts so the quantity of milk supplied balances the quantity demanded. In this way, the invisible hand coordinates the behavior of dairy farmers and milk drinkers. Once we realize that saving represents the supply of loanable funds and investment represents the demand, we can see how the invisible hand coordinates saving and investment. When the interest rate adjusts to balance supply and demand in the market for loanable funds, it coordinates the behavior of people who want to save (the suppliers of loanable funds) and the behavior of people who want to invest (the demanders of loanable funds).

We can use this model of the market for loanable funds to examine government policies that affect the economy's saving and investment. Because the model is just supply and demand in a particular market, use the three steps discussed in Chapter 4. First, decide whether the policy shifts the supply curve or the demand curve. Second, determine the direction of the shift. Third, use the supply-and-demand diagram to see how the equilibrium changes.

27-3b Policy 1: Saving Incentives

Many economists and policymakers have advocated increases in saving. Their argument is simple. One of the **Ten Principles of Economics** in Chapter 1 is that a country's standard of living depends on its ability to produce goods and services. And as the preceding chapter discussed, saving is an important long-run determinant of a nation's productivity. If the United States could somehow raise its saving rate, more resources would be available for capital accumulation, GDP would grow more rapidly, and over time, people would enjoy a higher standard of living.

Another of the **Ten Principles of Economics** is that people respond to incentives. Many economists have relied on this principle to suggest that the low rate of saving is at least partly attributable to tax laws that discourage saving. The U.S. federal government, as well as many state governments, collects revenue by taxing income, including interest and dividend income. To see the effects of this policy, consider a 25-year-old who saves $1,000 and buys a 30-year bond that pays an interest rate of 9 percent. Without taxes, the $1,000 grows to $13,268 when the individual reaches age 55. But if the interest income is taxed at a rate of, say, 33 percent, the after-tax interest rate is only 6 percent. In this case, the $1,000 grows to only $5,743 over the 30 years. The tax on interest income substantially reduces the future payoff from current saving and thereby reduces the incentive for people to save.

In response to this problem, some economists and lawmakers have proposed reforming the tax code to encourage greater saving. For example, one proposal is

to expand eligibility for special accounts, such as Individual Retirement Accounts, that allow people to shelter some of their saving from taxation. Let's consider the effect of such a saving incentive on the market for loanable funds, as illustrated in Figure 2.

First, which curve does this policy affect? Because the tax change alters the incentive for households to save at any given interest rate, it affects the quantity of loanable funds supplied at each interest rate. The supply curve for loanable funds shifts. The demand curve for loanable funds remains the same because the tax change does not directly affect the amount that borrowers want to borrow at any interest rate.

Second, which way does the supply curve shift? Because saving is being taxed less heavily than under current law, households increase their saving by consuming a smaller fraction of their income. Households use this additional saving to increase their deposits in banks or to buy more bonds. The supply of loanable funds increases, and the supply curve shifts to the right from S_1 to S_2, as shown in Figure 2.

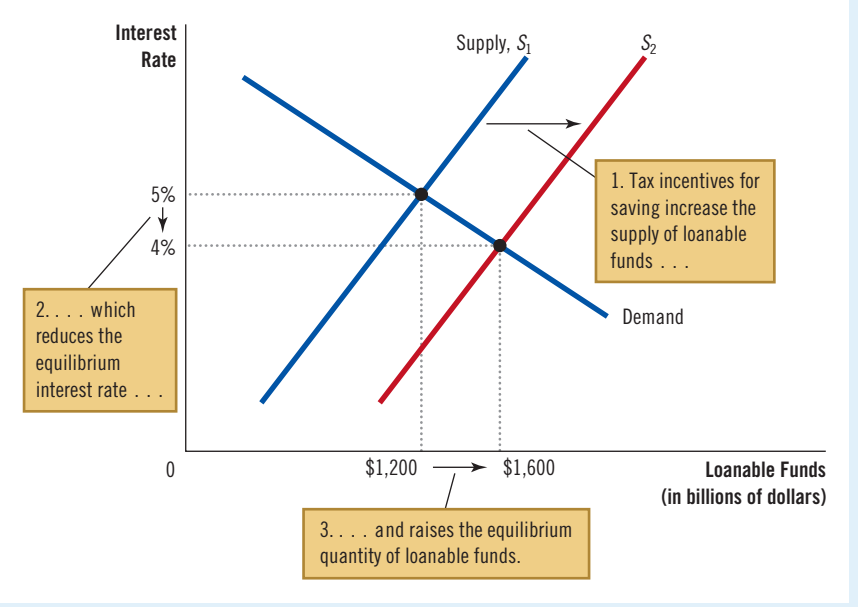
Finally, compare the old and new equilibria. In the figure, the increased supply of loanable funds reduces the interest rate from 5 percent to 4 percent. The lower interest rate raises the quantity of loanable funds demanded from $1,200 billion to $1,600 billion. That is, the shift in the supply curve moves the market equilibrium along the demand curve. With a lower cost of borrowing, households and firms are motivated to borrow more to finance greater investment. **In short, if a reform of the tax laws encourages greater saving, the result is lower interest rates and greater investment.**

This analysis of the effects of increased saving is widely accepted among economists, but there is less consensus about what kinds of tax changes should be enacted. Many economists endorse tax reform aimed at increasing saving to stimulate

Figure 2

Saving Incentives Increase the Supply of Loanable Funds

A change in the tax laws to encourage Americans to save more shifts the supply of loanable funds to the right from S_1 to S_2. As a result, the equilibrium interest rate falls, and the lower interest rate stimulates investment. Here, the equilibrium interest rate falls from 5 percent to 4 percent, and the equilibrium quantity of loanable funds saved and invested rises from $1,200 billion to $1,600 billion.



investment and growth. Yet others are skeptical that these tax changes would have a large effect on saving. These skeptics also doubt the equity of the proposed reforms. They argue that, in many cases, the benefits of the tax changes would accrue primarily to the wealthy, who are least in need of tax relief.

27-3c Policy 2: Investment Incentives

Suppose that Congress passes a tax reform aimed at making investment more attractive—for instance, by instituting an **investment tax credit**, as Congress has done from time to time. An investment tax credit gives a tax advantage to any firm building a new factory or buying a new piece of equipment. Let's consider the effect of such a tax reform on the market for loanable funds, as shown in Figure 3.

First, does the tax credit affect supply or demand? Because it rewards firms that borrow and invest in new capital, it alters investment at any interest rate and thereby changes the demand for loanable funds. But because the tax credit does not affect the amount that households save at any interest rate, it does not affect the supply of loanable funds.

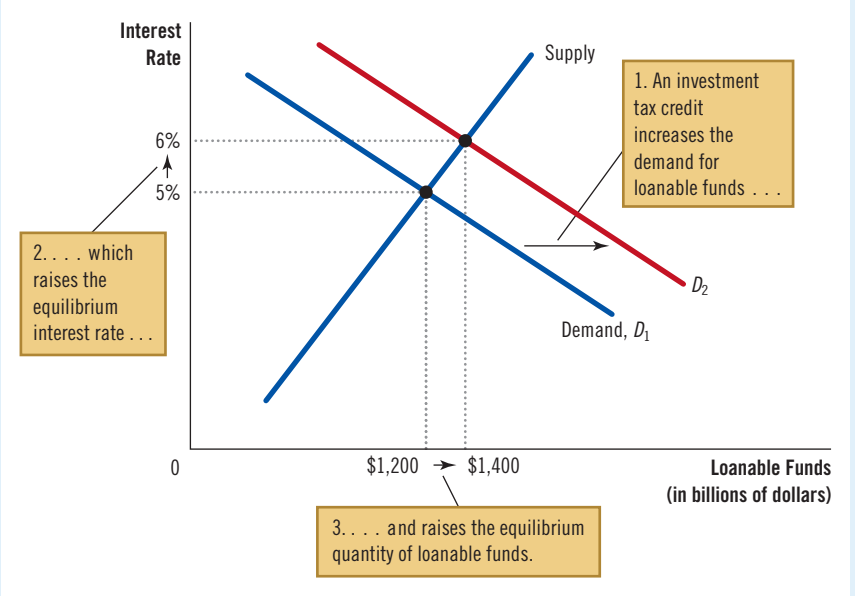
Second, which way does the demand curve shift? Because firms have an incentive to increase investment at any interest rate, the quantity of loanable funds demanded is higher at any interest rate. The demand curve for loanable funds moves to the right, as shown by the shift from D_1 to D_2 in the figure.

Third, consider how the equilibrium changes. In Figure 3, the increased demand for loanable funds raises the interest rate from 5 percent to 6 percent, and the higher interest rate, in turn, increases the quantity of loanable funds supplied from $1,200 billion to $1,400 billion as households respond by increasing the amount they save. This change in household behavior is represented as a movement along the supply curve. **Thus, if a reform of the tax laws encourages greater investment, the result is higher interest rates and greater saving.**

Figure 3

Investment Incentives Increase the Demand for Loanable Funds

If the passage of an investment tax credit encourages firms to invest more, the demand for loanable funds increases. As a result, the equilibrium interest rate rises, and the higher interest rate stimulates saving. Here, when the demand curve shifts from D_1 to D_2, the equilibrium interest rate rises from 5 percent to 6 percent, and the equilibrium quantity of loanable funds saved and invested rises from $1,200 billion to $1,400 billion.



Case Study

The Decline in Real Interest Rates from 1984 to 2020

In the evolution of the U.S. economy over the past few decades, one fact stands out as especially noteworthy: the large and fairly steady decline in real interest rates. Figure 4 illustrates the phenomenon. In the late 1980s and 1990s, the real interest rate was typically between 4 and 5 percent. In the 2010s, it was typically below 1 percent. In 2020, the real interest rate even dipped below zero. Over this period, many other nations experienced a similar trend.

What accounts for this decline, and what are its implications? Let's start with three reasons that saving might have increased, shifting the supply of loanable funds to the right:

- As income inequality has risen over the past few decades, resources have shifted from poorer households to richer ones. If the rich have higher propensities to save, more resources flow into capital markets.
- The Chinese economy has grown rapidly in recent years, and China has a high saving rate. This new, vast pool of saving flows into capital markets around the world.
- Events like the financial crisis of 2008 and the pandemic of 2020 are vivid reminders of how uncertain life can be. People may respond by increasing precautionary saving to prepare for such unfortunate events.

In addition, consider three reasons that investment may have declined, shifting the demand for loanable funds to the left:

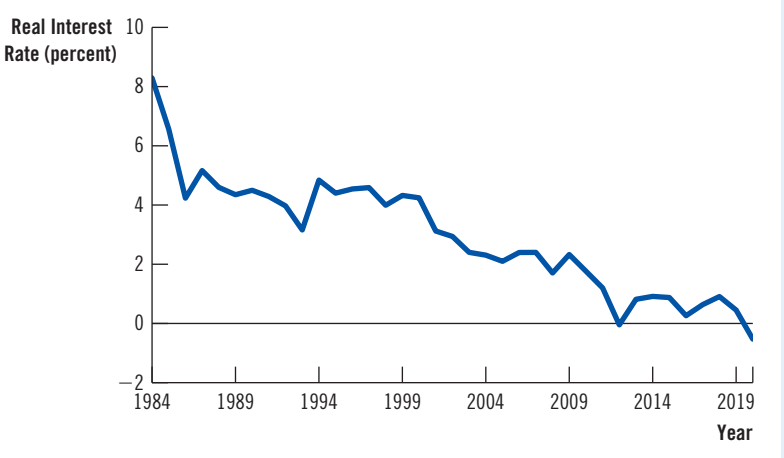
- Over the past several decades, average economic growth has slowed, due to a combination of lower productivity growth and lower population growth. A decline in growth reduces the demand for new capital investment.
- Old technologies, such as railroads and auto factories, required large capital investments. New technologies, like those developed in Silicon Valley, may be less capital-intensive.

Figure 4

The Decline in the Real Interest Rate

The real interest rate declined substantially from 1984 to 2020. The reason is a puzzle, though various hypotheses have been proposed.

Source: The Federal Reserve, the Department of Commerce, and the author's calculations. The real interest rate presented here is the yield on 10-year Treasury bonds minus the core inflation rate (based on the PCE deflator excluding food and energy) as a measure of expected inflation.



- Some economists have suggested that the U.S. economy is less competitive than it once was. Businesses with greater market power not only charge higher prices but also invest less.

Which of these hypotheses is right? Very likely, a combination of these forces is at work. When the supply of loanable funds shifts to the right and the demand shifts to the left, the impact on equilibrium saving and investment is ambiguous, but the effect on the equilibrium interest rate is clear: It falls.

Some of the implications of low interest rates are evident. For example, over the past century, a balanced portfolio of half stocks and half bonds has earned an average annual return of about 5 percent after inflation. Looking ahead as of 2021, a more plausible projection is a return of about 3 percent.

If that prospect is borne out, institutions like universities that use the return on their endowments to fund their activities will need to tighten their belts. It also means that individuals will need to rethink retirement saving. To support any level of spending for a 30-year retirement, a person's nest egg entering retirement needs to be 27 percent larger when the rate of return is 3 percent per year rather than 5. For much the same reason, public and private pension plans are probably more underfunded than current estimates suggest.

There are, however, also upsides to the decline in interest rates. Young families looking to buy homes, for example, benefit from the lower cost of mortgage financing.

In the end, the interest rate is just a price. A low price benefits those on the demand side of the market (young families taking out mortgages) and hurts those on the supply side (older individuals saving for retirement). If interest rates start to rise again, the winners and losers will be reversed. ●

27-3d Policy 3: Government Budget Deficits and Surpluses

A perpetual topic of political debate is the status of the government budget. Recall that a **budget deficit** is an excess of government spending over tax revenue. Governments finance budget deficits by borrowing in the bond market, and the accumulation of past government borrowing is called the **government debt**. A **budget surplus**, an excess of tax revenue over government spending, can be used to repay some of the government debt. If government spending exactly equals tax revenue, the government is said to have a **balanced budget**.

Imagine that the government starts with a balanced budget and then, because of an increase in government spending, starts to run a budget deficit. We can analyze the effects of the budget deficit by following our three steps in the market for loanable funds, as illustrated in Figure 5.

First, which curve shifts when the government starts running a budget deficit? Recall that national saving—the source of the supply of loanable funds—is composed of private and public saving. A change in the government budget balance represents a change in public saving and, therefore, in the supply of loanable funds. Because the budget deficit does not influence the amount that households and firms want to borrow to finance investment at any interest rate, it does not alter the demand for loanable funds.

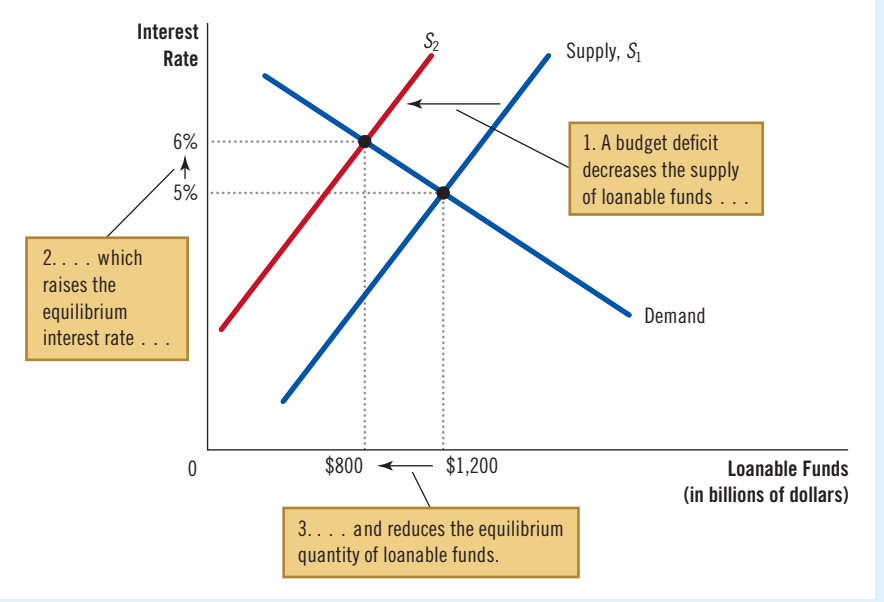
Second, in which direction does the supply curve shift? When the government runs a budget deficit, public saving is negative, so national saving declines. In other words, when the government borrows to finance its budget deficit, it reduces the supply of loanable funds available to finance investment. Thus, a budget deficit shifts the supply curve for loanable funds to the left from S_1 to S_2, as shown in Figure 5.

Third, compare the old and new equilibria. In the figure, when the budget deficit reduces the supply of loanable funds, the interest rate rises from 5 percent to

Figure 5

The Effect of a Government Budget Deficit

When the government spends more than it receives in tax revenue, the resulting budget deficit lowers national saving. The supply of loanable funds decreases, and the equilibrium interest rate rises. Thus, when the government borrows to finance its budget deficit, it crowds out households and firms that otherwise would borrow to finance investment. Here, when the supply curve shifts from S_1 to S_2, the equilibrium interest rate rises from 5 percent to 6 percent, and the equilibrium quantity of loanable funds saved and invested falls from $1,200 billion to $800 billion.



crowding out
a decrease in investment that results from government borrowing

6 percent. The quantity of loanable funds demanded then decreases from $1,200 billion to $800 billion as the higher interest rate discourages many demanders of loanable funds. Fewer families buy new homes, and fewer firms choose to build new factories. The fall in investment caused by government borrowing is represented by the movement along the demand curve and is called **crowding out**. That is, when the government borrows to finance its budget deficit, it crowds out private borrowers who are trying to finance investment.

In this standard model, the most basic lesson about budget deficits follows directly from their effects on the supply and demand for loanable funds: **When the government reduces national saving by running a budget deficit, the interest rate rises, and investment falls.** Because investment is important for long-run growth, government budget deficits reduce the economy's growth rate.

Why, you might ask, does a budget deficit affect the supply of loanable funds rather than the demand for them? After all, the government finances a budget deficit by selling bonds, thereby borrowing from the private sector. Why does increased borrowing by the government shift the supply curve, while increased borrowing by private investors shifts the demand curve? To answer this question, we need to examine more precisely the meaning of "loanable funds." The model as presented here takes this term to mean the **flow of resources available to fund private investment**; for that reason, a government budget deficit reduces the supply of loanable funds. If, instead, we had defined the term "loanable funds" to mean the **flow of resources available from private saving**, then the government budget deficit would increase demand rather than reduce supply. Changing the interpretation of the term would cause a semantic change in how we described the model, but the upshot of the analysis would be the same: In either case, a budget deficit increases the interest rate, thereby crowding out private borrowers who are relying on financial markets to fund private investment projects.

So far, we have examined a budget deficit that results from an increase in government spending, but a budget deficit that results from a tax cut has similar effects. A tax cut reduces tax revenue T and public saving, $T - G$. Private saving, $Y - T - C$, might increase because of lower T, but as long as households respond to the lower taxes by consuming more, C increases, so private saving rises by less than public saving falls. National saving ($S = Y - C - G$), the sum of public saving and private saving, declines. Once again, the budget deficit reduces the supply of loanable funds, drives up the interest rate, and crowds out borrowers trying to finance capital investments.

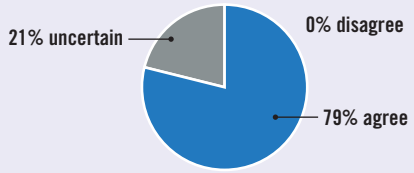
Having examined the impact of budget deficits, we can turn the analysis around and see the opposite effects of government budget surpluses. When the government collects more in tax revenue than it spends, it saves the difference by retiring some of the government debt. This budget surplus, or public saving, contributes to national saving. **Thus, a budget surplus increases the supply of loanable funds, reduces the interest rate, and stimulates investment.** Higher investment, in turn, means greater capital accumulation and more rapid economic growth.

Ask the Experts

Fiscal Policy and Saving

"Sustained tax and spending policies that boost consumption in ways that reduce the saving rate are likely to lower long-run living standards."

What do economists say?



Source: IGM Economic Experts Panel, July 8, 2013.

Case Study

The History of U.S. Government Debt

How indebted is the U.S. government? The answer to this question varies substantially over time. Figure 6 shows the debt of the U.S. federal government expressed as a percentage of U.S. GDP. It shows that the government debt has fluctuated from zero in 1836 to 106 percent of GDP in 1946.

The debt-to-GDP ratio is one gauge of the government's finances. Because GDP is a rough measure of the government's tax base, a declining debt-to-GDP ratio indicates that the government's indebtedness is shrinking relative to its ability to raise tax revenue. This suggests that the government is, in some sense, living within its means. By contrast, a rising debt-to-GDP ratio means that the government indebtedness is increasing relative to its ability to raise tax revenue. It is often interpreted as meaning that fiscal policy—government spending and taxes—cannot be sustained forever at current levels.

Throughout history, a primary cause of fluctuations in government debt has been war. When wars occur, government spending on national defense rises substantially to pay for soldiers and military equipment. Taxes sometimes rise as well but typically by much less than the increase in spending. The result is a budget deficit and increasing government debt. When the war is over, government spending declines, and the debt-to-GDP ratio starts declining as well.

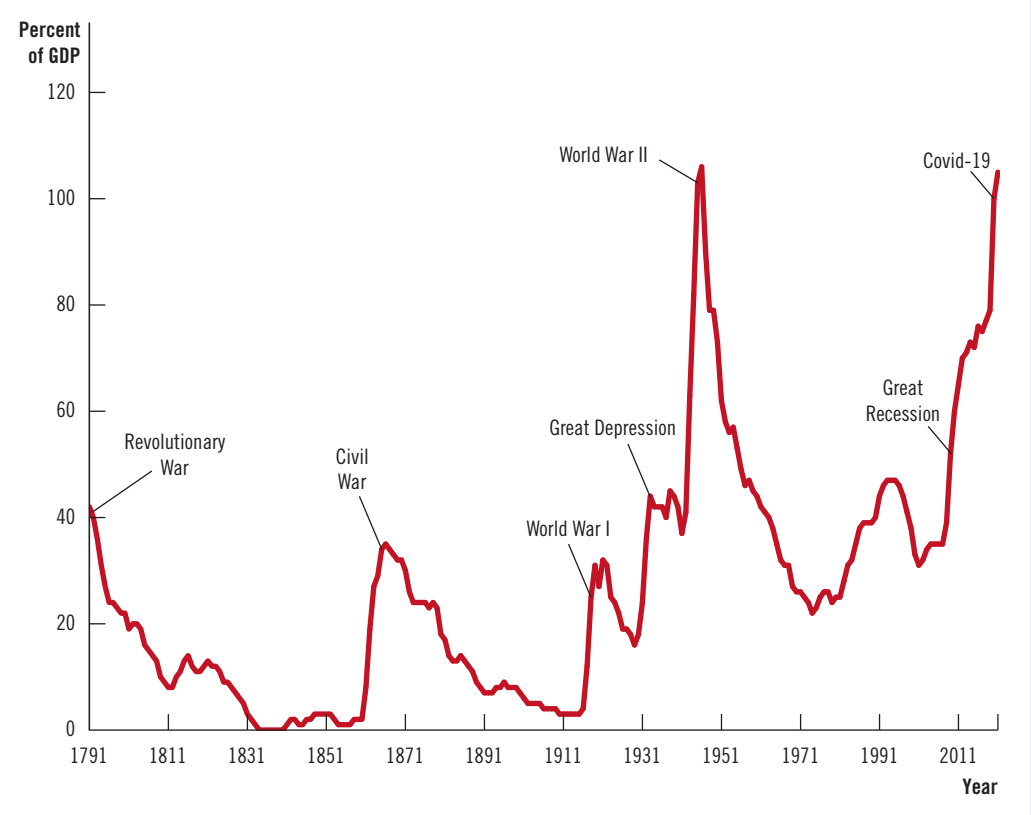
There are two reasons to believe that debt financing of war is an appropriate policy. First, it allows the government to keep tax rates smooth over time. Without debt financing, wars would require sharp increases in tax rates, which would cause a substantial decline in economic efficiency. Second, debt financing of wars shifts part of the cost of wars to future generations, who will have to pay off the government debt. Putting some of the tax burden on future generations is arguably fair if there is an enduring benefit from a war fought by a previous generation.

Figure 6

The U.S. Government Debt

The debt of the U.S. federal government, expressed here as a percentage of GDP, has varied throughout history. Wars and deep economic downturns are typically associated with substantial increases in government debt.

Source: U.S. Department of Treasury; U.S. Department of Commerce; and T. S. Berry, "Production and Population since 1789," Bostwick Paper No. 6, Richmond, 1988. The data here are for government debt held by the public, which excludes government debt held in government accounts, such as the Social Security trust fund.



Another notable cause of increases in government debt is deep economic downturns, such as the Great Depression of the 1930s, the Great Recession associated with the financial crisis of 2008–2009, and the Covid Recession of 2020. During downturns, government revenue automatically falls as receipts from income and payroll taxes decline. Spending on government programs such as unemployment insurance also rises automatically. In addition, policymakers usually enact policies to soften the downturn and mitigate the economic hardship, further increasing the budget deficit.

Going forward, many budget analysts are concerned about further increases in the debt-to-GDP ratio. As members of the large baby-boom generation reach retirement age, they will become eligible for Social Security and Medicare benefits, putting upward pressure on government spending. Without sizable increases in tax revenue or cuts in government spending, the U.S. federal government will likely experience substantially rising debt over the next few decades. According to a 2021 projection by the Congressional Budget Office, the debt-to-GDP ratio is on track to reach 140 percent by 2040, which is larger than anything that has been experienced historically. ●

FYI Financial Crises

In 2008 and 2009, the U.S. economy and many others around the world experienced a financial crisis, which led to a deep downturn in economic activity. We will examine these events in detail later in this book. For now, here are the key elements of typical financial crises.

The first element of a financial crisis is a large decline in the prices of some assets. In 2008 and 2009, that asset was real estate. House prices, after booming earlier in the decade, fell by about 30 percent over just a few years. Such a large decline in real estate prices had not been seen in the United States since the 1930s.

The second element of a financial crisis is widespread insolvencies at financial institutions. (A company is **insolvent** when its liabilities exceed the value of its assets.) In 2008 and 2009, many banks and other financial firms had, in effect, placed bets on house prices by holding mortgages backed by that real estate. When house prices plummeted, large numbers of homeowners stopped repaying their loans. These defaults pushed several major financial institutions toward bankruptcy.

The third element is a decline in confidence in financial institutions. Although some deposits in banks are insured by government policies, not all are. As insolvencies mounted in 2008 and 2009, every financial institution became a candidate for bankruptcy. Individuals and firms with uninsured deposits in those institutions pulled out their money. Needing cash to pay back depositors, banks started selling off assets (sometimes at reduced "fire-sale" prices) and cut back on new lending.

The fourth element is a credit crunch. With many financial institutions facing difficulties, prospective borrowers had trouble getting loans, even if they had profitable investment projects. In essence, the financial system had trouble directing the resources of savers into the hands of borrowers with the best investment opportunities.

The fifth element is an economic downturn. With people unable to obtain financing for new investment projects, the overall demand for goods and services declined. As a result, for reasons discussed more fully later in the book, national income fell, and unemployment rose.

The sixth and final element of a financial crisis is a vicious circle. The downturn reduced the profitability of many companies and the value of many assets, putting the economy back at step one. The problems in the financial system and the overall economy reinforced each other.

Financial crises, such as the one of 2008 and 2009, can have severe consequences. Fortunately, sooner or later, they end. The financial system eventually gets back on its feet, perhaps with help from government, and it returns to its normal function of financial intermediation. ■

QuickQuiz

6. If a popular TV show on personal finance convinces Americans to save more for retirement, the _____ curve for loanable funds would shift, driving the equilibrium interest rate _____.
   a. supply; up
   b. supply; down
   c. demand; up
   d. demand; down

7. If the business community becomes more optimistic about the profitability of capital, the _____ curve for loanable funds would shift, driving the equilibrium interest rate _____.
   a. supply; up
   b. supply; down
   c. demand; up
   d. demand; down

8. Which of the following policy actions would unambiguously reduce the supply of loanable funds and crowd out investment?
   a. an increase in taxes and a decrease in government spending
   b. a decrease in taxes together with an increase in government spending
   c. an increase in both taxes and government spending
   d. a decrease in both taxes and government spending

9. From 2019 to 2021, during the Covid-19 crisis, the ratio of government debt to GDP in the United States
   a. increased markedly.
   b. decreased markedly.
   c. was stable at a historically high level.
   d. was stable at a historically low level.

Answers are at the end of the chapter.

27-4 Conclusion

"Neither a borrower nor a lender be," Polonius advises his son in Shakespeare's *Hamlet*. If everyone followed Polonius's advice, this chapter would be unnecessary.

But few do. In our economy, people borrow and lend often, and usually for good reason. You may borrow one day to start your own business or buy a home. And people may lend to you in the hope that the interest you pay will allow them to enjoy a more prosperous retirement. The financial system's job is to coordinate all this borrowing and lending.

In many ways, financial markets are like others in the economy. The price of loanable funds—the interest rate—is governed by supply and demand, just as most other prices in the economy are. And we can analyze shifts in supply or demand in financial markets as we do in other markets. One of the **Ten Principles of Economics** in Chapter 1 is that markets are usually a good way to organize economic activity. This principle applies to financial markets as well. When financial markets bring the supply and demand for loanable funds into balance, they help allocate the economy's scarce resources to their most efficient uses.

In one way, however, financial markets are special. Financial markets, unlike most others, link the present and the future. Those who supply loanable funds—savers—do so because they want to convert some of their current income into future purchasing power. Those who demand loanable funds—borrowers—do so because they want to invest today and use the capital to produce goods and services in the future. Well-functioning financial markets are important not only for current generations but also for future generations who will inherit many of the benefits.

Chapter in a Nutshell

- The U.S. financial system is made up of many types of financial institutions, such as the bond market, the stock market, banks, and mutual funds. They all direct the resources of households that want to save some of their income into the hands of households and firms that want to borrow.
- National income accounting identities reveal important relationships among macroeconomic variables. In particular, for a closed economy, national saving must equal investment. Financial institutions are the mechanism through which the economy matches one person's saving with another person's investment.
- The interest rate is determined by the supply and demand for loanable funds. The supply comes from households that want to save some of their income and lend it out. The demand comes from households and firms that want to borrow for investment. To analyze how any policy or event affects the interest rate, consider how it affects the supply and demand for loanable funds.
- National saving equals private saving plus public saving. A government budget deficit represents negative public saving and, therefore, reduces national saving and the supply of loanable funds available to finance investment. When a government budget deficit crowds out investment, it reduces the growth of productivity and GDP.

Key Concepts

financial system, p. 554
financial markets, p. 554
bond, p. 554
stock, p. 555
financial intermediaries, p. 556
mutual fund, p. 557
national saving (saving), p. 559
private saving, p. 560
public saving, p. 560
budget surplus, p. 560
budget deficit, p. 560
market for loanable funds, p. 561
crowding out, p. 568

Questions for Review

1. What is the role of the financial system? Name and describe two markets that are part of the financial system in the U.S. economy. Name and describe two financial intermediaries.
2. Why is it important for people who own stocks and bonds to diversify their holdings? What type of financial institution makes diversification easier?
3. What is national saving? What is private saving? What is public saving? How are these three variables related?
4. What is investment? How is it related to national saving in a closed economy?
5. Describe a change in the tax code that might increase private saving. If this policy were implemented, how would it affect the market for loanable funds?
6. What is a government budget deficit? How does it affect interest rates, investment, and economic growth?

Problems and Applications

1. For each of the following pairs, which bond would you expect to pay a higher interest rate? Explain.
   a. a bond of the U.S. government or a bond of an Eastern European government
   b. a bond that repays the principal in year 2030 or a bond that repays the principal in year 2050
   c. a bond from Coca-Cola or a bond from a software company you run in your garage
   d. a bond issued by the federal government or a bond issued by New York State
2. Many workers hold large amounts of stock issued by the firms at which they work. Why do you suppose companies encourage this behavior? Why might people **not** want to hold stock in the company where they work?
3. Explain the difference between saving and investment as defined by a macroeconomist. Which of the following situations represent investment and which represent saving? Explain.
   a. Your family takes out a mortgage and buys a new house.
   b. You use your $200 paycheck to buy stock in AT&T.
   c. Your roommate earns $100 and deposits it in an account at a bank.
   d. You borrow $1,000 from a bank to buy a car to use in your pizza delivery business.
4. Suppose GDP is $8 trillion, taxes are $1.5 trillion, private saving is $0.5 trillion, and public saving is $0.2 trillion. Assuming this economy is closed, calculate consumption, government purchases, national saving, and investment.
5. Economists in Funlandia, a closed economy, have collected the following information about the economy for a particular year:

$$Y = 10,000$$
$$C = 6,000$$
$$T = 1,500$$
$$G = 1,700$$

   The economists also estimate that the investment function is:

$$I = 3,300 - 100r,$$

   where r is the country's real interest rate, expressed as a percentage. Calculate private saving, public saving, national saving, investment, and the equilibrium real interest rate.
6. Suppose that Intel is considering building a new chip-making factory.
   a. Assuming that Intel needs to borrow money in the bond market, why would an increase in interest rates affect Intel's decision about whether to build the factory?
   b. If Intel has enough of its own funds to finance the new factory without borrowing, would an increase in interest rates still affect Intel's decision about whether to build the factory? Explain.

7. Three students have each saved $1,000. Each has an investment opportunity in which he or she can invest up to $2,000. Here are the rates of return on the students' investment projects:

Harry	5 percent
Ron	8 percent
Hermione	20 percent

   a. If borrowing and lending are prohibited so each student can only use personal saving to finance his or her own investment project, how much will each student have a year later when the project pays its return?
   b. Now suppose their school opens up a market for loanable funds in which students can borrow and lend among themselves at an interest rate r. What would determine whether a student would choose to be a borrower or lender in this market?
   c. Among these three students, what would be the quantity of loanable funds supplied and quantity demanded at an interest rate of 7 percent? At 10 percent?
   d. At what interest rate would the loanable funds market among these three students be in equilibrium? At this interest rate, which student(s) would borrow and which student(s) would lend?
   e. At the equilibrium interest rate, how much does each student have a year later after the investment projects pay their return and loans have been repaid? Compare your answers with those you gave in part (a). Who benefits from the existence of the loanable funds market—the borrowers or the lenders? Is anyone worse off?

8. Suppose the government borrows $20 billion more next year than this year.
   a. Use a supply-and-demand diagram to analyze this policy. Does the interest rate rise or fall?
   b. What happens to investment? To private saving? To public saving? To national saving? Compare the size of the changes to the $20 billion of extra government borrowing.
   c. How does the elasticity of supply of loanable funds affect the size of these changes?
   d. How does the elasticity of demand for loanable funds affect the size of these changes?
   e. Suppose households believe that greater government borrowing today implies higher taxes to pay off the government debt in the future. What does this belief do to private saving and the supply of loanable funds today? Does it increase or decrease the effects you discussed in parts (a) and (b)?

9. This chapter explains that investment can be increased both by reducing taxes on private saving and by reducing the government budget deficit.
   a. Why is it difficult to implement both of these policies at the same time?
   b. What would you need to know about private saving to determine which of these two policies would be the more effective way to raise investment?

QuickQuiz Answers

1. **d** 2. **d** 3. **b** 4. **d** 5. **c** 6. **b** 7. **c** 8. **b** 9. **a**

Chapter

28

The Basic Tools of Finance

At some point, the financial system is bound to be important in your life. You may deposit your savings in a bank account, or you may take out a loan to cover tuition or buy a house. After you have a job, your employer may start a retirement account for you, and you'll be able to decide whether to invest the funds in stocks, bonds, or other financial instruments. If you invest on your own, you can choose whether to put your money in established companies such as Coca-Cola, newer ones such as Airbnb, or the entire stock market through an index fund. And in the media, you will hear reports about whether stocks are up or down, along with the often-feeble attempts to explain why the market behaves as it does.

In almost all the financial decisions you will make during your life, you will encounter two related elements: time and risk. As the preceding two chapters showed, the financial system coordinates the economy's saving and investment, which are crucial determinants of growth. Most fundamentally, the financial system concerns decisions and actions undertaken today that will affect our lives in an unknown future. When people allocate their saving and when

BLACKCAT/E+/GETTY IMAGES; GOWITHSTOCK/SHUTTERSTOCK.COM

firms pick investment projects, their decisions are based on guesses about the likely results. The actual outcome could end up very different from what was expected.

finance
the field that studies how people make decisions regarding the allocation of resources over time and the handling of risk

This chapter introduces some tools for understanding the financial decisions that people make. The field of **finance** develops these tools in detail, and you may later take courses that focus on this topic. But because the financial system is so important, many of the insights of finance are central to understanding how the economy works. The tools of finance can also help you think through some of the decisions that you will make in life.

This chapter takes up three topics: how to compare sums of money from different points in time, how to manage risk, and how to determine the value of an asset such as a share of stock.

28-1 Present Value: Measuring the Time Value of Money

Imagine that someone offers to give you $100 today or $100 in 10 years. Which would you choose? This question is easy. Getting $100 today is better because you can deposit the money in a bank, still have it in 10 years, and earn interest on the $100 along the way. The lesson: Money today is more valuable than the same amount of money in the future.

Now consider a harder question. Imagine that someone offers you $100 today or $200 in 10 years. Which would you choose? To answer this question, you need some way to compare sums of money from different points in time. Economists do this with a concept called present value. The **present value** of any future sum of money is the amount of money that, given current interest rates, would be needed today to produce that future sum.

present value
the amount of money today needed to produce a future amount of money, given prevailing interest rates

To learn how to use the concept of present value, let's work through a couple of simple examples:

Question: If you put $100 in a bank account today, how much will it be worth in N years? That is, what will be the **future value** of this $100?

future value
the amount of money in the future that an amount of money today will yield, given prevailing interest rates

Answer: Use r to denote the interest rate expressed in decimal form (so an interest rate of 5 percent means $r = 0.05$). Suppose that interest is paid annually and that it remains in the bank account to earn more interest—a process called **compounding**. Then the $100 will become

compounding
the accumulation of a sum of money in, say, a bank account, where the interest earned remains in the account to earn additional interest in the future

$(1 + r) \times \$100$	after 1 year,
$(1 + r) \times (1 + r) \times \$100 = (1 + r)^2 \times \$100$	after 2 years,
$(1 + r) \times (1 + r) \times (1 + r) \times \$100 = (1 + r)^3 \times \$100$	after 3 years, …
$(1 + r)^N \times \$100$	after N years.

For example, if we invest at an interest rate of 5 percent for 10 years, then the future value of the $100 will be $(1.05)^{10} \times \$100$, or $163.

Question: Now suppose you are going to be paid $200 in N years. What is the **present value** of this future payment? That is, how much would you have to deposit in a bank right now to yield $200 in N years?

Answer: To answer this question, turn the previous answer on its head. In the last question, we computed a future value from a present value by **multiplying** by the factor $(1 + r)^N$. To compute a present value from a future value, **divide** by the factor $(1 + r)^N$. The present value of $200 to be paid in N years is $\$200/(1 + r)^N$. If that amount is deposited in a bank today, after N years it will become $(1 + r)^N \times [\$200/(1 + r)^N]$,

which equals $200. For instance, if the interest rate is 5 percent, the present value of $200 to be paid in 10 years is $200/(1.05)^{10}, or $123. This means that $123 deposited today in a bank account earning 5 percent interest would be worth $200 after 10 years.

This illustrates the general formula:

- If r is the interest rate, then an amount X to be received in N years has a present value of $X/(1 + r)^N$.

Because the possibility of earning interest reduces the present value below the amount X, the process of finding the present value of a future sum of money is called **discounting**. This formula shows precisely how much future sums should be discounted.

Let's now return to our initial question: Should you choose $100 today or $200 in 10 years? Based on the present value calculation using an interest rate of 5 percent, you should prefer the $200 in 10 years. The future $200 has a present value of $123, which is greater than $100. You are better off waiting for the future sum.

Notice that the answer to our question depends on the interest rate. If the interest rate were 8 percent, then the $200 in 10 years would have a present value of $200/(1.08)^{10}, which is only $93. In this case, you should take the $100 today. Why should the interest rate matter? Because the higher the interest rate, the more you can earn by depositing your money in a bank, so the more attractive getting $100 today becomes.

At this point, a curious reader might ask: Which interest rate should be used in these calculations—the real or the nominal? These differ when the price level is changing. The answer is, it depends. If the future sum being discounted is stated in nominal terms (that is, as a specified number of dollars), the nominal interest rate is called for. But if the future sum is stated in real terms (that is, as a specified number of inflation-adjusted dollars, measured in terms of today's prices), the real interest rate is the right one to use. Discounting and present value apply in both cases, but the calculation must be done consistently.

The concept of present value is useful in many applications, including the decisions that companies face when evaluating investment projects. For instance, imagine that General Motors is thinking about building a new factory. Suppose that the factory will cost $100 million today and will yield the company $200 million in 10 years. Should General Motors undertake the project? This decision is exactly like the one we have been studying. To make its choice, the company should compare the present value of the $200 million return with the $100 million cost.

The decision, therefore, will depend on the interest rate. If the interest rate is 5 percent, then the present value of the $200 million return from the factory is $123 million, and the company should choose to pay the $100 million cost. But if the interest rate is 8 percent, then the present value of the return is only $93 million, and the company should forgo the project. The concept of present value helps explain why investment—and the quantity of loanable funds demanded—declines when the interest rate rises.

Here is another application of present value: Suppose you win a million-dollar lottery and are given a choice between $20,000 a year for 50 years (totaling $1,000,000) or an immediate payment of $500,000. Which is more valuable? You need to calculate the present value of the stream of payments. Suppose the interest rate is 5 percent. After performing 50 calculations like those above (one for each payment) and adding up the results, you would learn that the present value of this million-dollar prize at a 5 percent interest rate is only $383,000. The immediate payment of $500,000 is the better choice. The million dollars may seem like more money, but the future cash flows, once discounted to the present, are worth far less.

FYI

The Magic of Compounding and the Rule of 70

Suppose one country has an average growth rate of 1 percent per year, while another has an average growth rate of 3 percent per year. At first, this gap might not seem like a big deal. What difference can 2 percent make?

The answer is: a big difference. Growth rates that seem small when written in percentage terms are large after they are compounded for many years.

Consider an example. Suppose that two college graduates—William and Sarah—both take their first jobs at the age of 22 with an annual salary of $50,000. William lives in an economy where all incomes grow at 1 percent per year, and Sarah lives in one where incomes grow at 3 percent per year. Straightforward calculations show what happens. Forty years later, when both are 62 years old, William earns $74,000 a year, while Sarah earns $163,000. Because of that difference of 2 percentage points in the growth rate, Sarah's salary is more than twice William's.

An old rule of thumb, called the **rule of 70**, clarifies the effects of compounding. According to the rule of 70, if some amount grows at a rate of x percent per year, then that amount doubles in approximately $70/x$ years. In William's economy, incomes grow at 1 percent per year, so it takes about 70 years for incomes to double. In Sarah's economy, incomes grow at 3 percent per year, so it takes about 70/3, or 23, years for incomes to double.

The rule of 70 applies not only to a growing economy but also to a growing savings account. Here is an example: In 1791, Ben Franklin died and left $5,000 to be invested for a period of 200 years to benefit medical students and scientific research. If this money had earned 7 percent per year (which would, in fact, have been possible), the investment would have doubled in value every 10 years. Over 200 years, it would have doubled 20 times. At the end of 200 years of compounding, the investment would have been worth $2^{20} \times \$5{,}000$, which is about $5 billion. (In fact, Franklin's $5,000 grew to only $2 million over 200 years because some money was spent along the way.)

Franklin understood compounding, which he described as follows: "Money makes money. And the money that money makes, makes money." If this process continues for many years, it can lead to spectacular results. ■

QuickQuiz

1. If the interest rate is zero, then $100 to be paid in 10 years has a present value that is
   a. less than $100.
   b. exactly $100.
   c. more than $100.
   d. indeterminate.
2. If the interest rate is 10 percent, then the future value in 2 years of $100 today is
   a. $80.
   b. $83.
   c. $120.
   d. $121.
3. If the interest rate is 10 percent, then the present value of $100 to be paid in 2 years is
   a. $80.
   b. $83.
   c. $120.
   d. $121.

Answers are at the end of the chapter.

28-2 Managing Risk

Life is full of gambles. When you go skiing, you risk breaking your leg. When you drive somewhere, you risk getting into a car accident. When you put some of your savings in the stock market, you risk losing your money from a fall in stock prices. The rational response to risk is not to avoid it at any cost but to take it into account in your decision making. Let's consider how you might do that as you make financial decisions.

28-2a Risk Aversion

Most people are **risk averse**. This means more than that people dislike bad things happening to them. It means that they dislike bad things more than they like comparable good things.

risk aversion
a dislike of uncertainty

For example, suppose your friend Felicity offers you the following deal. She will toss a coin. If it comes up heads, Felicity will pay you $1,000. But if it comes up tails, you will have to pay her $1,000. Would you accept the bargain? You wouldn't if you were risk averse. For a risk-averse person, the pain of losing the $1,000 would exceed the pleasure of winning $1,000. Because the two outcomes are equally likely, you would be worse off on average.

Economists have developed models of risk aversion using the concept of **utility**, which is a person's subjective measure of well-being or satisfaction. As the utility function in Figure 1 shows, every level of wealth provides a certain amount of utility. But the utility function gets flatter as wealth increases, meaning it exhibits the property of diminishing marginal utility: The more wealth a person has, the less utility each additional dollar yields. Because of this, the utility forfeited from losing the $1,000 bet exceeds the utility gained from winning it. In other words, diminishing marginal utility is the reason most people are risk averse.

Risk aversion helps explain many things in the economy. Let's consider three of them: insurance, diversification, and the risk-return trade-off.

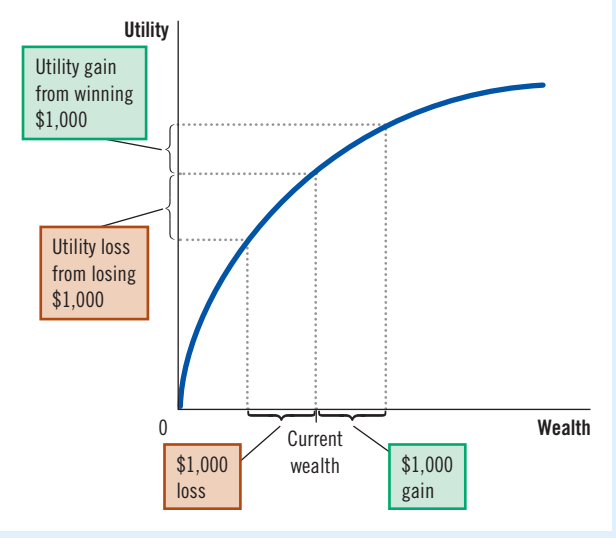
28-2b The Markets for Insurance

Buying insurance is one way to deal with risk. With insurance contracts, a person facing a risk pays a fee to an insurance company, which agrees to accept all or part of the risk. Insurance takes many forms: Car insurance covers the risk of an auto accident, fire insurance covers the risk of your house burning down, health insurance covers the risk of needing expensive medical treatment, and life insurance covers the risk of dying young and leaving your family without enough money. There is also insurance against the risk of outliving your resources: For a fee paid today, an insurance company will pay you an **annuity**—a regular income for the rest of your life.

Figure 1

The Utility Function

Utility, a subjective measure of satisfaction, depends on wealth. As wealth rises, the utility function becomes flatter, reflecting the property of diminishing marginal utility. Because of diminishing marginal utility, a $1,000 loss decreases utility by more than a $1,000 gain increases it.



In a sense, every insurance contract is a gamble. It is possible that you will not be in an auto accident, that your house will not burn down, and that you will not need expensive medical treatment. In most years, you will pay the insurance company the premium and get nothing in return except peace of mind. The insurance company is counting on most people not making claims on their policies; otherwise, it couldn't pay out large claims to the unlucky few and still stay in business.

For the overall economy, the role of insurance is not to eliminate the risks inherent in life but to spread them around more efficiently. Consider fire insurance. Buying it does not reduce the risk of losing your home in a fire. But if that unlucky event occurs, the insurance company compensates you. The risk, rather than being borne by you alone, is shared among thousands of insurance-company shareholders.

The markets for insurance suffer from two problems that impede their ability to spread risk. One problem is **adverse selection**: High-risk people benefit more from insurance protection and, therefore, are more likely to buy insurance than low-risk people. A second problem is **moral hazard**: After people buy insurance, they have less incentive to be careful about risky behavior because the insurance company will cover much of their losses. Insurance companies are aware of these pitfalls, but they cannot entirely avoid them. An insurance company cannot perfectly distinguish between high-risk and low-risk customers, and it cannot monitor all of its customers' risky behaviors. The price of insurance reflects the expected risks that the insurance company will face after the insurance is bought. The high price of insurance is why some people, especially those who know themselves to be low risk, decide against buying it and, instead, endure life's uncertainty on their own.

28-2c Diversification of Firm-Specific Risk

In 2001, Enron, a large and once widely respected company, went bankrupt amid accusations of fraud and accounting irregularities. Several of the company's top executives were prosecuted and ended up in prison. The saddest part of the story, however, involved Enron's thousands of lower-level employees. Not only did they lose their jobs, but many lost their life savings as well. The employees had put about two-thirds of their retirement funds in Enron stock, which became worthless.

If there is one piece of practical advice that finance offers risk-averse people, it is the following: "Don't put all your eggs in one basket." You may have heard this folk wisdom before, but finance has turned it into a science. It is called **diversification**.

diversification
the reduction of risk achieved by replacing a single risk with a large number of smaller, imperfectly correlated risks

The market for insurance is an example of diversification. Imagine a town with 10,000 homeowners, each facing the risk of a house fire. If someone starts an insurance company and each person becomes both a shareholder and a policyholder of the company, they all reduce their risk through diversification. It is better to face 1/10,000 of the risk of 10,000 possible fires rather than the entire risk of a single fire in your own home. Unless the entire town catches fire at once, the downside that each person faces is much smaller.

When people use their savings to buy financial assets, they can reduce risk through diversification. A person who buys stock in a company is betting on that company's future profitability. That bet is inherently risky because it is hard to predict the long-term profitability of any single company. Microsoft evolved from a start-up by some geeky teenagers into one of the world's most valuable companies in only a few years; Enron went from one of the world's most respected companies to an almost worthless one in only a few months. Happily, shareholders need not tie their own fortunes to that of any single company. Risk can be reduced by placing many small, imperfectly correlated bets rather than a small number of large ones.

Figure 2 shows how the risk of a stock portfolio depends on how many stocks it contains. Risk is measured here by a statistic called the **standard deviation**, which you may have learned about in a math or statistics class. The standard deviation measures the volatility of a variable—that is, how much the variable can be expected to fluctuate. When the standard deviation of a portfolio's return is large, the return is likely to be volatile, and people holding the portfolio face a sizable risk that the return will fall well below what they expected.

The figure shows that the risk of a stock portfolio falls substantially as the number of stocks increases. For a portfolio with a single stock, the standard deviation is 49 percent. Going from 1 stock to 10 stocks eliminates about half the risk. Going from 10 stocks to 20 stocks reduces the risk by another 10 percent. As the number of stocks continues to increase, the risk continues to fall, although the reductions in risk beyond 20 to 30 stocks are small.

Notice that it is impossible to eliminate all risk by increasing the number of stocks in the portfolio. Diversification can eliminate **firm-specific risk**—the uncertainty associated with a specific company. But it cannot eliminate **market risk**—the uncertainty associated with the entire stock market. For example, when the economy goes into a recession, most companies experience falling sales, reduced profit, and lower stock returns. Diversification reduces the risk of holding stocks but does not eliminate it.

firm-specific risk
risk that affects only a single company

market risk
risk that affects all companies in the stock market

28-2d The Trade-Off between Risk and Return

One of the **Ten Principles of Economics** in Chapter 1 is that people face trade-offs. The most relevant of these for understanding financial decisions is the trade-off between risk and return.

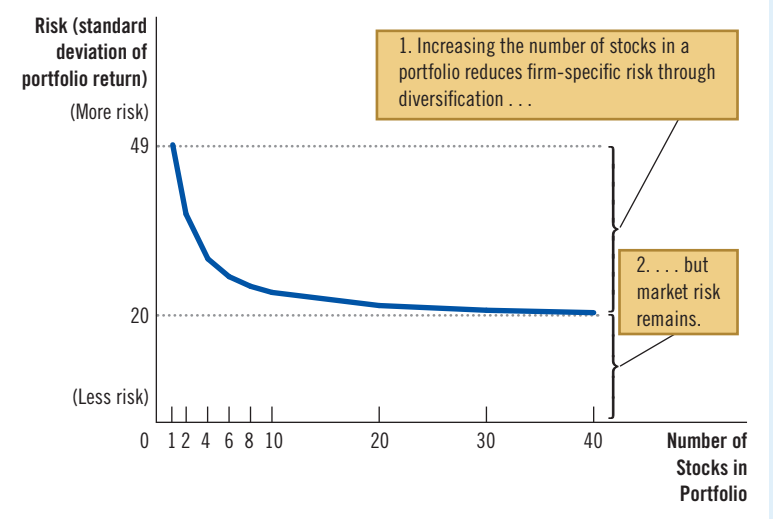
As we have seen, holding stocks is inherently risky, even in a diversified portfolio. But risk-averse people are willing to accept this uncertainty because they are compensated for doing so. Historically, stocks have offered much higher rates of return than alternative financial assets such as bonds and bank savings accounts.

Figure 2

Diversification Reduces Risk

This figure shows how the risk of a portfolio, measured here by a statistic called the **standard deviation**, depends on the number of stocks in the portfolio. Investors are assumed to put an equal amount in each of the stocks. Increasing the number of stocks reduces but does not eliminate the portfolio risk.

Source: Adapted from Meir Statman, "How Many Stocks Make a Diversified Portfolio?" *Journal of Financial and Quantitative Analysis* 22 (September 1987): 353–364.



Over the past two centuries, stocks have generated an average real return of about 8 percent per year, while short-term government bonds have paid a real return of only 3 percent per year.

When allocating their savings, people choose how much risk they are willing to accept to earn a higher expected return. For example, consider a person allocating a portfolio between two asset classes:

- The first asset class is a diversified group of risky stocks with an average annual return of 8 percent and a standard deviation of 20 percent. You may recall from a math or statistics class that a normal random variable stays within 2 standard deviations of its average about 95 percent of the time. Here, 2 standard deviations mean fluctuations of ±40 percent. Thus, while returns are centered around 8 percent, they vary between a 48 percent gain to a 32 percent loss 95 percent of the time. (And 5 percent of the time, the gain or loss is larger.)
- The second asset class is a safe alternative with an annual return of 3 percent and a standard deviation of zero. That is, this asset always pays exactly 3 percent. The safe alternative can be either a bank savings account or a government bond.

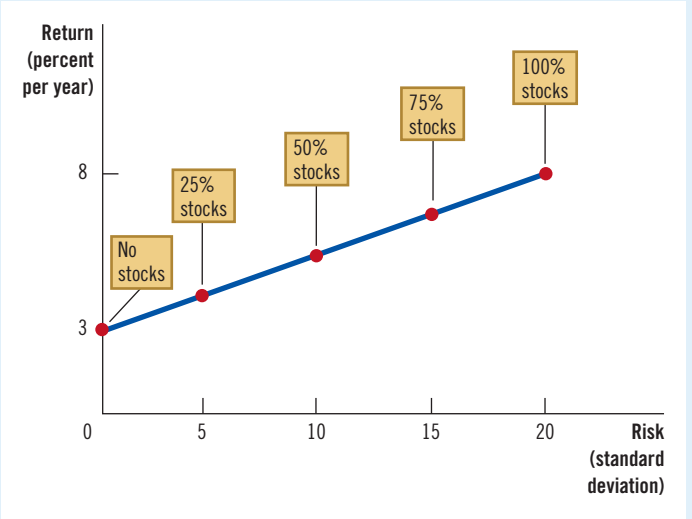
Figure 3 illustrates the trade-off between risk and return. Each point in this figure represents a particular allocation of a portfolio between risky stocks and the safe asset. The figure shows that the more a person puts into stocks, the greater both the risk and the return.

Acknowledging the risk-return trade-off does not, by itself, tell us what a person should do. The choice of a particular combination of risk and return depends on a person's risk aversion, which is a matter of individual preference. But it is important for stockholders to recognize that the higher expected return comes at the price of higher risk.

Figure 3

The Trade-Off between Risk and Return

When people increase the percentage of their savings invested in stocks, they increase the return they can expect to earn along with the risks they can expect to face.



QuickQuiz

4. The ability of insurance to spread risk is limited by
   a. risk aversion and moral hazard.
   b. risk aversion and adverse selection.
   c. moral hazard and adverse selection.
   d. risk aversion only.
5. The benefit of diversification when constructing a portfolio is that it can eliminate
   a. adverse selection.
   b. risk aversion.
   c. firm-specific risk.
   d. market risk.
6. The extra return that stocks earn over bonds (on average) compensates stockholders for
   a. the greater market risk that stockholding entails.
   b. the greater firm-specific risk that stockholding entails.
   c. the higher taxes levied on stockholders.
   d. the higher brokerage costs incurred buying stocks.

Answers are at the end of the chapter.

28-3 Asset Valuation

With a basic understanding of the two building blocks of finance—time and risk—let's apply this knowledge. This section considers a simple question: What determines the price of a share of stock? As for most prices, the answer is supply and demand. But that is not the end of the story. To understand stock prices, we need to think more deeply about what determines a person's willingness to pay for a share of stock.

28-3a Fundamental Analysis

Imagine that you have decided to put 60 percent of your savings into stock and that, to achieve diversification, you want to buy 20 different stocks. If you check the stock listings, you will find thousands of stocks. How should you pick the 20 for your portfolio?

When you buy stock, you are buying shares in a business. To decide which businesses you want to own, it is natural to consider two things: the value of that share of the business and the price at which the shares are being sold. If the price is more than the value, the stock is said to be **overvalued**. If the price and the value are equal, the stock is said to be **fairly valued**. And if the price is less than the value, the stock is said to be **undervalued**. Undervalued stocks are a bargain because you pay less than the business is worth. When choosing 20 stocks for your portfolio, you should look for undervalued stocks.

But that is easier said than done. Learning the price of the company's stock is easy: You can just look it up. Determining the value of the company is the hard part. The term **fundamental analysis** refers to the detailed study of a company to estimate its value. Many Wall Street firms hire stock analysts to conduct fundamental analysis and offer advice about which stocks to buy.

fundamental analysis the study of a company's accounting statements and future prospects to determine its value

The value of a stock to stockholders is what they get from owning it, which includes the present value of the stream of dividend payments and the final sale price. Recall that **dividends** are the cash payments that a company makes to its shareholders. A company's ability to pay dividends, as well as the value of the stock when stockholders sell their shares, depends on the company's ability to earn profits. Its profitability, in turn, depends on many factors: the demand for its product, the amount and kinds of capital it has in place, the degree of competition it confronts, the extent of unionization of its workers, the loyalty of its customers,

FYI

Key Numbers for Stock Watchers

When following the stock of a company, you should keep an eye on three key numbers. These numbers are reported in some newspapers and many online services:

- **Price.** The single most important piece of information about a stock is the price of a share. News services usually present several prices. The "last" price is the price at which the stock more recently traded. The "previous close" is the price of the last transaction that occurred by the close of the previous day of trading. A news service may also give the "high" and "low" prices over the past day of trading and, sometimes, over the past week, month, or year. It may also report the change from the previous day's closing price.
- **Dividend.** Corporations pay out some of their profits to their stockholders; this amount is called the **dividend**. (Profits not paid out are called **retained earnings**, which corporations can use to finance capital investments, increase cash reserves, or buy back their own shares.) News services often report the dividend paid over the previous year for each share of stock. They sometimes report the **dividend yield**, which is the dividend expressed as a percentage of the stock's price.
- **Price-earnings ratio.** A corporation's earnings, or accounting profit, is the amount of revenue it receives for the sale of its products minus its costs of production as measured by its accountants. **Earnings per share** is the company's total earnings divided by the number of shares of stock outstanding. The **price-earnings ratio**, often called the P/E, is the price of one share of a corporation's stock divided by the corporation's earnings per share over the past year. Historically, price-earnings ratios have averaged about 16, though they often deviate substantially from this norm. A high P/E indicates that a corporation's stock is expensive relative to its recent earnings, suggesting either that people expect earnings to rise in the future or that the stock is overvalued. Conversely, a low P/E indicates that a corporation's stock is cheap relative to its recent earnings, suggesting either that people expect earnings to fall or that the stock is undervalued.

Why do financial services report all these data? Many people who invest their savings in stock follow these numbers closely when deciding which stocks to buy and sell. By contrast, other stockholders follow a buy-and-hold strategy: They buy the stock of well-run companies, hold it for long periods, and do not respond to daily fluctuations. ■

the government regulations and taxes it faces, and so on. The goal of fundamental analysis is to take these factors into account to determine how much a share of stock in the company is worth.

If you want to rely on fundamental analysis to pick a stock portfolio, there are three ways to do it. One way is to do all the necessary research yourself by, for instance, reading the companies' annual reports. A second way is to rely on the advice of Wall Street analysts. A third is to buy shares in a mutual fund, which has a manager who makes decisions for you.

28-3b The Efficient Markets Hypothesis

There is another way to choose 20 stocks for your portfolio: Pick them randomly by, say, putting the stock pages on your bulletin board and throwing darts at them. This approach may sound crazy, but there is reason to believe that it won't lead you too far astray. That reason is called the **efficient markets hypothesis**.

efficient markets hypothesis
the theory that asset prices reflect all publicly available information about the value of an asset

To understand this theory, the starting point is to realize that each company listed on a major stock exchange is followed closely by numerous portfolio managers who run mutual funds, pension funds, and university endowments and who are actively seeking the best return. Every day, these managers monitor news stories and conduct fundamental analysis to try to determine a stock's value. Their job is to buy a stock when its price falls below its fundamental value and to sell it when its price rises above its fundamental value.

The second piece to the efficient markets hypothesis is that the equilibrium of supply and demand sets the market price. This means that, at the market price, the

number of shares being offered for sale exactly equals the number of shares that people want to buy. In other words, at the market price, the number of people who think the stock is overvalued exactly balances the number who think it's undervalued. As judged by the typical person in the market, then, all stocks are fairly valued all the time.

According to this theory, the stock market exhibits **informational efficiency**: It reflects all available information about the value of an asset. Stock prices change when information changes. When good news about a company's prospects becomes known, the company's value and stock price both rise. When a company's prospects deteriorate, its value and price both fall. But at any moment, the market price is the best guess of the company's value based on available information.

informational efficiency
the description of asset prices that rationally reflect all available information

One implication of the efficient markets hypothesis is that stock prices should follow a **random walk**, meaning that, unless you have inside information, changes in stock prices should be impossible to predict. If, based on publicly available information, a person could predict that a stock price would rise by 10 percent tomorrow, the stock market must be failing to incorporate that information today. According to the theory, the only thing that can move a company's stock price is news that changes the market's perception of the company's value. But news is inherently unpredictable—otherwise, it wouldn't really be news. As a result, changes in stock prices should be unpredictable as well.

random walk
the path of a variable whose changes are impossible to predict

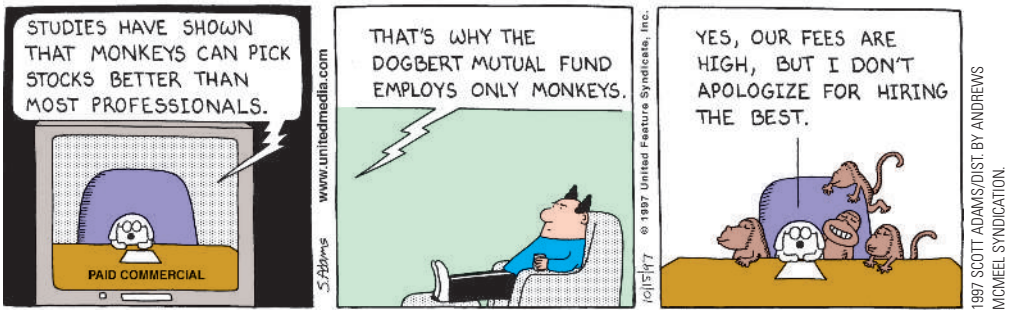
If the efficient markets hypothesis is correct, then there is little point in spending many hours studying the business page to decide which 20 stocks to add to your portfolio. If prices reflect all available information, no stock is a better buy than any other. The best you can do is to buy a diversified portfolio.

Case Study

Random Walks and Index Funds

The efficient markets hypothesis is a theory about how financial markets work. The theory may not be completely true: As the next section explains, there is reason to doubt that stockholders are always rational and that stock prices are informationally efficient at every moment. Nonetheless, the efficient markets hypothesis describes the world much better than you might expect.

There is much evidence that stock prices follow, even if not exactly a random walk, something close to it. For example, you might be tempted to buy stocks that have recently risen and avoid stocks that have recently fallen (or perhaps just the opposite). But statistical studies have shown that following such trends (or bucking them) fails to systematically outperform the market. The correlation between how well a stock does one year and how well it does the following year is about zero.



Some of the best evidence for the efficient markets hypothesis comes from the performance of index funds. An index fund is a mutual fund that buys all the stocks in a stock index. The performance of these funds can be compared with that of actively managed mutual funds, where professional portfolio managers pick stocks based on extensive research and supposed expertise. In essence, index funds buy all stocks and offer investors the return on the average stock, while actively managed funds seek to buy only the best stocks and outperform the market averages.

In practice, however, active managers usually fail to beat index funds. For example, in the 15-year period ending December 31, 2021, 86 percent of stock mutual funds performed worse than a broadly based index fund holding all stocks traded on U.S. stock exchanges. Over this period, the average annual return on stock funds fell short of the return on the index fund by 1.07 percentage points. Most active portfolio managers trail the market because they trade more frequently, incurring more trading costs, and because they charge greater fees to compensate for their research and expertise.

What about the 14 percent of managers who did beat the market? Perhaps they are smarter than average, or perhaps they were luckier. If 5,000 people each flip a coin ten times, on average, about five people will flip ten heads. These five might claim an exceptional coin-flipping skill, but they would have trouble replicating the feat. Similarly, studies have shown that mutual

Ask the Experts

Diversified Investing

"In general, absent any inside information, an equity investor can expect to do better by holding a well-diversified, low-fee, passive index fund than by holding a few stocks."

What do economists say?

Source: IGM Economic Experts Panel, January 28, 2019.

In the News

The Perils of Investing with a Y Chromosome

Overconfidence is bad when investing, and this problem is especially common among men.

Women May Be Better Investors Than Men. Let Me Mansplain Why.

By Ron Lieber

Merrill was a guy, and so was Lynch. Goldman? A dude, and Sachs as well. Charles Schwab is a man, and so was E.F. Hutton. Gordon Gekko was an alpha male. And Jordan Belfort, the Wolf of Wall Street? Total bro.

Heroes or villains, winners or losers, real or imagined, our iconic investors are very, very male. But that's a mistake—because it turns out that women are often better at investing.

Fidelity offered up the latest evidence this month: Over a 10-year period, its female customers earned, on average, 0.4 percentage points more annually than their male counterparts. That may not seem like a lot, but over a few decades, it can add up to tens of thousands of dollars or more.

"Invest like a woman is what you learn from this," said Lorna Kapusta, head of women investors and customer engagement at Fidelity.

This isn't the first time that researchers have found women to be the better investors. The surprising thing about this phenomenon, however, is that neither women nor men seem to be aware of it—and they end up depriving themselves of some lessons that might help both genders invest better.

Fidelity's analysis covered 5.2 million customer accounts (some people had more than one), from 2011 to 2020. It looked at individual retirement accounts, 529 plans and basic brokerage accounts that individuals (not financial advisers) controlled, but not workplace accounts like 401(k)s. No strategies were excluded: Those who traded individual stocks were tracked along with those who stuck to mutual funds.

The source of women's superior returns is the way they trade. Or, rather, how they don't.

fund managers with a history of superior performance usually fail to maintain it in subsequent periods.

The efficient markets hypothesis says that it is impossible to beat the market over the long haul. Many studies confirm that beating the market is, at best, extremely difficult. The efficient markets hypothesis may not describe the world perfectly, but it offers a useful perspective. ●

28-3c Market Irrationality

The efficient markets hypothesis assumes that people buying and selling stock rationally process the information they have about the stock's underlying value. But are participants in the stock market really that rational? Or do stock prices sometimes deviate from reasonable expectations of their true value?

There is a long tradition suggesting that fluctuations in stock prices are partly psychological. In the 1930s, the economist John Maynard Keynes suggested that asset markets are driven by the "animal spirits" of investors—irrational waves of optimism and pessimism. In the 1990s, as the stock market soared to new heights, Fed Chair Alan Greenspan questioned whether the boom reflected "irrational exuberance." Stock prices did subsequently fall sharply, but whether the exuberance of the 1990s was irrational given the information available at the time remains open to debate. When the price of an asset rises far above what appears to be its fundamental value, the market is said to be experiencing a **speculative bubble**.

The possibility of speculative bubbles in the stock market arises in part because the value of the stock to a stockholder depends not only on the stream of dividend

Female Fidelity customers bought and sold half as much as male customers. Vanguard saw similar patterns over the same decade-long period when examining workplace retirement accounts that it manages; at least 50 percent more men traded in them than women did every year during that time.

This is very bad. In a now classic paper that appeared in The Journal of Finance in 2000, titled "Trading Is Hazardous to Your Wealth," two professors, Brad M. Barber and Terrance Odean, proved just that. From 1991 to 1996, individual investors who traded the most earned an annual return that was 6.5 percentage points worse than the overall performance of the stock market.

The following year, the two professors tackled trading and gender in a different paper called "Boys Will Be Boys." Sure, women traded more than they should too, and from 1991 to 1997, their trading reduced their net returns by 1.72 percentage points per year. But the even more frequent buying and selling men engaged in caused them to take a 2.65 percentage point hit—more than twice the male underperformance that Fidelity found years later.

Why do men trade too much? Professors Barber and Odean chalked it up to overconfidence. And where does overconfidence come from? William J. Bernstein, a neurologist who turned his attention to investing years ago, points to testosterone.

The hormone causes three problems for investors: It decreases fear, increases greed and very much contributes to overconfidence. "It does wonderful things for muscle mass and reflex time but doesn't do much for judgment," he said.

If you fear too little, you're more likely to get hit hard when markets fall, since you'll have too much money in the wrong kinds of investments. Similarly, too much greed can lead to too much risk.

As for overconfidence, Mr. Bernstein, who is the author of books, including "The Investor's Manifesto," suggests a self-administered test question: How certain am I of what I'm doing? "In finance, if you're certain of anything, you're out of your mind," he said. ■

Source: *New York Times*, October 29, 2021.

Questions to Discuss

1. In your personal experience, do you find that men are more likely than women to be overconfident?
2. How might investors protect themselves from their own overconfidence?

payments but also on its final sale price. Some people might be willing to pay more than a stock is worth today if they expect others to pay even more for it tomorrow. From this perspective, when evaluating a stock, you are less concerned about the true value of the business than about what other people will think the business is worth in the future.

Economists disagree about the frequency and importance of departures from rational pricing. Believers in market irrationality point out (correctly) that the stock market often moves in ways that are hard to explain based on news that might alter a rational valuation. Believers in the efficient markets hypothesis point out (correctly) that it is hard to know the true, rational valuation of a company, so one should not quickly jump to the conclusion that any particular valuation is irrational. Moreover, if the market were irrational, a rational person might be able to take advantage of this fact and beat the market; yet as the previous case study discussed, beating the market over sustained periods is nearly impossible.

QuickQuiz

7. The goal of fundamental analysis is to
   a. determine the true value of a company.
   b. put together a diversified portfolio.
   c. predict changes in investor irrationality.
   d. eliminate investor risk aversion.
8. According to the efficient markets hypothesis,
   a. excessive diversification can reduce an investor's expected portfolio returns.
   b. changes in stock prices are impossible to predict from public information.
   c. actively managed mutual funds should generate higher returns than index funds.
   d. the stock market moves based on the changing animal spirits of investors.
9. Historically, index funds have had _____ than most actively managed mutual funds.
   a. higher fees
   b. less diversification
   c. larger tax burdens
   d. better returns

Answers are at the end of the chapter.

28-4 Conclusion

This chapter has developed some of the basic tools that people should (and often do) use to make financial decisions. The concept of present value tells us that a dollar tomorrow is less valuable than a dollar today, and it provides a way to compare sums of money at different points in time. The theory of risk management says that the future is uncertain and that risk-averse people can take precautions to guard against this uncertainty. The study of asset valuation tells us that the stock price of any company should reflect its expected future profitability.

Most tools of finance are well established, but there is controversy about the validity of the efficient markets hypothesis and whether stock prices are, in reality, rational estimates of a company's true value. Rational or not, large movements in stock prices have important macroeconomic implications. Stock market fluctuations often go hand in hand with fluctuations in the economy more broadly. We revisit the stock market when we study economic fluctuations later in the book.

Chapter in a Nutshell

- Because savings can earn interest, a sum of money today is more valuable than the same sum in the future. A person can compare sums from different times using the concept of present value. The present value of any future sum is the amount that would be needed today, given prevailing interest rates, to produce that future sum.
- Because of diminishing marginal utility, most people are risk averse. Risk can be reduced by buying insurance, diversifying holdings, and choosing a portfolio with lower risk and lower return.
- The value of an asset equals the present value of the cash flows the owner will receive. For a share of stock, these cash flows include the stream of dividends and the final sale price. According to the efficient markets hypothesis, financial markets process available information rationally, so a stock price always equals the best estimate of the value of the underlying business. Some economists question the efficient markets hypothesis, however, and say that irrational psychological factors influence asset prices.

Key Concepts

finance, p. 576
present value, p. 576
future value, p. 576
compounding, p. 576
risk aversion, p. 579
diversification, p. 580
firm-specific risk, p. 581
market risk, p. 581
fundamental analysis, p. 583
efficient markets hypothesis, p. 584
informational efficiency, p. 585
random walk, p. 585

Questions for Review

1. The interest rate is 7 percent. Use the concept of present value to compare $200 to be received in 10 years and $300 to be received in 20 years.
2. What benefit do people get from the market for insurance? What two problems impede the insurance market from working perfectly?
3. What is diversification? Does a stockholder get a greater benefit from diversification when going from 1 stock to 10 stocks or when going from 100 stocks to 120 stocks?
4. Between stocks and government bonds, which type of asset has more risk? Which pays a higher average return?
5. What factors should a stock analyst think about in determining the value of a share of stock?
6. Describe the efficient markets hypothesis, and give a piece of evidence consistent with this hypothesis.
7. Explain the view of those economists who are skeptical of the efficient markets hypothesis.

Problems and Applications

1. According to an old myth, Native Americans sold the island of Manhattan about 400 years ago for $24. If they had invested this amount at an interest rate of 7 percent per year, how much, approximately, would they have today?
2. A company has an investment project that would cost $10 million today and yield a payoff of $15 million in 4 years.
   a. Should the firm undertake the project if the interest rate is 11 percent? 10 percent? 9 percent? 8 percent?
   b. Can you figure out the exact interest rate at which the firm would be indifferent between undertaking and forgoing the project? (This interest rate is called the project's **internal rate of return**.)

3. Bond A pays $8,000 in 20 years. Bond B pays $8,000 in 40 years. (To keep things simple, assume that these are zero-coupon bonds, meaning the $8,000 is the only payment the bondholder receives.)
   a. If the interest rate is 3.5 percent, what is the value of each bond today? Which bond is worth more? Why? (Hint: You can use a calculator, but the rule of 70 should make the calculation easy.)
   b. If the interest rate increases to 7 percent, what is the value of each bond? Which bond has a larger **percentage** change in value?
   c. Based on the example above, fill in the two blanks in this sentence: "The value of a bond [rises/falls] when the interest rate increases, and bonds with a longer time to maturity are [more/less] sensitive to changes in the interest rate."
4. Your bank account pays an interest rate of 8 percent. You are considering buying a share of stock in XYZ Corporation for $110. After 1, 2, and 3 years, it will pay a dividend of $5. You expect to sell the stock after 3 years for $120. Is XYZ a good investment? Support your answer with calculations.
5. For each of the following kinds of insurance, give an example of behavior that reflects **moral hazard** and another example of behavior that reflects **adverse selection**.
   a. health insurance
   b. car insurance
   c. life insurance
6. Which kind of stock would you expect to pay the higher average return: stock in an industry that is very sensitive to economic conditions (such as an automaker) or stock in an industry that is relatively insensitive to economic conditions (such as a water company)? Why?
7. A company faces two kinds of risk. A firm-specific risk is that a competitor might enter its market and take some of its customers. A market risk is that the economy might enter a recession, reducing sales. Which of these two risks would more likely cause the company's shareholders to demand a higher return? Why?
8. When company executives buy and sell stock based on private information that they obtain as part of their jobs, they are engaged in **insider trading**.
   a. Give an example of inside information that might be useful for buying or selling stock.
   b. Those who trade stocks based on inside information usually earn very high rates of return. Does this fact violate the efficient markets hypothesis?
   c. Insider trading is illegal. Why do you suppose that is?
9. Jamal has a utility function $U = W^{1/2}$, where W is his wealth in millions of dollars, and U is the utility he obtains from that wealth. In the final stage of a game show, the host offers Jamal a choice between (A) $4 million for sure and (B) a gamble that pays $1 million with a probability of 0.6 and $9 million with a probability of 0.4.
   a. Graph Jamal's utility function. Is he risk averse? Explain.
   b. Does A or B offer Jamal the higher expected prize? Explain your reasoning with appropriate calculations. (Hint: The expected value of a random variable is the weighted average of the possible outcomes, where the probabilities are the weights.)
   c. Does A or B offer Jamal the higher expected utility? Again, show your calculations.
   d. Should Jamal pick A or B? Why?

QuickQuiz Answers

1. **b** 2. **d** 3. **b** 4. **c** 5. **c** 6. **a** 7. **a** 8. **b** 9. **d**

Chapter

29

Unemployment

Losing a job can be the most distressing economic event in a person's life. Most people count on their paychecks to maintain their standard of living, and many get a sense of personal accomplishment from their work. A job loss may mean immediate financial trouble, anxiety about the future, and reduced self-esteem. Understandably enough, when politicians campaign for office, they often claim that their proposed policies will create jobs.

Unemployment can be not only a personal tragedy but also a macroeconomic one. Previous chapters discussed some of the forces that influence the level and growth of a country's standard of living, such as saving and investment, the rule of law, political stability, educational attainment, openness to trade, and technological progress. Another critical determinant of a country's standard of living is the degree of unemployment it experiences. People who want to work but can't find a job are not producing goods and services. Unemployment is a drag on economic prosperity.

SHIFT DRIVE/SHUTTERSTOCK.COM; ERDARKVE+/GETTY IMAGES

Economists divide the problem of joblessness into two pieces—one that persists in the long run and one that occurs in the short run. The **natural rate of unemployment** is the term for the amount of unemployment that the economy normally experiences. **Cyclical unemployment** refers to the year-to-year fluctuations in unemployment around its natural rate and is closely associated with short-run fluctuations in economic activity. We examine cyclical unemployment when we study short-run fluctuations later in this book. This chapter focuses on the determinants of an economy's natural rate of unemployment. As we will see, the label **natural** does not mean that this rate of unemployment is desirable. Nor does it mean that it is constant over time or across places or that it is impervious to economic policy. It just means that this unemployment does not go away on its own, even in the long run.

The chapter begins with three questions about unemployment data: How does the government measure the economy's rate of unemployment? What problems arise in interpreting these data? How long are the unemployed typically out of work?

It then discusses why economies always have some unemployment and how policymakers can help the unemployed. We consider four explanations for the natural rate of unemployment: job search, minimum-wage laws, unions, and efficiency wages. Long-run unemployment does not arise from a single cause but, instead, reflects a variety of related forces. As a result, there is no easy recipe for policymakers to reduce the natural rate of unemployment and alleviate the hardships of the unemployed. But the study of unemployment clarifies the available options and the trade-offs that policymakers face.

29-1 Identifying Unemployment

Let's start by examining more precisely what the term **unemployment** means.

29-1a How Is Unemployment Measured?

In the United States, measuring unemployment is the job of the Bureau of Labor Statistics (BLS), a part of the Department of Labor. Every month, the BLS produces data on unemployment and other aspects of the labor market, including types of employment, length of the average workweek, and the duration of unemployment. These data come from a survey of about 60,000 households, called the Current Population Survey.

Based on the answers to the survey questions, the BLS places each adult (age 16 and older) in surveyed households into one of three categories:

- **Employed:** This category includes those who worked as paid employees, worked in their own business, or worked without pay in a family member's business. Both full-time and part-time workers are counted. Those who had jobs from which they were temporarily absent because of vacation, illness, bad weather, or similar reasons also fall within this category.
- **Unemployed:** Those who were not employed, were available for work, and had tried to find employment during the previous four weeks are in this group. It also includes those waiting to be recalled to a job after a temporary layoff.
- **Not in the labor force:** This category includes those who fit into neither of the first two categories, such as full-time students, homemakers, and retirees.

Figure 1 shows the breakdown into these categories for December 2021.

Once the BLS has categorized everyone in the survey, it computes statistics to summarize the state of the labor market. The BLS defines the **labor force** as the sum of the employed and the unemployed:

$$\text{Labor force} = \text{Number of employed} + \text{Number of unemployed}.$$

labor force
the total number of workers, including both the employed and unemployed

The **unemployment rate** is the percentage of the labor force that is unemployed:

$$\text{Unemployment rate} = \frac{\text{Number of unemployed}}{\text{Labor force}} \times 100.$$

unemployment rate
the percentage of the labor force that is unemployed

The BLS calculates unemployment rates for the entire adult population and for demographic groups defined by race, gender, and so on.

The BLS uses the same survey to produce data on labor-force participation. The **labor-force participation rate** measures the percentage of the total adult population of the United States that is in the labor force:

$$\text{Labor-force participation rate} = \frac{\text{Labor force}}{\text{Adult population}} \times 100.$$

labor-force participation rate
the percentage of the adult population that is in the labor force

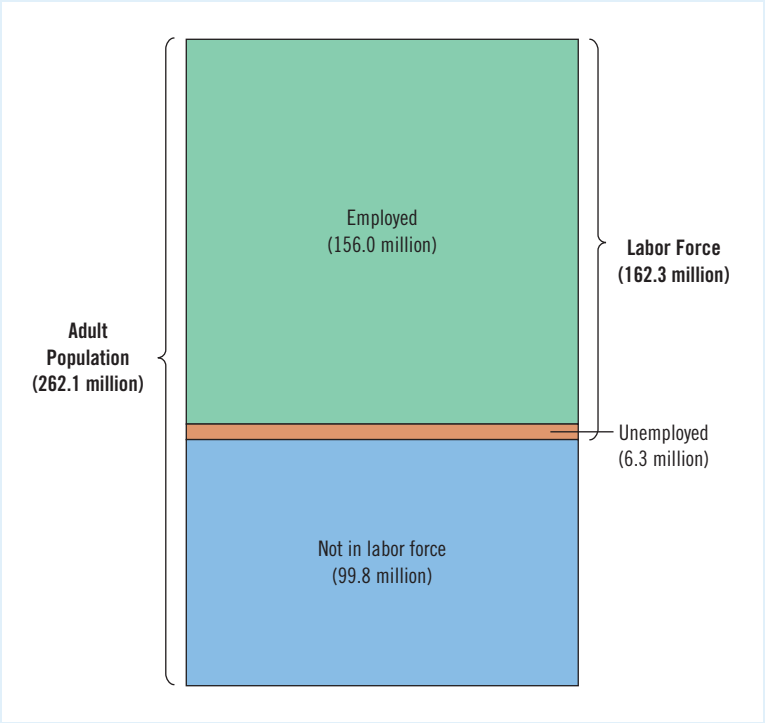
This statistic tells us the fraction of the population that participates in the labor market. The labor-force participation rate, like the unemployment rate, is computed for both the entire adult population and specific groups.

Figure 1

The Breakdown of the Population in December 2021

The Bureau of Labor Statistics divides the adult population into three categories: employed, unemployed, and not in the labor force.

Source: Bureau of Labor Statistics.



To see how these data are calculated, consider the figures for December 2021. At that time, 156.0 million people were employed, and 6.3 million people were unemployed. The labor force was

$$\text{Labor force} = 156.0 + 6.3 = 162.3 \text{ million}.$$

The unemployment rate was

$$\text{Unemployment rate} = (6.3/162.3) \times 100 = 3.9 \text{ percent}.$$

Because the adult population was 262.1 million, the labor-force participation rate was

$$\text{Labor-force participation rate} = (162.3/262.1) \times 100 = 61.9 \text{ percent}.$$

Hence, in December 2021, 61.9 percent of the U.S. adult population were participating in the labor market, and 3.9 percent of those labor-market participants were without work.

natural rate of unemployment the normal rate of unemployment around which the unemployment rate fluctuates

cyclical unemployment the deviation of unemployment from its natural rate

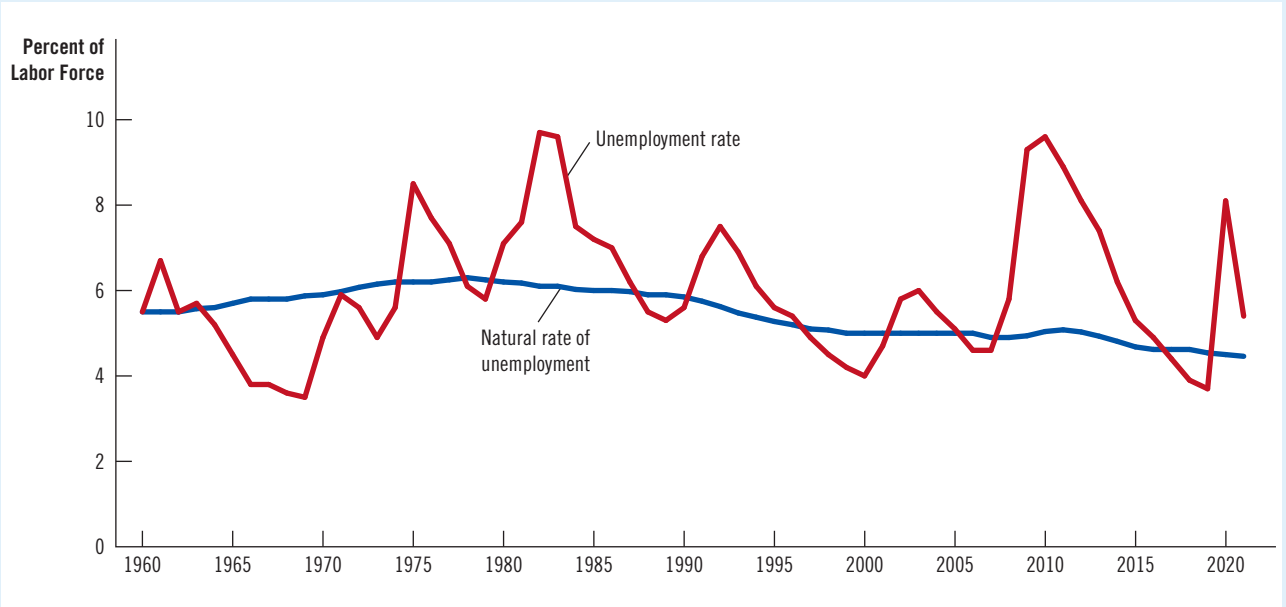
The BLS data on the labor market allows economists and policymakers to monitor changes in the economy over time. Figure 2 shows the unemployment rate in the United States since 1960. The figure shows that the economy always has some unemployment and that the amount changes from year to year. As noted earlier, the normal rate of unemployment around which the unemployment rate fluctuates is called the **natural rate of unemployment**, and the deviation of unemployment from its natural rate is called **cyclical unemployment**. The natural rate of unemployment

Figure 2

Unemployment Rate since 1960

This graph uses annual data on the U.S. unemployment rate to show the percentage of the labor force without a job. The natural rate of unemployment is the normal level of unemployment around which the unemployment rate fluctuates.

Source: U.S. Department of Labor, Congressional Budget Office.



shown in the figure is a series produced by economists at the Congressional Budget Office. For 2021, they estimated a natural rate of 4.5 percent, compared with the actual unemployment rate of 5.4 percent. The rest of this chapter ignores short-run fluctuations in unemployment around its natural rate and examines why there is always some unemployment.

Case Study

Labor-Force Participation of Women and Men in the U.S. Economy

The past century has seen a dramatic transformation in women's role in American society. This change has many causes. In part, it is attributable to the mass adoption of new technologies, such as the washing machine, clothes dryer, refrigerator, freezer, and dishwasher, which reduced the time required for routine household tasks. In part, it is attributable to improved birth control, which reduced the number of children born to the typical family. And in part, it is attributable to shifting political and social attitudes, which in turn were facilitated by the advances in technology and birth control. Together, these developments have had a profound impact on society in general and on the economy in particular.

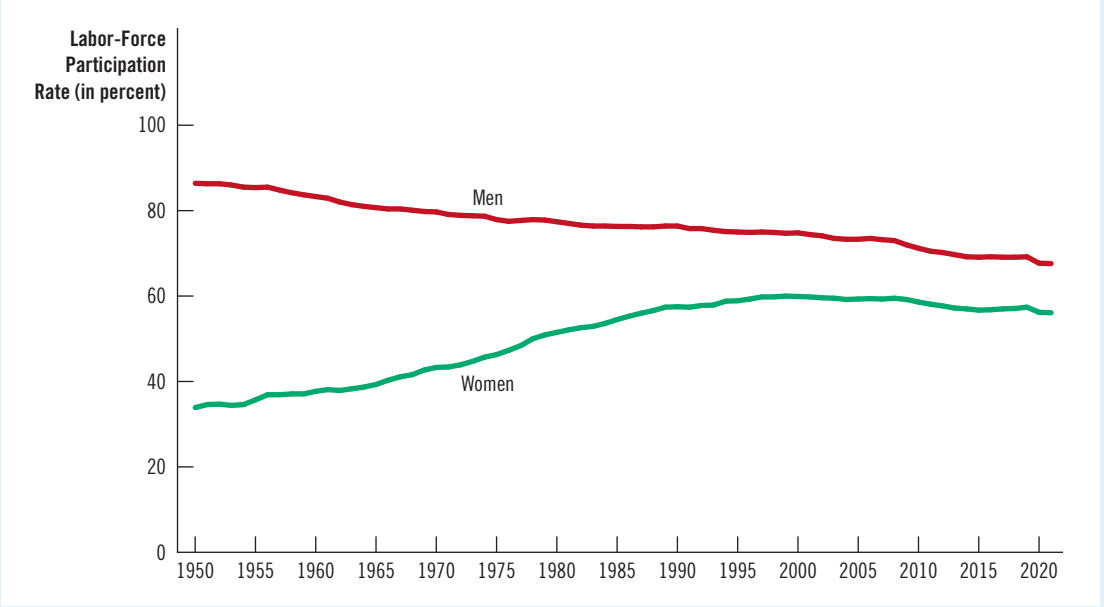
That impact is clear in the data on labor-force participation. Figure 3 shows the labor-force participation rates of men and women in the United States since 1950. Just after World War II, men and women had very different experiences. Only 33 percent of women were working or looking for work, compared with 87 percent of

Figure 3

Labor-Force Participation Rates for Men and Women since 1950

This figure shows the percentage of adult men and women who are members of the labor force. It shows that, over the past several decades, women have entered the labor force, and men have left it.

Source: U.S. Department of Labor.



men. Since then, this difference in participation rates has gradually diminished, as growing numbers of women have entered the labor force and some men have left it. Data for 2021 show that 56 percent of women were in the labor force, compared with 68 percent of men. As measured by labor-force participation, men and women are now playing more equal roles in the economy.

The increase in women's labor-force participation is easy to explain, but the fall in men's may seem puzzling. There are several reasons for this decline. First, young men stay in school longer than their fathers and grandfathers did. Second, older men are enjoying longer periods of retirement. Third, with more women employed, more fathers stay at home to raise their children. Full-time students, retirees, and stay-at-home dads are all counted as being out of the labor force. ●

29-1b Does the Unemployment Rate Measure What We Want It to Measure?

Measuring the amount of unemployment in the economy is not straightforward. It is easy to distinguish between people who work full-time and those who don't work at all. But among those who are not working, it is hard to tell who is unemployed and who is not in the labor force.

Movements into and out of the labor force are common. More than one third of the unemployed are recent entrants into the labor force. These entrants include young workers looking for their first jobs. They also include, in greater numbers, people who had left the labor force but have now returned to look for work. Moreover, not all unemployment ends with the job seeker finding work. Almost half of all spells of unemployment end when the jobless person stops looking for work and leaves the labor force.

Because people move into and out of the labor force so often, statistics on unemployment are difficult to interpret. On the one hand, some of those who report being unemployed may not be trying hard to find a job. They may be calling themselves unemployed because they want to qualify for a government program that gives financial assistance to the unemployed or because they are working but are paid "under the table" to avoid taxes on their earnings. (This may have been especially important in 2020 and much of 2021, when government assistance programs during the pandemic gave some people extra incentives to call themselves unemployed.) It may be more accurate to view these individuals as out of the labor force or, in some cases, employed. On the other hand, some of those who report being out of the labor force may want to work. These individuals may have given up looking for a job after an unsuccessful search. Such individuals, called **discouraged workers**, do not show up in unemployment statistics, even though they are truly prospective workers without jobs.

discouraged workers individuals who would like to work but have given up looking for a job

Because of these and other problems, the BLS calculates several measures of labor underutilization, which are presented in Table 1. In the end, it is best to view the official unemployment rate as a useful but imperfect measure of joblessness.

29-1c How Long Are the Unemployed without Work?

In judging how serious the problem of unemployment is, one question to consider is whether unemployment is typically a short-term or long-term condition. If unemployment is short-term, one might conclude that it is not a big problem. Workers may require a few weeks between jobs to find the openings that best suit their tastes and skills. Yet if unemployment is long-term, one might conclude that it is serious indeed. Workers unemployed for many months are more likely to suffer economic and psychological hardship.

Table 1

Measures of Labor Underutilization

The table shows various measures of joblessness for the U.S. economy. The data are for December 2021.

Source: U.S. Department of Labor.

Measure and Description		Rate
U-1	Persons unemployed 15 weeks or longer, as a percent of the civilian labor force (includes only very long-term unemployed)	1.7%
U-2	Job losers and persons who have completed temporary jobs, as a percent of the civilian labor force (excludes job leavers)	1.9
U-3	Total unemployed, as a percent of the civilian labor force (official unemployment rate)	3.9
U-4	Total unemployed, plus discouraged workers, as a percent of the civilian labor force plus discouraged workers	4.3
U-5	Total unemployed plus all marginally attached workers, as a percent of the civilian labor force plus all marginally attached workers	4.9
U-6	Total unemployed, plus all marginally attached workers, plus total employed part-time for economic reasons, as a percent of the civilian labor force plus all marginally attached workers	7.3

Note: The Bureau of Labor Statistics defines terms as follows.

- **Marginally attached workers** are people neither working nor looking for work but indicate that they want and are available for a job and have looked for work sometime in the recent past.
- **Discouraged workers** are marginally attached workers who have given a job-market-related reason for not currently looking for a job.
- **Persons employed part-time for economic reasons** are those who want and are available for full-time work but have had to settle for a part-time schedule.

Economists have devoted much energy to studying data on the duration of unemployment spells. In this work, they have uncovered a result that is important, subtle, and seemingly contradictory: **Most spells of unemployment are short, but most unemployment observed at any given time is long-term.**

To see how this statement can be true, consider an example. Suppose that you visited the government's unemployment office every week for a year to survey the unemployed. Each week you find four unemployed workers. Three are the same individuals for the whole year, while the fourth person changes every week. Based on this experience, would you say that unemployment is typically short-term or long-term?

Simple calculations help answer this question. In this example, you meet a total of 55 unemployed people over the course of a year, 52 who are unemployed for one week and 3 who are unemployed for the full year. This means that 52/55, or 95 percent, of unemployment spells end in one week. Yet whenever you walk into the unemployment office, three of the four people you meet will be unemployed for the entire year. So, even though 95 percent of unemployment spells end in one week, 75 percent of the unemployment observed at any moment is attributable to

those who are unemployed for a full year. In this example, as in the world, most spells of unemployment are short, but most unemployment observed at any time is long-term.

This subtle conclusion implies that economists and policymakers must be careful when interpreting data on unemployment and when designing policies to help the unemployed. Most people who become unemployed will soon find jobs. Yet most of the economy's unemployment problem is attributable to the relatively few workers who are jobless for long periods.

29-1d Why Are There Always Some People Unemployed?

We have discussed the measurement of unemployment, the problems that arise in interpreting unemployment statistics, and the duration of unemployment. Yet we have not explained why economies experience unemployment in the first place.

Let's go back to supply and demand. In the standard model of a competitive market, as introduced in Chapter 4, the goods being sold are all the same, and the price adjusts to bring the quantity supplied and the quantity demanded into balance. For some purposes, this model can be applied to the labor market: Labor is the good, and the price is the wage. But like all models, this one has its limitations. It cannot fully describe the labor market because, at the model's equilibrium, there is no unemployment.

In the world, there are always some workers without jobs, even when the economy is doing well. Unemployment never falls to zero but, instead, fluctuates around its natural rate. To explain unemployment, the remaining sections of this chapter consider why actual labor markets depart from the benchmark model of supply and demand.

To preview the conclusions, there are four explanations for unemployment. The first emphasizes that workers and jobs are diverse, so it takes time for workers to find the jobs that best suit them and for firms to find the workers that best

The Jobs Number

When the Bureau of Labor Statistics announces the unemployment rate at the beginning of every month, it also gives the number of jobs that the economy gained or lost in the previous month. As an indicator of short-run economic trends, the jobs number gets as much attention as the unemployment rate.

Where does the jobs number come from? You might guess that it comes from the same survey of 60,000 households that yields the unemployment rate. And indeed, the household survey does produce data on total employment. The jobs number that gets the most attention, however, comes from a separate survey of 160,000 business establishments that have over 40 million workers on their payrolls. The results of the establishment survey are announced at the same time as the results of the household survey.

Both surveys yield information about total employment, but the results are not always the same. One reason is that the establishment survey has a larger sample, which makes it more reliable. Another is that the surveys are not measuring exactly the same thing. For example, a person who has two part-time jobs at different companies is counted as one employed person in the household survey but as two jobs in the establishment survey. As another example, people running their own small businesses are counted as employed in the household survey but do not show up at all in the establishment survey because it counts only employees on business payrolls.

The establishment survey is closely watched for its data on jobs, but it says nothing about unemployment. To measure the number of unemployed, we need to know how many people without jobs are trying to find them. The household survey is the only source of that information. ■

fill their needs. The unemployment that results from the search process is called **frictional unemployment**, and it is often thought to explain relatively short spells of unemployment.

The next three explanations have a common theme: Unemployment arises when, in some labor markets, the number of jobs available is less than the number of people looking for work. In an ideal labor market, this never occurs because wages equilibrate supply and demand. But sometimes wages are stuck above the equilibrium level, and the quantity of labor supplied exceeds the quantity demanded. Unemployment of this sort is called **structural unemployment**, and it is usually thought to explain longer spells of unemployment.

We will discuss three reasons for above-equilibrium wages: minimum-wage laws, unions, and efficiency wages. Wages are set above the equilibrium level by the government in the first case, by workers in the second, and by firms in the third.

frictional unemployment
unemployment that results because it takes time for workers to search for the jobs that best suit their tastes and skills

structural unemployment
unemployment that results because the number of jobs available in some labor markets is insufficient to provide a job for everyone who wants one

QuickQuiz

1. The population of Ectenia is 100 people: 40 work full-time, 20 work half-time but would prefer to work full-time, 10 are looking for a job, 10 would like to work but are so discouraged that they have given up looking, 10 are not interested in working because they are full-time students, and 10 are retired. What is the number of unemployed?
   a. 10
   b. 20
   c. 30
   d. 40

2. Using the numbers in the preceding question, what is the size of Ectenia's labor force?
   a. 50
   b. 60
   c. 70
   d. 80

Answers are at the end of the chapter.

29-2 Job Search

One reason economies experience unemployment is **job search**, the process of workers finding appropriate jobs. If all workers and all jobs were the same, identifying a good match would be quick and easy. But workers differ in their tastes and skills, jobs differ in their attributes, and information about job candidates and job vacancies disseminates slowly among the economy's many firms and households. As this process unfolds, some people spend time unemployed.

job search
the process by which workers find appropriate jobs given their tastes and skills

29-2a Why Some Frictional Unemployment Is Inevitable

The labor market is in a constant state of churn. As firms experience rising and falling fortunes, their demand for labor shifts, resulting in frictional unemployment. For example, when consumers decide that they prefer cars from Tesla over those from Ford, Tesla increases employment, and Ford lays off workers. The former Ford workers must now look for new jobs, and Tesla must decide which workers to hire to fill its vacancies. The result of this transition is a period of unemployment for some workers.

Similarly, employment can rise in one region while it falls in another. Consider what happens when the price of oil declines in world markets. Oil-producing firms

in Texas and North Dakota respond to the lower price by cutting back production and employment. Meanwhile, cheaper gasoline spurs car sales, so auto-producing firms in Michigan and Ohio expand production and employment. The opposite happens when the price of oil rises. Changes in the composition of demand among industries or regions are called **sectoral shifts**. Because it takes time for workers to search for jobs in the new sectors, sectoral shifts cause unemployment.

Changing patterns of international trade can also lead to frictional unemployment. Chapter 3 explained that nations export goods for which they have a comparative advantage and import goods for which other nations have an advantage. Comparative advantage, however, need not be stable over time. As the world economy evolves, nations may find themselves importing and exporting different goods than they have in the past. Workers will need to move among industries. As they make this transition, they may find themselves unemployed for a while.

Frictional unemployment is inevitable because the economy is always changing. For example, in the U.S. economy from 2010 to 2020, employment fell by 235,000 in newspaper and book publishing, 238,000 in the hotel industry, and 534,000 in state and local government. During the same period, employment rose by 650,000 in manufacturing, 1.8 million in construction, and 3.0 million in healthcare. This churning of the labor force is normal in a well-functioning, dynamic economy. Because workers tend to move toward industries in which they are most valuable, the long-run result for the overall economy is higher productivity and living standards. But along the way, workers in declining industries find themselves out of work and searching for new jobs. The result is frictional unemployment.

29-2b Public Policy and Job Search

Even if frictional unemployment is inevitable, the precise amount is not. The faster information spreads about job openings and worker availability, the more rapidly the economy can match workers and firms. The Internet, for instance, may help facilitate job search and reduce frictional unemployment. In addition, public policy may be able to reduce the time it takes unemployed workers to find new jobs, thereby lowering the economy's natural rate of unemployment.

Government programs try to facilitate job search in various ways. One is through government-run employment agencies, which give out information about job vacancies. Another is through training programs, which aim to ease workers' transition from declining to growing industries and help disadvantaged groups escape poverty. Advocates of these programs believe that they make the economy operate more efficiently by keeping the labor force more fully employed and that they reduce the inequities inherent in a constantly changing market economy.

Critics of these programs question whether the government should play a role in the process of job search. They argue that it is better to let the private market match workers and jobs and that the government is no better—and most likely worse—at disseminating the right information to the right workers and deciding what kinds of worker training would be most valuable. In fact, most job search in the U.S. economy takes place without government intervention. Newspaper ads, online job sites, college career offices, headhunters, and word of mouth all help spread information about job openings and job candidates. Similarly, much worker education is done privately, through schools or on-the-job training.

unemployment insurance
a government program that partially protects the incomes of workers who become unemployed

29-2c Unemployment Insurance

One government program that increases the amount of frictional unemployment, without intending to do so, is **unemployment insurance**. This program offers

workers partial protection against job loss. The unemployed who quit their jobs, were fired for cause, or just entered the labor force are not eligible. Benefits are paid only to the unemployed who were laid off because their previous employers no longer needed their skills. The terms of the program vary over time and across states, but typical workers covered by unemployment insurance in the United States receive 50 percent of their former wages for 26 weeks. This insurance is often increased in national downturns, and that was especially true during the coronavirus pandemic of 2020–2021.

While unemployment insurance reduces the hardship of unemployment, it also increases the amount of unemployment. The explanation is based on one of the **Ten Principles of Economics** in Chapter 1: People respond to incentives. Because unemployment benefits stop when a worker takes a new job, the unemployed may devote less effort to job search and may be more likely to turn down unattractive job offers. In addition, because unemployment insurance makes unemployment less onerous, workers may be less likely to seek guarantees of job security when they negotiate with employers over the terms of their employment.

Many studies by labor economists have analyzed the incentive effects of unemployment insurance. One study examined an experiment run by the state of Illinois in 1985. When unemployed workers applied to collect unemployment insurance benefits, the state randomly selected some of them and offered each a $500 bonus if they found new jobs within 11 weeks. This group was then compared with a control group that did not receive the incentive. The average spell of unemployment for the group offered the bonus was 7 percent shorter than the average spell for the control group. This experiment shows that the design of the unemployment insurance system influences the effort that the unemployed devote to job search.

Several other studies examined search effort by following a group of workers over time. Unemployment insurance benefits, rather than lasting forever, usually run out after six months or one year. These studies found that when the unemployed become ineligible for benefits, the probability of them finding a new job rises markedly. This suggests that receiving unemployment insurance benefits reduces the search effort of the unemployed.

Even though unemployment insurance has been shown to reduce search effort and raise unemployment, we should not necessarily conclude that the policy is ill advised. The program achieves its primary goal of reducing the income uncertainty that workers face. In addition, when workers turn down unattractive job offers, they can look for jobs that better suit their tastes and skills. Some economists argue that unemployment insurance improves the ability of the economy to match each worker with the most appropriate job.

QuickQuiz

3. The main policy goal of unemployment insurance is to reduce the
   a. search effort of the unemployed.
   b. income uncertainty that workers face.
   c. role of unions in wage setting.
   d. amount of frictional unemployment.

4. One unintended consequence of unemployment insurance is that it reduces the
   a. search effort of the unemployed.
   b. income uncertainty that workers face.
   c. role of unions in wage setting.
   d. amount of frictional unemployment.

Answers are at the end of the chapter.

29-3 Minimum-Wage Laws

While frictional unemployment results from the process of job search, structural unemployment occurs when the number of jobs is insufficient for the number of workers.

To understand structural unemployment, let's begin by recalling how minimum-wage laws can cause unemployment. Minimum wages are not the main reason for unemployment in the U.S. economy, but their analysis is a good starting point because it can be used to understand the other reasons for structural unemployment.

Figure 4 reviews how a minimum wage affects the outcome in a competitive labor market. When a minimum wage is binding, it forces the wage to remain above the level that balances supply and demand. Compared with the equilibrium that would otherwise prevail, the quantity of labor supplied is higher, and the quantity of labor demanded is lower. There is a surplus of labor. Because more workers are willing to work than there are jobs, some workers are unemployed.

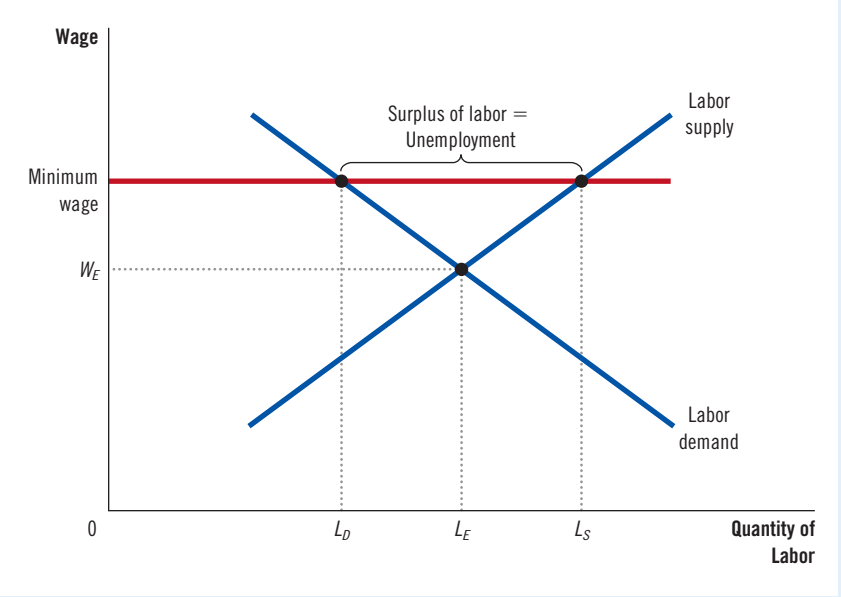
In the U.S. economy, minimum-wage laws affect a small percentage of the labor force. Most workers have wages well above the legal minimum, so the law does not prevent their wages from adjusting to balance supply and demand. Minimum-wage laws matter most for those with low levels of skill and experience, such as teenagers. Because their equilibrium wages are more likely to fall below the legal minimum, it is mainly among these workers that minimum-wage laws help explain unemployment.

Yet Figure 4 illustrates a more general lesson: **If the wage is kept above the equilibrium level for any reason, unemployment results.** The remaining two sections of this chapter consider two reasons beyond minimum-wage laws for above-equilibrium wages: unions and efficiency wages. The basic economics in these cases is similar to that shown in Figure 4, but these explanations of structural unemployment apply to many more workers.

Figure 4

Unemployment from a Wage above the Equilibrium Level

In this labor market, supply and demand are balanced at the wage W_E. At this equilibrium wage, the quantity of labor supplied and the quantity of labor demanded both equal L_E. But if the wage is forced to remain above the equilibrium level, perhaps because of a minimum-wage law, the quantity of labor supplied rises to L_S, and the quantity of labor demanded falls to L_D. The resulting surplus of labor, $L_S - L_D$, represents unemployment.



Before moving on, however, note the key difference between structural and frictional unemployment. Structural unemployment arises from an above-equilibrium wage, while frictional unemployment arises from the process of job search. The need for job search is not due to the failure of wages to balance labor supply and labor demand but instead follows from the great diversity of workers and jobs. When unemployment is frictional, workers are **searching** for the jobs that best suit their tastes and skills. By contrast, when unemployment is structural, the quantity of labor supplied exceeds the quantity demanded, and workers are **waiting** for jobs to open up.

Who Earns the Federal Minimum Wage?

In 2021, the Department of Labor released a study showing what kinds of workers reported earnings at or below the federal minimum wage in 2020, when it was $7.25 per hour. (A reported wage below the minimum wage is possible because some workers are exempt from the statute, because enforcement is imperfect, and because some workers round down when reporting their wages on surveys.) Here is a summary of the findings:

- In 2020, 73 million workers were paid at hourly rates (as opposed to being salaried or self-employed), representing 55 percent of the labor force. Among hourly paid workers, about 1.5 percent reported wages at or below the prevailing federal minimum. Overall, the federal minimum wage directly affects about 0.8 percent of all workers.
- Minimum-wage workers tend to be young. Among employed teenagers (ages 16 to 19) paid by the hour, about 5 percent earned the minimum wage or less, compared with 1 percent of hourly paid workers age 25 and older.
- Minimum-wage workers tend to be less educated. Among hourly paid workers age 16 and older, about 2 percent of those without a bachelor's degree earned the minimum wage or less, compared with about 1 percent of those with a bachelor's degree.
- Minimum-wage workers are more likely to be working part-time. Among part-time workers (those who usually work less than 35 hours per week), 4 percent were paid the minimum wage or less, compared with 1 percent of full-time workers.
- The industry with the highest proportion of workers with reported hourly wages at or below the minimum wage was leisure and hospitality (8 percent). About three-fifths of all workers paid at or below the minimum wage were employed in this industry, primarily in restaurants and other food services. For many of these workers, tips supplement their hourly wages.
- The percentage of hourly paid workers earning the prevailing federal minimum wage or less has changed substantially over time. It has declined from 13.4 percent in 1979, when data collection first began on a regular basis, to 1.5 percent in 2020. One reason for this change is that the federal minimum wage has not kept up with inflation. If it had, the minimum wage in 2020 would have been about $10 rather than $7.25 per hour. At a higher level, the minimum wage becomes a binding price floor for more workers.

Finally, note that about half of the U.S. states have minimum wages above the federal one, sometimes substantially so. For workers in these states, the state minimum wage is the more relevant one. ●

QuickQuiz

5. In a competitive labor market, an increase in the minimum wage results in a(n) _____ in the quantity of labor supplied and a(n) _____ in the quantity of labor demanded.
   a. increase; increase
   b. increase; decrease
   c. decrease; increase
   d. decrease; decrease

6. Approximately what percent of U.S. workers are directly affected by the federal minimum wage?
   a. 1
   b. 6
   c. 12
   d. 25

Answers are at the end of the chapter.

29-4 Unions and Collective Bargaining

union
a worker association that bargains with employers over wages, benefits, and working conditions

A **union** is a worker association that bargains with employers over wages, benefits, and working conditions. In the 1940s and 1950s, when union membership in the United States was at its peak, about 33 percent of the U.S. labor force was unionized. Today, less than 11 percent of U.S. workers belong to unions. In many European countries, however, unions play a large role. In Belgium, Norway, and Sweden, more than 50 percent of workers belong to unions. In France, Italy, and Germany, most workers have wages set by collective bargaining by law, even though only some of these workers are themselves union members. In these cases, wages are not determined by the equilibrium of supply and demand in competitive labor markets.

29-4a The Economics of Unions

The word "cartel" is usually applied to a combination of firms that join forces to limit competition, but in many ways, it applies to unions as well. Like any cartel, a union is a group of sellers acting together with the goal of exerting their joint market power. Most workers in the U.S. economy deal with their employers as individuals on issues like wages, benefits, and working conditions. By contrast, workers in a union do so collectively so they have more influence over the outcome. The process by which unions and firms agree on the terms of employment is called **collective bargaining**.

collective bargaining
the process by which unions and firms agree on the terms of employment

strike
the organized withdrawal of labor from a firm by a union

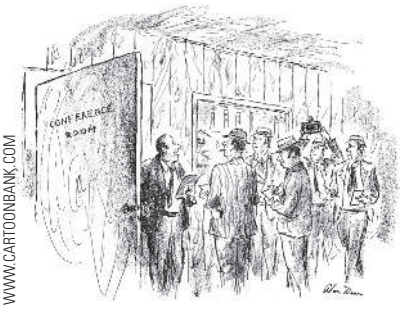
When a union bargains with a firm, it asks for higher wages, better benefits, and better working conditions than the firm might offer in the absence of a union. If the union and the firm do not reach an agreement, the union can organize a withdrawal of labor from the firm, called a **strike**. Because a strike reduces production, sales, and profit, a firm facing a strike threat is likely to agree to pay higher wages than it otherwise would. Economists who study the effects of unions typically find that union workers earn about 10 to 20 percent more than similar workers who do not belong to unions.

When a union raises the wage above the equilibrium level, it increases the quantity of labor supplied and reduces the quantity of labor demanded, resulting in unemployment. Workers who are employed at the higher wage are better off. Those who would otherwise be employed but are now unemployed are worse off. For this reason, unions are often thought to create conflict between different groups of workers—between the **insiders** who benefit from high union wages and the **outsiders** who do not get the union jobs.

The outsiders can respond to their status in one of two ways. Some remain unemployed and wait for the chance to become insiders and earn the high union wage. Others take jobs in firms that are not unionized. Thus, when unions raise wages in one part of the economy, the supply of labor increases in other parts of the economy. This increase in labor supply, in turn, reduces wages in industries that are not unionized. In other words, workers in unions reap the benefit of collective bargaining, while workers not in unions bear some of the cost.

The role of unions in the economy depends in part on the laws that govern union organization and collective bargaining. Normally, explicit agreements among members of a cartel are illegal. When firms that sell similar products agree to set high prices, the agreement is considered a "conspiracy in restraint of trade," and the government prosecutes the firms for violating the antitrust laws. Unions, however, are exempt from these laws. The policymakers who wrote the antitrust and labor laws believed that workers needed greater market power as they bargained with employers. Indeed, various laws are designed to encourage the formation of unions. The National Labor Relations Act (enacted in 1935 and subsequently amended) prohibits employers from interfering in certain ways with workers trying to organize unions, and in unionized companies, it requires employers and unions to bargain in good faith when negotiating the terms of employment.

Legislation affecting the market power of unions is a perennial topic of political debate. For instance, state lawmakers sometimes debate **right-to-work laws**, which bar a union and employer from requiring workers to financially support the union. Absent such laws, a union can seek an agreement during collective bargaining that requires all employees to pay union dues (for union members) or an agency fee (for nonmembers) as a condition of employment. As of 2021, about half of the U.S. states had right-to-work laws, and some members of Congress have proposed national right-to-work legislation. Lawmakers in Washington have also considered laws that would either make strikes more possible or prohibit them in some situations. For example, one proposal would prevent firms from hiring permanent replacements (as opposed to temporary replacements) for workers on strike. If such a law were enacted, striking workers would no longer face the threat of losing their jobs to permanent replacements, making strikes more viable and thereby increasing unions' market power. Another proposal would bar strikes in the airline and railroad industries by requiring unions and employers at the end of collective bargaining to resolve their remaining disagreements through arbitration. How these policy debates are resolved will help determine the future of the union movement.



ALAN DUNN/THE NEW YORKER COLLECTION/
WWW.CARTOONBANK.COM

"Gentlemen, nothing stands in the way of a final accord except that management wants profit maximization and the union wants more moola."

29-4b Are Unions Good or Bad for the Economy?

Economists disagree about whether unions are good or bad for the economy as a whole. Let's consider both sides of the debate.

Critics argue that unions have the downsides of any cartel. When unions raise wages above the level that would prevail in competitive markets, they reduce the quantity of labor demanded, cause some workers to be unemployed, and reduce the wages in the rest of the economy. The resulting allocation of labor, critics say, is both inefficient and inequitable. It is inefficient because high union wages reduce employment in unionized firms below the efficient, competitive level. It is inequitable because some workers benefit at the expense of other workers.

Advocates contend that unions are a necessary antidote to the market power of the firms that hire workers. The extreme case of this market power is the "company town," where a single firm does most of the hiring in a geographical region. In a company town, if workers do not accept the wages and working conditions that the firm offers, they have little choice but to move or stop working and suffer the consequences. In the absence of a union, therefore, the firm could use its market power to pay lower wages and offer worse working conditions than it would if it had to compete with other firms for the same workers. In this case, a union may be necessary to check the firm's market power and protect the workers from being at the mercy of the firm's owners.

Advocates of unions also claim that unions help firms respond efficiently to workers' concerns. When a worker takes a job, the worker and the firm must agree on many attributes of the job in addition to the wage: hours of work, overtime, vacations, sick leave, health benefits, promotion schedules, job security, and so on. By representing workers' views on these issues, unions help firms provide the right mix of job attributes. Even if unions have the adverse effect of pushing wages above the equilibrium level and causing unemployment, they have the benefit of ensuring that firms have a content and productive workforce.

In the end, there is no consensus about the overall impact of unions on the economy. Like many institutions, their influence is probably beneficial in some circumstances and adverse in others.

QuickQuiz

7. In the United States, unionized workers are paid about _____ percent more than similar nonunion workers.
   a. 2
   b. 5
   c. 15
   d. 40

8. In many European nations, unions
   a. are considered cartels in violation of antitrust laws.
   b. conspire with firms to keep wages below competitive levels.
   c. are social clubs without any economic impact.
   d. play a much larger role than they do in the United States.

Answers are at the end of the chapter.

FYI Mismatch as a Source of Structural Unemployment

Unemployment is sometimes said to result from a mismatch between the kinds of workers that are seeking jobs and the kinds that firms are looking to hire. If the available workers are trained as truck drivers but firms want computer technicians, that is a mismatch. If the available workers stopped their educations before completing high school but firms want workers with college degrees, that is also a mismatch. If the available workers live in Texas but firms are hiring in Florida, that is a mismatch as well.

In an ideal labor market governed by supply and demand, wages would adjust to these mismatches. Wages would fall for truck drivers, high school dropouts, and Texans, and rise for computer technicians, college graduates, and Floridians. Sufficient wage adjustment would restore equilibrium without unemployment.

In the real world, wages don't always adjust quickly and fully to supply and demand. As a result, when mismatches arise, the quantity of labor supplied can exceed the quantity demanded in some labor markets. In our examples, there would be a surplus of truck drivers, dropouts, and Texans. This is a type of structural unemployment. It occurs when supply or demand shifts and, for various reasons, some wages are stuck above the equilibrium level. ■

29-5 The Theory of Efficiency Wages

efficiency wages above-equilibrium wages paid by firms to increase worker productivity

In addition to job search, minimum-wage laws, and unions, the theory of **efficiency wages** suggests a fourth reason that economies experience some unemployment. According to this theory, firms operate more efficiently if wages are above the equilibrium level. Therefore, it may be profitable for firms to keep wages high despite a surplus of labor.

In some ways, the unemployment that arises from efficiency wages is like the unemployment that arises from minimum-wage laws and unions. In all three cases, unemployment is the result of wages above the level that balances supply and demand. Yet there is also an important difference. Minimum-wage laws and unions prevent firms from lowering wages in the presence of a surplus of workers. Efficiency-wage theory states that such a constraint on firms is unnecessary because, in some cases, firms may want to keep wages above the equilibrium level.

Why might a firm prefer high wages? Normally, profit-maximizing firms want to keep costs—and therefore wages—as low as possible. The novel insight of efficiency-wage theory is that higher wages can make a firm more profitable by increasing the efficiency of its workers.

There are several types of efficiency-wage theory. Each suggests a different mechanism for how wages affect worker efficiency.

29-5a Worker Health

The first and simplest type of efficiency-wage theory emphasizes the link between wages and worker health. Better-paid workers eat a more nutritious diet, and workers who eat a better diet are healthier and more productive. A firm may find it profitable to pay high wages to ensure its workers are healthy and productive.

This type of efficiency-wage theory can be relevant for explaining above-equilibrium wages and unemployment in less developed countries where inadequate nutrition is a problem. In these countries, firms may fear that cutting wages would adversely influence their workers' health and productivity. Worker health concerns are less relevant for firms in rich countries such as the United States, where the equilibrium wages for most workers are well above the level needed for an adequate diet.

29-5b Worker Turnover

A second type of efficiency-wage theory emphasizes the link between wages and worker turnover. Workers quit jobs for many reasons: to take jobs at other firms, to move to other parts of the country, to leave the labor force, and so on. The frequency with which they quit depends on the entire set of incentives they face, including the benefits of leaving and the benefits of staying. The more a firm pays its workers, the less often its workers will choose to leave. A firm can reduce turnover by paying higher wages.

Firms care about turnover because hiring and training new workers is costly. Moreover, even after they are trained, new workers are not as productive as experienced ones. Higher turnover, therefore, tends to increase production costs. Firms may find it profitable to pay wages above the equilibrium level to reduce worker turnover.

29-5c Worker Quality

A third type of efficiency-wage theory emphasizes the link between wages and worker quality. All firms want workers who are talented. But because firms cannot perfectly gauge the quality of applicants, hiring has a degree of randomness to it.

When a firm pays high wages, it attracts a better pool of workers to apply for its jobs and thereby increases the quality of its workforce. If the firm responded to a surplus of labor by reducing the wage, the most competent applicants—who are more likely to have better opportunities than less competent applicants—may choose not to apply. If this influence of the wage on worker quality is strong enough, it may be profitable for the firm to pay a wage above the level that balances supply and demand.

29-5d Worker Effort

A fourth type of efficiency-wage theory emphasizes the link between wages and worker effort. In many jobs, workers have discretion over how hard to work, and some may choose to work as little as possible. Because workplace monitoring is costly and imperfect, firms cannot quickly catch all shirkers.

One way to deter shirking is to pay wages above the equilibrium level. High wages make workers more eager to keep their jobs and motivate them to put forward their best effort. If the wage were at the level that balanced supply and demand, workers would have less reason to work hard because if they were fired, they could quickly find new jobs at the same wage. Therefore, firms may raise wages above the equilibrium level to incentivize workers not to shirk their responsibilities.

29-5e Worker Morale

A fifth and final type of efficiency-wage theory suggests that high wages improve worker morale and that content workers are more productive. This version deviates from economists' standard concept of rationality and is based on social norms and people's sense of fair play. Workers may consider it fair for a profitable firm to share its good fortune with its employees, even if that means paying an above-equilibrium wage. A firm may share that sense of fairness. Or it may simply recognize that its workers are less productive if they think they are being treated unfairly. If so, paying a high wage can be in the firm's best interest.

Case Study

Henry Ford and the Amazing $5-a-Day Wage

Henry Ford was a complex man. Historians tell us that he was a racist and anti-Semite, but he was also an industrial visionary. As founder of the Ford Motor Company, he introduced modern techniques of production. Rather than building cars with small teams of skilled craftsmen, Ford built cars on assembly lines in which unskilled workers were taught to perform the same simple tasks repeatedly. The output of this assembly process

was the Model T Ford, often considered the first affordable automobile, which made car travel accessible to middle-class families.

In 1914, Ford introduced another innovation: the $5 workday. This might not seem like much today, but back then, $5 was about twice the going wage. It was also far above the wage that balanced supply and demand. When the new $5-a-day wage was announced, long lines of job seekers formed outside the Ford factories. The number of workers willing to work at this wage far exceeded the number that Ford needed.

Ford's high-wage policy had many of the effects predicted by efficiency-wage theory. Turnover fell, absenteeism fell, and productivity rose. Workers were so much more efficient that Ford's production costs declined despite higher wages. Thus, paying a wage above the equilibrium level was profitable for the firm. A historian of the early Ford Motor Company wrote, "Ford and his associates freely declared on many occasions that the high-wage policy turned out to be good business. By this, they meant that it had improved the discipline of the workers, given them a more loyal interest in the institution, and raised their personal efficiency." Ford himself called the $5-a-day wage "one of the finest cost-cutting moves we ever made."

Why did it take a Henry Ford to introduce this efficiency wage? Why were other firms not already taking advantage of this seemingly profitable business strategy? According to some analysts, Ford's decision was linked to his early use of the assembly line. Workers organized in an assembly line are highly interdependent. If one worker is absent or works slowly, others are less able to complete their own tasks. While assembly lines made production more efficient, they also increased the importance of low worker turnover, high worker quality, and high worker effort. Paying efficiency wages may have been a better strategy for the Ford Motor Company than for other businesses at the time.

Above-equilibrium wages, however, are not unique to Ford. According to a 2018 article in *The California Sun*, the fast-food chain In-N-Out Burger paid its store managers on average more than $160,000 a year, about triple the industry average. Why? Denny Warnick, vice president of operations, said that the policy dated back to the company's founders, who wanted to make quality service a central focus. "Paying their associates well was just one way to help maintain that focus, and those beliefs remain firmly in place with us today," he said. Like Henry Ford, the owners of In-N-Out Burger seem to pay high wages to promote worker efficiency. ●

QuickQuiz

9. According to the theory of efficiency wages,
   a. firms may find it profitable to pay above-equilibrium wages.
   b. an excess supply of labor puts downward pressure on wages.
   c. sectoral shifts are the main source of frictional unemployment.
   d. right-to-work laws reduce the bargaining power of unions.

10. When a firm pays an efficiency wage, it may
    a. have trouble attracting enough workers.
    b. have to monitor its workers more closely.
    c. experience declines in worker quality.
    d. find that its workers quit less frequently.

Answers are at the end of the chapter.

In the News

Efficiency Wages in Practice

During the Covid recession of 2020, many companies cut or froze employees' pay. But research shows that paying workers more can sometimes boost the bottom line.

How Higher Wages Can Increase Profits

By Ray Fisman and Michael Luca

Pay cuts and salary freezes have become an unfortunate hallmark of the Covid-19 recession. Over seven million employees have seen their wages drop since March, and a great many others have had their pay frozen. But a handful of companies have bucked this trend and increased pay despite the economic crisis. In November, yogurt maker Chobani announced that it was raising its workers' lowest hourly wage from $13 to $15; the floor was set at $18 in high-cost centers like New York. E-furniture retailer Wayfair followed suit last week with a $15 minimum.

These and other companies present such wage hikes as enlightened capitalism—a way to help employees during difficult times while also buying loyalty and goodwill that translate into higher productivity and lower turnover. It's the latest entry in a century-old debate over whether companies can improve productivity and profits by paying their workers more. Economists refer to this possibility as the efficiency-wage theory: the idea that wages increased to above market level can effectively pay for themselves through increased worker motivation and retention. There is increasing evidence that efficiency-wage proponents may be right: Higher wages can at times boost the bottom line and—crucially for the current moment—pay cuts can elicit employee backlash and even sabotage.

The simple economics of efficiency wages were intuited by Henry Ford in 1914 with his idea of the $5 daily wage—more than double the pay at neighboring factories—for an eight-hour shift (down from the then-standard nine hours). Ford expected the high wages to make employees more engaged and harder-working, and if they couldn't meet his exacting standards, there was a long line of job-seekers outside the Highland Park, Mich., plant waiting to take their place.

Modern efficiency-wage theory is more subtle. For example, Nobel laureate George Akerlof (husband of incoming Treasury Secretary Janet Yellen, herself a pioneer in the study of efficiency wages) introduced the notion of "gift exchange": If employers are "nicer" than they need to be—by paying above-market wages, for example—workers will reciprocate by being more productive than is required merely to keep their jobs.

Some of the best evidence for the benefits of higher pay appears in a recently released working paper by Harvard University doctoral students Natalia Emanuel and Emma Harrington that examined wages and productivity among warehouse workers at a Fortune 500 online retailer (kept anonymous in the study). The researchers looked at the effects of a 2019 pay increase that looks a lot like the ones recently announced by Chobani and Wayfair—from about $16 an hour to $18. Prior to the increase, employees moved an average of 4.92 boxes per hour. A $1 pay increase boosted this figure by a third of a box. Higher wages also led to a large drop in employee turnover: a $1 increase reduced the quit rate by 19%.

Given the cost savings from not having to hire and train new employees, combined with improved productivity, the raise more than paid

29-6 Conclusion

This chapter discussed how unemployment is measured and why economies always experience some degree of unemployment. We have seen how job search, minimum-wage laws, unions, and efficiency wages can help explain why some workers do not have jobs. Which of these four explanations for the natural rate of unemployment are the most relevant for the U.S. economy and other economies around the world? There is no easy way to tell. Economists differ in which of these explanations of unemployment they emphasize.

The analysis in this chapter yields an important lesson: Although the economy will always have some unemployment, its natural rate changes over time. Many events and policies can alter the amount of unemployment the economy typically experiences. As the information revolution changes the process of job search, as Congress and state legislatures adjust the minimum wage, as workers form or quit unions, and as firms change their reliance on efficiency wages, the natural rate of unemployment evolves. Unemployment is not a simple problem with a simple solution. But how we choose to organize our society can profoundly influence how prevalent a problem it is.

for itself, boosting the company's bottom line in addition to improving employees' lives. The result wasn't only true for warehouse workers: The study found similar productivity and turnover improvements from higher wages for the company's customer service representatives as well.

The win-win of higher wages for this particular company suggests that pay had been set inefficiently low prior to the 2019 pay hike. If this mistake could be made by a Fortune 500 business, other businesses might do well to consider whether they could benefit from a similar policy.

Other research has found that the impact of pay increases depends in part on how they are communicated to employees. In work by one of us published in the journal Management Science in 2016 (with coauthors Duncan Gilchrist and Deepak Malhotra), we set out to understand why such pay raises have the potential to increase productivity. In an experiment conducted via the freelancing platform oDesk (now called Upwork), we found that workers hired for $4 an hour worked no harder than those hired for $3 an hour. (At the time, oDesk drew its largest share of freelancers from India, and both rates in the experiment were high compared to similar jobs on the platform.) However, giving employees an unexpected raise from $3 an hour to $4 after they were hired elicited greater effort.

For managers who are tempted to take advantage of slack in the labor market to cut wages to the bare minimum, it's worth considering evidence on the effects of pay cuts. A forthcoming paper in Management Science by Jason Sandvik, Richard Saouma, Nathan Seegert and Christopher Stanton looks at one company's decision to rebalance compensation in a way that ended up cutting pay for a subset of employees at one division. Other divisions did not rebalance their compensation at the same time and thus acted as a "control group" in the study.

The researchers obtained the human resources records of over 2,033 sales agents at the company (again kept anonymous by the researchers). Consistent with the findings of the Harvard study on warehouse workers, the authors found that employees who received pay cuts were more likely to leave the company. More troubling for company profits, the ones who left had been the most productive salespeople. So just as higher pay can pay for itself, lowering pay may have costs that offset much or all of the savings.

A recent working paper by Decio Coviello, Erika Deserranno and Nicola Persico offered evidence that pay cuts may even lead employees to be deliberately unproductive. The researchers found that phone sales representatives at a large U.S. retailer reacted to a 2014 pay cut by peddling items that customers didn't really want. How did they know? Because the drop in pay was accompanied by an increase in sales of items that were ultimately returned for refunds. Whether it was due to carelessness or sabotage, company revenues suffered.

Of course, not all firms can raise their wages in 2021. For businesses that are struggling to stay afloat, like the many mom-and-pop shops that have been hit hard by the Covid-19 recession, pay cuts may ultimately be the only option. And some businesses might choose to rely on extensive monitoring or complicated performance contracts to motivate workers instead.

But when times are tough, it's especially important to think carefully about both the costs and the benefits of higher wages. There's been considerable discussion about the moral case for higher wages, but there is a strong business case as well, since high wages have the potential to increase productivity and ultimately profits. In the transition to the new economic normal after the pandemic, doing what's right may also be what's best for the bottom line. ■

Mr. Fisman is a professor of economics at Boston University. Mr. Luca is a professor of business administration at Harvard Business School.

Source: *The Wall Street Journal*, January 23, 2021.

Chapter in a Nutshell

- The unemployment rate is the percentage of those who would like to work but don't have jobs. The Bureau of Labor Statistics calculates this statistic monthly based on a survey of thousands of households.
- The unemployment rate is an imperfect measure of joblessness. Some people who call themselves unemployed may actually not want to work, and some who would like to work are not counted as unemployed because they have left the labor force after an unsuccessful search.
- In the U.S. economy, most people who become unemployed find work within a short period of time. Nonetheless, most unemployment observed at any given time is attributable to the few people who are unemployed for long periods.
- One reason for unemployment is the time it takes workers to search for jobs that best suit their tastes and skills. Unemployment insurance, a government policy designed to protect the incomes of workers who lose their jobs, increases frictional unemployment by reducing the search effort of the unemployed.
- A second reason the economy always has some unemployment is minimum-wage laws. By raising the wage of unskilled and inexperienced workers above the equilibrium level, these laws increase the quantity of labor supplied and reduce the quantity

demanded. The resulting surplus of labor represents unemployment.

- A third reason for unemployment is the market power of unions. When unions push the wages in unionized industries above the equilibrium level, they create a surplus of labor.
- A fourth reason for unemployment is suggested by the theory of efficiency wages. According to this theory, firms find it profitable to pay wages above the equilibrium level. Higher wages can improve worker health, reduce worker turnover, raise worker quality, increase worker effort, and boost worker morale.

Key Concepts

labor force, p. 593
unemployment rate, p. 593
labor-force participation rate, p. 593
natural rate of unemployment, p. 594
cyclical unemployment, p. 594
discouraged workers, p. 596
frictional unemployment, p. 599
structural unemployment, p. 599
job search, p. 599
unemployment insurance, p. 600
union, p. 604
collective bargaining, p. 604
strike, p. 604
efficiency wages, p. 607

Questions for Review

1. What are the three categories into which the Bureau of Labor Statistics divides everyone? How does the BLS compute the labor force, the unemployment rate, and the labor-force participation rate?
2. Is unemployment typically short-term or long-term? Explain.
3. Why is frictional unemployment inevitable? How might the government reduce the amount of frictional unemployment?
4. Are minimum-wage laws a better explanation for structural unemployment among teenagers or among college graduates? Why?
5. How do unions affect the natural rate of unemployment?
6. What claims do advocates of unions make to argue that unions are good for the economy?
7. Explain five ways in which a firm might increase its profits by raising the wages it pays.

Problems and Applications

1. In April 2020, at the trough of the Covid-19 recession, the Bureau of Labor Statistics announced that of all adult Americans, 133,320,000 were employed, 23,038,000 were unemployed, and 103,538,000 were not in the labor force. Use this information to calculate:
   a. the adult population.
   b. the labor force.
   c. the labor-force participation rate.
   d. the unemployment rate.
2. Explain whether each of the following events increases, decreases, or has no effect on the unemployment rate and the labor-force participation rate.
   a. After a long search, Jon finds a job.
   b. Tyrion, a full-time college student, graduates and is immediately employed.
   c. After an unsuccessful job search, Arya gives up looking and retires.
   d. Daenerys quits her job to become a stay-at-home mom.
   e. Sansa has a birthday, becomes an adult, but has no interest in working.
   f. Jaime has a birthday, becomes an adult, and starts looking for a job.
   g. Cersei dies while enjoying retirement.
   h. Jorah dies working long hours at the office.
3. Go to the website of the Bureau of Labor Statistics (http://www.bls.gov). What is the national unemployment rate right now? Find the unemployment rate for the demographic group that best fits a description of you (for example, based on age, sex, and race). Is it higher or lower than the national average? Why do you think this is so?
4. Between January 2012 and January 2019, U.S. employment increased by 17.3 million workers, but the number of unemployed workers declined by only 6.3 million. How are these numbers consistent with each other? Why might one expect a reduction in the number of people counted as unemployed to be smaller than the increase in the number of people employed?

5. Economists use labor-market data to evaluate how well an economy is using its most valuable resource—its people. Two closely watched statistics are the unemployment rate and the employment–population ratio (calculated as the percentage of the adult population that is employed). Explain what happens to each of these statistics in the following scenarios. In your opinion, which statistic is the more meaningful gauge of how well the economy is doing?
   a. An auto company goes bankrupt and lays off its workers, who immediately start looking for new jobs.
   b. After an unsuccessful search, some of the laid-off workers quit looking for new jobs.
   c. Numerous students graduate from college but cannot find work.
   d. Numerous students graduate from college and immediately begin new jobs.
   e. A stock market boom induces newly enriched 60-year-old workers to take early retirement.
   f. Advances in healthcare prolong the life of many retirees.

6. Are the following workers more likely to experience short-term unemployment or long-term unemployment? Explain.
   a. a construction worker who is laid off because of bad weather
   b. a manufacturing worker who loses a job at a plant in an isolated area
   c. a stagecoach-industry worker who is laid off because of competition from railroads
   d. a short-order cook who loses a job when a new restaurant opens across the street
   e. an expert welder with little formal education who loses a job when the company installs automatic welding machinery

7. Using a diagram of the labor market, show the effect of an increase in the minimum wage on the wage paid to workers, the number of workers supplied, the number of workers demanded, and the amount of unemployment.

8. Consider an economy with two labor markets—one for manufacturing workers and one for service workers. Suppose initially that neither is unionized.
   a. If manufacturing workers formed a union, what would you expect to happen to the wages and employment in manufacturing?
   b. How would these changes in the manufacturing labor market affect the supply of labor in the market for service workers? What would happen to the equilibrium wage and employment in this labor market?

9. Structural unemployment is sometimes said to result from a mismatch between the job skills that employers want and the job skills that workers have. To explore this idea, consider an economy with two industries: auto manufacturing and aircraft manufacturing.
   a. If workers in these two industries require similar amounts of training, and if workers at the beginning of their careers can choose which industry to train for, what would you expect to happen to the wages in these two industries? How long would this process take? Explain.
   b. Suppose that one day, the economy opens itself to international trade and, as a result, starts importing autos and exporting aircraft. What would happen to the demand for labor in these two industries?
   c. Suppose that workers in one industry cannot be quickly retrained for the other. How would these shifts in demand affect equilibrium wages both in the short run and in the long run?
   d. If, for some reason, wages fail to adjust to the new equilibrium levels, what would occur?

10. Suppose that Congress passes a law requiring employers to provide employees some benefit (such as healthcare) that raises the cost of an employee by $4 per hour.
   a. What effect does this employer mandate have on the demand for labor? (In answering this and the following questions, be quantitative when you can.)
   b. If employees place a value on this benefit exactly equal to its cost, what effect does this employer mandate have on the supply of labor?
   c. If the wage can freely adjust to balance supply and demand, how does this law affect the wage and the level of employment? Are employers better or worse off? Are employees better or worse off?
   d. Suppose that, before the mandate, the wage in this market was $3 above the minimum wage. In this case, how does the employer mandate affect the wage, the level of employment, and the level of unemployment?
   e. Now suppose that workers do not value the mandated benefit at all. How does this alternative assumption change your answers to parts (b) and (c)?

QuickQuiz Answers

1. **a** 2. **c** 3. **b** 4. **a** 5. **b** 6. **a** 7. **c** 8. **d** 9. **a** 10. **d**

Chapter

30

The Monetary System

When you buy a meal at Rosie's Restaurant, you get something of value—a full stomach. To pay for this service, you might hand Rosie several worn-out pieces of greenish paper decorated with strange symbols, government buildings, and the portraits of famous dead Americans. Or you might hand her a single piece of paper with a bank's name and your autograph. Or you may compensate Rosie using a plastic card or a cell phone app. Whether you pay by cash, check, debit card, or mobile transfer, the restaurateur is happy to work hard to satisfy your gastronomical desires in exchange for these tokens of value, which, in and of themselves, are worthless.

Anyone who has lived in a modern economy is familiar with this social custom. Even though paper currency and digital debits have no intrinsic value, Rosie is confident that, in the future, some third person will accept them in exchange for something that she does value. And that third person is confident that some fourth person will accept the money, with the knowledge that yet a fifth person will accept the money, and so on. To Rosie and other people, your cash, check, debit card charge, or electronic payment represents a claim to goods and services in the future.

AZMANJAKA/E+/GETTY IMAGES ; CARLOS FLORES/EYEEM/GETTY IMAGES

The social custom of using money for transactions is extraordinarily useful. Imagine, for a moment, that an economy had no item widely accepted in exchange for goods and services. People would have to rely on **barter**—the exchange of one good or service for another—to obtain the things they need. To get a meal at Rosie's Restaurant, for instance, you would have to offer Rosie something of immediate value to her. You could offer to wash some dishes, mow her lawn, or give her your family's secret recipe for meatloaf. An economy that relies on barter has trouble allocating its scarce resources efficiently. In such an economy, trade requires the **double coincidence of wants**—the unlikely occurrence that two people each have a good or service that the other wants.

The existence of money makes trade much easier. Rosie does not care whether you can produce a valuable good or service for her. She is happy to accept your money, knowing that other people will do the same for her. This convention allows trade to be roundabout. Rosie accepts your money and uses it to pay Tony, her chef; Tony uses his paycheck to send his daughter Ava to day care; the day care center uses Ava's tuition to pay Mira, a teacher; and Mira hires you to mow her lawn. As money flows from person to person, it facilitates production and trade, allowing people to specialize in what they do best and raising everyone's standard of living.

This chapter begins to examine the role of money in an economy. We discuss what money is, the forms that money takes, how the banking system helps create money, and how the government controls the quantity of money in circulation. The rest of this book devotes much effort to studying how changes in the quantity of money affect various economic variables, including inflation, interest rates, production, and employment. Consistent with the long-run focus of the previous four chapters, the next chapter examines the long-run effects of changes in the quantity of money. The short-run effects of monetary changes are a more complex topic, which we take up later. This chapter builds the foundation for this further analysis.

30-1 The Meaning of Money

What is money? This might seem like an odd question. When you read that Jeff Bezos has a lot of money, you know what that means: He is so rich that he can buy almost anything he wants. In this sense, the term **money** is used to mean **wealth**.

money
the set of assets in an economy that people regularly use to buy goods and services

Economists, however, use the word in a more specific sense: **Money** is the set of assets in the economy that people regularly use to buy goods and services from each other. The cash in your wallet is money because you can use it to buy a meal at a restaurant or a shirt at a store. By contrast, the large share of Amazon that makes up much of Jeff Bezos's wealth is not considered a form of money. Mr. Bezos could not buy a meal or a shirt with this wealth without first obtaining some cash. According to the economist's definition, money includes only those few types of wealth that sellers regularly accept in exchange for goods and services.

30-1a The Functions of Money

Money has three functions: It is a **medium of exchange**, a **unit of account**, and a **store of value**. These functions together distinguish money from other assets, such as stocks, bonds, real estate, art, and even baseball cards. Let's examine each of these three functions.

medium of exchange
an item that buyers give to sellers when they want to purchase goods and services

A **medium of exchange** is an item that buyers give to sellers when they purchase goods and services. When you buy a shirt, the store gives you the shirt, and you give the store some money. This transfer of money from buyer to seller allows the

transaction to take place. When you go shopping, you are confident that stores will accept your money for the items they sell because money is the commonly accepted medium of exchange.

A **unit of account** is the yardstick people use to post prices and record debts. When you shop, you might find that a shirt costs $60 and a hamburger costs $6. It would be accurate to say that a shirt costs 10 hamburgers and a hamburger costs 1/10 of a shirt, but prices are rarely quoted in this way. Similarly, if you borrow from a bank, your future loan repayments will be set in dollars, not in a quantity of goods and services. When most people want to measure and record economic value, they use money as the unit of account.

unit of account
the yardstick people use to post prices and record debts

A **store of value** is an item that people can use to transfer purchasing power from the present to the future. When sellers accept money today in exchange for goods and services, they can hold the money and become buyers at another time. Money is not the only store of value in the economy: A person can also transfer purchasing power from the present to the future by holding nonmonetary assets such as stocks and bonds. The term **wealth** refers to the total of all stores of value, including both monetary and nonmonetary assets.

store of value
an item that people can use to transfer purchasing power from the present to the future

Economists use the term **liquidity** to describe the ease with which an asset can be converted into the economy's medium of exchange. Because money is the economy's medium of exchange, it is the most liquid asset available. Other assets vary widely in liquidity. Most stocks and bonds can be sold easily at a low cost, so they are relatively liquid assets. By contrast, selling a house, a Rembrandt painting, or a 1948 Joe DiMaggio baseball card requires more time and effort, so these assets are less liquid.

liquidity
the ease with which an asset can be converted into the economy's medium of exchange

When people decide how to allocate wealth, they balance the liquidity of each possible asset against the asset's usefulness as a store of value. Money is the most liquid asset, but it is far from perfect as a store of value. When prices rise, the value of money falls. In other words, when goods and services become more expensive, a dollar buys less. This link between the price level and the value of money is key to understanding how money affects the economy, a topic we start to explore in the next chapter.

30-1b The Kinds of Money

When money takes the form of a commodity with intrinsic value, it is called **commodity money**. The term **intrinsic value** means that the item would have value even if it were not used as money. An example of commodity money is gold. Gold has intrinsic value because it is used in industry and for making jewelry. Today, we no longer use gold as money, but historically, gold was a common form of money because it is relatively easy to carry, measure, and verify for impurities. When an economy uses gold as money (or uses paper money that is convertible into a specific amount of gold on demand), it is said to be operating under a **gold standard**.

commodity money
money that takes the form of a commodity with intrinsic value

Another example of commodity money is cigarettes. In prisoner-of-war camps during World War II, prisoners traded goods and services with one another using cigarettes as the store of value, unit of account, and medium of exchange. Similarly, as the Soviet Union was breaking up in the late 1980s, cigarettes sometimes replaced the ruble as the preferred currency in Moscow. In both cases, nonsmokers accepted cigarettes in an exchange, knowing they could use them to buy other goods and services.

Money without intrinsic value is called **fiat money**. A **fiat** is an order or decree, and fiat money is established as money by government decree. For example, compare the paper dollars in your wallet (printed by the U.S. government) with the paper dollars from a game of Monopoly (printed by the Hasbro game company). Why can you use the first to pay your bill at a restaurant but not the second?

fiat money
money without intrinsic value that is used as money by government decree

FYI

Cryptocurrencies: A Fad or the Future?

In recent years, the world has seen a proliferation of a new kind of money called **cryptocurrencies**. These currencies rely on cryptography—the science of encoding information—to create a medium of exchange that exists only in electronic form. They use a technology called **blockchain** to maintain a public ledger that records transactions.

The first of these cryptocurrencies, introduced in 2009, was **bitcoin**. It was conceived by a person or group who used the name Satoshi Nakamoto. Nakamoto authored and circulated a white paper establishing the bitcoin protocol, but Nakamoto's identity is otherwise unknown. According to the protocol, people create bitcoins by using computers to solve complex mathematical problems. The number of bitcoins that can be "mined" in this way is supposedly limited to 21 million units. Once created, bitcoins can be used in exchange. They can be bought and sold for U.S. dollars on organized bitcoin exchanges, where supply and demand determine the dollar price of a bitcoin. People can hold bitcoins as a store of value, and they can use bitcoins to buy things from any vendor who is willing to accept them. In essence, Nakamoto created a virtual asset from nothing at all and ensured holders of it that its supply would be forever limited.

Bitcoins are neither commodity money nor fiat money. Unlike commodity money, they have no intrinsic value. You can't use bitcoins for anything other than exchange. Unlike fiat money, they are not created by government decree. Fans of bitcoin embrace this new form of money because it exists apart from government. Some bitcoin users are engaged in illicit transactions such as the drug trade and benefit from the anonymity that bitcoin transactions offer.

During bitcoin's brief history, its dollar value has fluctuated wildly. In 2010, the price of a bitcoin ranged between 5 cents to 39 cents. The price rose above $1 in 2011 and above $1,000 in 2013 before falling below $500 in 2014. Over the following years, the dollar value of a bitcoin skyrocketed, though the ride was bumpy. Bitcoin reached $60,000 in April 2021, fell back to $31,000 in July 2021, rose to $67,000 in November 2021, and then fell again to $20,000 in July 2022. Meanwhile, other cryptocurrencies were introduced, such as Ethereum, Litecoin, Ripple, and Zcash, providing competition for bitcoin. These other cryptocurrencies differ from bitcoin in the details of their protocols, but like bitcoin, they have all exhibited large price swings. Some newer cryptocurrencies, known as stablecoin, are pegged to the dollar. Yet sometimes the peg has been impossible to sustain.

The long-term success of cryptocurrencies depends on whether they succeed in performing the functions of money: a store of value, a unit of account, and a medium of exchange. Many economists are skeptical. The great volatility of the dollar prices of most cryptocurrencies makes them a risky way to hold wealth and an inconvenient measure in which to post prices. Few retailers accept them in exchange, at least so far. As a result, cryptocurrencies have been excluded from standard measures of the quantity of money.

Cryptocurrencies may be the money of the future or a passing fad. Or they may become just a new niche asset class. ■

The answer is that the U.S. government has declared its dollars to be valid money. Each paper dollar in your wallet reads: "This note is legal tender for all debts, public and private."

Although the government is central to establishing and regulating a system of fiat money (by prosecuting counterfeiters, for example), other factors are also required for the success of a monetary system. To a large extent, the acceptance of an asset as money depends as much on expectations and social convention as on government decree. The Soviet government in the 1980s didn't abandon the ruble as the official currency. Yet the people of Moscow often preferred American dollars or cigarettes because they believed that these alternative monies were more reliable stores of value and would continue to be accepted as mediums of exchange.

30-1c Money in the U.S. Economy

As we will see, the quantity of money circulating in the economy, called the **money stock**, has a powerful influence on many economic variables. But first, we need to ask a basic question: What is the quantity of money? Suppose you were given the task of measuring how much money there is in the U.S. economy. What would you include in your measure?

The most obvious asset to include is **currency**—the paper bills and coins in public hands. Currency is still the most widely accepted medium of exchange in our economy. There is no doubt that it is part of the money stock.

currency
the paper bills and coins in the hands of the public

Yet currency is not the only asset that you can use to buy goods and services. Many businesses also accept personal checks. Wealth held in your checking account is almost as convenient for buying things as wealth held in your wallet. To measure the money stock, therefore, you might want to add **demand deposits**—balances in bank accounts that depositors can access on demand simply by writing a check or swiping a debit card.

demand deposits
balances in bank accounts that depositors can access on demand by writing a check

Once you include balances in checking accounts in the money stock, you are led to consider other accounts held at banks and other financial institutions. Bank depositors usually cannot write checks against the balances in their savings accounts, but they can easily transfer funds from savings into checking accounts. In addition, depositors in money market mutual funds can often write checks against their balances or use them to pay bills electronically, though the number and size of these payments may be limited. Arguably, these other accounts should be counted as part of the U.S. money stock.

In a complex economy, it is hard to clearly distinguish between assets that can be called "money" and those that cannot. Without a doubt, the coins in your pocket are part of the money stock, and the Empire State Building is not. But many assets lie between these extremes, making the choice less clear. Because different analysts can reasonably disagree about where to draw the line between monetary and nonmonetary assets, various measures of the money stock are available.

For the U.S. economy, the two most widely followed measures of the money stock are M1 and M2. M1 includes currency, demand deposits at banks, and some other liquid deposits such as the balances in savings accounts. M2 includes everything in M1 plus small time deposits and money market funds (except those held in restricted retirement accounts). For the purposes of this book, the differences between the various measures of money aren't crucial. The important point is that the money stock includes not only currency but also deposits in banks and other financial institutions that can be readily accessed and used to buy goods and services.

Why Credit Cards Aren't Money

It might seem natural to include credit cards as part of the economy's stock of money. After all, people often use credit cards to make purchases. Aren't credit cards, therefore, a medium of exchange?

At first, this argument may seem persuasive, but credit cards are excluded from all measures of the quantity of money. The reason is that credit cards are not really a method of payment but rather a way of **deferring** payment. When you buy a meal with a credit card, the bank that issued the card pays the restaurant what it is due. At a later date, you will repay the bank (perhaps with interest). When the time comes to pay your credit card bill, you will probably do so by writing a check against your checking account or by drawing money from it electronically. The balance in this checking account is part of the economy's stock of money.

Notice that despite their physical similarity, credit cards are different from debit cards, which automatically withdraw funds from a bank account to pay for items bought. Rather than allowing the user to postpone payment for a purchase, a debit card gives the user immediate access to deposits in a bank account. In this sense, a debit card is more like a check than a credit card. The account balances that lie behind debit cards are included in measures of the quantity of money.

Even though credit cards are not considered a form of money, they are important for analyzing the monetary system. People who have credit cards can pay many of their bills together at the end of the month rather than sporadically as they make purchases. As a result, people with credit cards probably hold less money on average than those without credit cards. The wide availability of credit cards—along with the electronic payment systems often linked to them—may reduce the amount of money that people choose to hold. ■

Case Study

Where Is All the Currency?

One puzzle about the money stock of the U.S. economy concerns the amount of currency. In November 2021, there was $2.1 trillion of currency outstanding. To put this number in perspective, we can divide it by 262 million, the number of adults (age 16 and older) in the United States. This calculation implies that there is more than $8,000 of currency per adult. Most people are surprised by this figure because they carry far less in their wallets.

Who is holding all this currency? No one knows for sure, but there are two plausible explanations.

The first is that much of the currency is held abroad. In foreign countries without a stable monetary system, people often prefer U.S. dollars to domestic assets. Estimates suggest that more than half of U.S. dollars circulate outside the United States.

The second explanation is that much of the currency is held by drug dealers, tax evaders, and other criminals. For most people in the U.S. economy, currency is not a particularly good way to hold wealth: Not only can currency be lost or stolen, but it also does not earn interest, whereas a bank deposit does. Thus, most people hold only small amounts of currency. But criminals find currency more appealing. They may avoid putting their money in banks, for instance, because a bank deposit gives police a paper trail they can use to trace illegal activities. For criminals, currency may be the best store of value available. ●

QuickQuiz

1. Fiat money is
   a. a type of money with intrinsic value.
   b. a type of money set by government decree.
   c. any asset used as the medium of exchange.
   d. any asset used as the unit of account.

2. The money stock includes all of the following EXCEPT
   a. metal coins.
   b. paper currency.
   c. lines of credit accessible with credit cards.
   d. bank balances accessible with debit cards.

Answers are at the end of the chapter.

30-2 The Federal Reserve System

Federal Reserve (Fed) the central bank of the United States

central bank an institution designed to oversee the banking system and regulate the quantity of money in the economy

Whenever an economy uses a system of fiat money, as the U.S. economy does, some agency must be responsible for managing the system. In the United States, that entity is the **Federal Reserve**, often simply called the **Fed**. If you look at the top of a dollar bill, you will see that it is called a "Federal Reserve Note." The Fed is an example of a **central bank**—an institution designed to oversee the banking system and regulate the quantity of money. Other major central banks around the world include the Bank of England, the Bank of Japan, and the European Central Bank.

30-2a The Fed's Organization

The Federal Reserve was created in 1913 after a series of bank failures in 1907 convinced Congress that the United States needed a central bank to ensure the health of the nation's banking system. Today, the Fed is run by its Board of Governors, which has up to seven members appointed by the president and confirmed by the Senate. The governors have 14-year terms. Just as federal judges are given lifetime appointments

to insulate them from politics, Fed governors are given long terms to give them independence from short-term political pressures when they formulate monetary policy.

Among the members of the Board of Governors, the most important is the chair. The chair directs the Fed staff, presides over board meetings, and testifies regularly about Fed policy in front of congressional committees. The president appoints the chair to a 4-year term. As this book was going to press, the chair of the Fed was Jerome Powell, who was nominated to the job by President Donald Trump in 2017 and reappointed by President Joe Biden in 2021.

The Federal Reserve System consists of the Federal Reserve Board in Washington, D.C., and twelve regional Federal Reserve Banks located in major cities around the country. The presidents of the regional banks are chosen by each bank's board of directors, whose members are typically drawn from the region's banking and business community.

The Fed has two related jobs. The first is to regulate banks and ensure the health of the banking system. In particular, the Fed monitors each bank's financial condition and facilitates bank transactions by clearing checks. It also acts as a bank's bank. That is, the Fed makes loans to banks when banks themselves need funds. The Fed acts as a **lender of last resort**—a lender to those who cannot borrow anywhere else—to maintain stability in the overall banking system.

The Fed's second job is to control the quantity of money available in the economy, called the **money supply**, which in the short run is closely connected to the level of interest rates. Decisions by policymakers concerning the money supply and interest rates constitute **monetary policy**. At the Federal Reserve, monetary policy is made by the Federal Open Market Committee (FOMC). The FOMC meets about every six weeks in Washington, D.C., to discuss the condition of the economy and consider changes in monetary policy.

money supply
the quantity of money available in the economy

monetary policy
the setting of the money supply by policymakers in the central bank

30-2b The Federal Open Market Committee

The Federal Open Market Committee consists of the members of the Board of Governors and five of the twelve regional bank presidents. All twelve regional presidents attend each FOMC meeting, but only five get to vote. Voting rights rotate among the twelve regional presidents. The president of the New York Fed always gets a vote, however, because New York is the country's traditional financial center and because all Fed purchases and sales of government bonds are conducted at the New York Fed's trading desk.

Through the decisions of the FOMC, the Fed has the power to increase or decrease the number of dollars in the economy. In simple metaphorical terms, imagine the Fed expanding the money supply by printing dollar bills and dropping them around the country from a helicopter. Similarly, imagine it reducing the money supply with a giant vacuum cleaner that sucks dollar bills out of people's wallets. In reality, the Fed's methods for changing the money supply are complex and subtle, but the helicopter-vacuum metaphor is a good first step to an understanding of monetary policy.

Later, this chapter discusses the many ways the Fed can influence the money supply without using a helicopter or vacuum, but it is worth noting here that the Fed's primary tool has historically been the **open-market operation**—the purchase and sale of U.S. government bonds. Recall that a U.S. government bond is a certificate of indebtedness of the federal government. To increase the money supply, the Fed can create dollars and use them to buy government bonds from the public in the bond markets. After the purchase, these dollars are in public hands. In this way, the Fed's open-market purchase of bonds increases the money supply. Conversely, to decrease the money supply, the Fed can sell government bonds from its portfolio

to the public in the bond markets. After the sale, the dollars the Fed receives for the bonds are out of public hands. The Fed's open-market sale of bonds thereby decreases the money supply. In recent years, the Fed has relied less on open-market operations and more on other tools of monetary policy, as we will see.

Central banks are important institutions because changes in the money supply can profoundly affect the economy. One of the **Ten Principles of Economics** in Chapter 1 is that prices rise when the government prints too much money. Another of the **Ten Principles of Economics** is that society faces a short-run trade-off between inflation and unemployment. The power of the Fed rests on these principles. For reasons discussed more fully in coming chapters, the Fed's policy decisions are key determinants of inflation in the long run and employment and production in the short run. The Fed chair has been called the second most powerful person in the United States.

QuickQuiz

3. Which of the following is NOT true about the Federal Reserve?
   a. It was established by the U.S. Constitution.
   b. It regulates the banking system.
   c. It lends to banks.
   d. It can own government bonds.

4. If the Fed wants to increase the money supply, it can
   a. raise income tax rates.
   b. reduce income tax rates.
   c. buy bonds in open-market operations.
   d. sell bonds in open-market operations.

Answers are at the end of the chapter.

30-3 Banks and the Money Supply

MICK STEVENS

"I've heard a lot about money, and now I'd like to try some."

So far, we have introduced the concept of "money" and discussed how the Fed controls the money supply. This explanation is correct, but it is incomplete because it omits the key role that banks play in the monetary system.

Recall that the money you hold includes both currency (the bills in your wallet and coins in your pocket) and demand deposits (the balance in your checking account). Because demand deposits are held in banks, the behavior of banks can influence the quantity of demand deposits and, therefore, the money supply. This section examines how banks affect the money supply and complicate the Fed's job of controlling the money supply.

30-3a The Simple Case of 100-Percent-Reserve Banking

To see how banks influence the money supply, let's first imagine a world without any banks at all. In this simple world, currency is the only form of money. To be concrete, suppose that the total quantity of currency is $100. The supply of money is, therefore, $100.

Now suppose that someone opens a bank, appropriately called First National Bank. First National Bank is only a depository institution—that is, it accepts deposits but does not make loans. The purpose of the bank is to give depositors a safe place to keep their money. Whenever people make a deposit, the bank keeps the money in its vault until the depositors withdraw it, write checks, or use a debit card to access their balances. Deposits that banks have received but have not loaned out are called **reserves**. In this imaginary economy, all deposits are held as reserves, so this system is called **100-percent-reserve banking**.

reserves
deposits that banks have received but have not loaned out

We can express the financial position of First National Bank with a **T-account**, which is a simplified accounting statement that shows a bank's assets and liabilities. Here is the T-account for First National Bank if the economy's entire $100 of money is deposited in the bank:

First National Bank			
Assets		**Liabilities**	
Reserves	$100.00	Deposits	$100.00

On the left side of the T-account are the bank's assets of $100 (the reserves it holds in its vaults). On the right side are the bank's liabilities of $100 (the amount it owes to its depositors). Because the assets and liabilities balance, this accounting statement is called a **balance sheet**.

Now consider this economy's money supply. Before the start of First National Bank, it is the $100 of currency that people are holding. After the bank opens and people deposit their currency, the money supply is the $100 of demand deposits. (There is no longer any currency outstanding because it is all in the bank vault.) Each deposit reduces currency and raises demand deposits by exactly the same amount, leaving the money supply unchanged. **Thus, if banks hold all deposits in reserve, banks do not influence the supply of money.**

30-3b Money Creation with Fractional-Reserve Banking

Eventually, the people who run First National Bank may reconsider their policy of 100-percent-reserve banking. Leaving all that money idle in their vaults seems unnecessary. Why not lend some of it out and earn a profit by charging interest on the loans? Families buying houses, firms building new factories, and students paying for college would all be willing to pay interest to borrow some of that money for a while. First National Bank has to keep some reserves so that currency is available if depositors want to make withdrawals. But if the flow of new deposits is roughly the same as the flow of withdrawals, First National needs to keep only a fraction of its deposits in reserve. And so First National adopts a system called **fractional-reserve banking**.

fractional-reserve banking
a banking system in which banks hold only a fraction of deposits as reserves

The fraction of total deposits that a bank holds as reserves is called the **reserve ratio**. This ratio is determined by a combination of government regulation and bank policy. As the chapter will discuss more fully later, the Fed traditionally set a minimum amount of reserves that banks must hold, called a **reserve requirement**. In addition, banks may hold reserves above the legal minimum, called **excess reserves**, so they can be more confident that they will not run short of cash. For our purpose here, we take the reserve ratio as given to examine how fractional-reserve banking influences the money supply.

reserve ratio
the fraction of deposits that banks hold as reserves

Suppose that First National has a reserve ratio of 1/10, or 10 percent. This means that it keeps 10 percent of its deposits in reserve and lends out the rest. Now look again at the bank's T-account:

First National Bank			
Assets		**Liabilities**	
Reserves	$10.00	Deposits	$100.00
Loans	90.00		

First National still has $100 in liabilities because making the loans did not alter the bank's obligation to its depositors. But now, the bank has two kinds of assets: $10 of reserves in its vault and $90 of loans. (These loans are liabilities of the people borrowing from First National, but they are assets of the bank because the borrowers will repay the loans.) In total, First National's assets still equal its liabilities.

Once again, consider the economy's supply of money. Before First National makes any loans, the money supply is the $100 of deposits. Yet when First National lends out some of these deposits, the money supply increases. The depositors still have demand deposits totaling $100, but now the borrowers hold $90 in currency. The money supply (which equals currency plus demand deposits) equals $190. **Thus, when banks hold only a fraction of deposits in reserve, the banking system creates money.**

At first, this creation of money by fractional-reserve banking may seem too good to be true: It appears that the bank has created money out of thin air. To make this feat seem less miraculous, note that when First National Bank lends out some of its reserves and creates money, it does not create any wealth. Loans from First National give the borrowers some currency, which augments their ability to buy goods and services. Yet the borrowers are also incurring debts, so the loans do not make them richer. In other words, as a bank creates the asset of money, it also creates a corresponding liability for those who borrowed the created money. At the end of this process of money creation, the economy is more liquid in the sense that there is more of the medium of exchange, but the economy is no wealthier than before.

30-3c The Money Multiplier

The creation of money does not stop with First National Bank. Suppose the borrower from First National uses the $90 to buy something from someone who then deposits the currency in Second National Bank. Here is the T-account for Second National Bank:

Second National Bank			
Assets		**Liabilities**	
Reserves	$ 9.00	Deposits	$90.00
Loans	81.00		

After the deposit, Second National has liabilities of $90. If Second National also has a reserve ratio of 10 percent, it keeps assets of $9 in reserve and makes $81 in loans. In this way, Second National creates an additional $81 of money. If this $81 is eventually deposited in Third National Bank, which also has a reserve ratio of 10 percent, Third National keeps $8.10 in reserve and makes $72.90 in loans. Here is the T-account for Third National Bank:

Third National Bank			
Assets		**Liabilities**	
Reserves	$ 8.10	Deposits	$81.00
Loans	72.90		

The process goes on and on. Each time that money is deposited and a bank loan is made, more money is created.

After all this money creation, how much money is there in this economy? Let's add it up:

Original deposit	= \$100.00
First National lending	= \$90.00 (= .9 × \$100.00)
Second National lending	= \$81.00 (= .9 × \$90.00)
Third National lending	= \$72.90 (= .9 × \$81.00)
•	•
•	•
•	•
Total money supply	= \$1,000.00

It turns out that even though this process of money creation can continue forever, it does not create an infinite amount of money. If you laboriously add the infinite sequence of numbers in the preceding example, you find that the \$100 of reserves generates \$1,000 of money. The amount of money that results from each dollar of reserves is called the **money multiplier**. In this economy, where the \$100 of reserves generates \$1,000 of money, the money multiplier is 10.

money multiplier the amount of money that results from each dollar of reserves

What determines the size of the money multiplier? It turns out the answer is simple: **The money multiplier is the reciprocal of the reserve ratio.** If R is the reserve ratio for all banks in the economy, then each dollar of reserves generates $1/R$ dollars of money. In our example, $R = 1/10$, so the money multiplier is 10.

Consider why this reciprocal formula for the money multiplier makes sense. If a bank holds \$1,000 in deposits, then a reserve ratio of 1/10 (10 percent) means that the bank must hold \$100 in reserves. The money multiplier turns this idea around: If the banking system as a whole holds a total of \$100 in reserves, it can have only \$1,000 in deposits. In other words, if R is the ratio of reserves to deposits at each bank (that is, the reserve ratio), then the ratio of deposits to reserves in the banking system (that is, the money multiplier) must be $1/R$.

This formula shows how the amount of money banks create depends on the reserve ratio. If the reserve ratio were only 1/20 (5 percent), then the banking system would have 20 times as much in deposits as in reserves, implying a money multiplier of 20. Each dollar of reserves would generate \$20 of money. Similarly, if the reserve ratio were 1/4 (25 percent), deposits would be 4 times reserves, the money multiplier would be 4, and each dollar of reserves would generate \$4 of money. **The higher the reserve ratio, the less of each deposit banks loan out, and the smaller the money multiplier.** In the special case of 100-percent-reserve banking, the reserve ratio is 1, the money multiplier is 1, and banks do not make loans or create money.

30-3d Bank Capital, Leverage, and the Financial Crisis of 2008–2009

The previous sections presented a simplified explanation of how banks work. The reality of modern banking is more complex, and this complexity played an important role in the financial crisis of 2008 and 2009. To understand such financial crises, we need a bit more background about how banks actually function.

In the bank balance sheets examined so far, a bank accepts deposits and either uses those deposits to make loans or holds them as reserves. More realistically, though, a bank gets financial resources not only from accepting deposits but also, like other companies, from issuing equity and debt. The resources that a bank obtains from

bank capital
the resources a bank's owners have put into the institution

issuing equity to its owners are called **bank capital**. A bank uses these financial resources in a variety of ways to generate profit for its owners. In addition to making loans and holding reserves, it can buy financial securities, such as stocks and bonds.

Here is a somewhat more realistic example of a bank's balance sheet:

More Realistic National Bank			
Assets		**Liabilities and Owners' Equity**	
Reserves	$200	Deposits	$800
Loans	700	Debt	150
Securities	100	Capital (owners' equity)	50

On the right side of this balance sheet are the bank's liabilities and capital (also called **owners' equity**). This bank obtained $50 in capital from its owners. It also issued $150 of debt and took in $800 of deposits. It used the total of $1,000 in three ways, listed on the left side of the balance sheet, which shows the bank's assets. This bank held $200 in reserves, made $700 in loans, and used $100 to buy financial securities, such as government or corporate bonds. The bank decides how to allocate its resources among asset classes based on their risk and return as well as on any regulations that restrict its choices.

By the rules of accounting, the reserves, loans, and securities on the left side of the balance sheet always equal, in total, the deposits, debt, and capital on the right side. There is no magic in this equality. It occurs because the value of the owners' equity is, by definition, the value of the bank's assets (reserves, loans, and securities) minus the value of its liabilities (deposits and debt). Therefore, the left and right sides of the balance sheet always sum to the same total.

leverage
the use of borrowed money to supplement existing funds for investment purposes

Many businesses rely on **leverage**, the use of borrowed money to supplement existing funds for the purposes of investment. Whenever a business uses debt to finance an investment project, it is applying leverage. For banks, leverage is particularly important because borrowing and lending are at the heart of what they do. To better understand banking, let's look more closely at how leverage works.

leverage ratio
the ratio of assets to bank capital

The **leverage ratio** is the ratio of the bank's total assets to bank capital. In this example, the leverage ratio is $1,000/$50, or 20. A leverage ratio of 20 means that for every dollar of capital the bank owners have contributed, the bank has $20 of assets. Of the $20 of assets, $19 are financed with borrowed money—either by taking in deposits or issuing debt.

You may have learned in a science class that a lever can amplify a force: A boulder you cannot move with your arms alone will move more easily if you use a lever. A similar amplification occurs with bank leverage. To see how, suppose that the bank's assets increase in value by 5 percent because, say, some of the securities held by the bank rise in price. The $1,000 of assets is now worth $1,050. Because the depositors and debt holders are still owed $950, bank capital increases from $50 to $100 ($1050 – $950). Thus, when the leverage ratio is 20, a 5-percent appreciation in asset values increases the owners' equity by 100 percent.

The same principle works on the downside but with troubling consequences. Suppose that some people who borrowed from the bank default on their loans, reducing the value of the bank's assets by 5 percent to $950. Once again, the depositors and debt holders are still owed $950, so the value of the owners' equity falls to zero ($950 – $950). Thus, when the leverage ratio is 20, a 5-percent fall in the value of the bank's assets leads to a 100-percent fall in bank capital. If the value of assets

were to fall by more than 5 percent, the bank's assets would fall below its liabilities. In this case, the bank would be **insolvent**, and it would be unable to pay off its debt holders and depositors in full.

Bank regulators require banks to hold a certain amount of capital. The goal of such a **capital requirement** is to ensure that banks will be able to pay off their depositors (without having to resort to government-provided deposit insurance funds). The amount of capital required depends on the kind of assets a bank holds. Other things being equal, a bank that holds risky assets such as loans to borrowers whose credit is of dubious quality is required to hold more capital than a bank that holds safe assets such as government bonds.

capital requirement a government regulation specifying a minimum amount of bank capital

Economic turmoil can result when banks find themselves with too little capital to satisfy capital requirements. An example of this phenomenon arose in 2007 and 2008, when many banks incurred sizable losses on some of their assets—specifically, mortgage loans and securities backed by mortgage loans. The capital shortage induced banks to reduce lending, a phenomenon called a **credit crunch**, which contributed to a severe downturn in economic activity. (Chapter 34 discusses this event more fully.) To address the problem, the U.S. Treasury, working with the Fed, put many billions of dollars of public funds into the banking system to increase bank capital. As a result, the U.S. taxpayer temporarily became a part owner of many banks. The goal of this unusual policy was to recapitalize the banking system so that lending could return to a more normal level. And, in fact, by late 2009, it did.

QuickQuiz

5. Isabella takes $100 of currency from her wallet and deposits it into her checking account. If the bank adds the entire $100 to reserves, the money supply _____, but if the bank lends out some of the $100, the money supply _____.
   a. increases; increases even more
   b. increases; increases by less
   c. is unchanged; increases
   d. decreases; decreases by less
6. If the reserve ratio is ¼ and the central bank increases the quantity of reserves in the banking system by $120, the money supply increases by
   a. $90.
   b. $150.
   c. $160.
   d. $480.
7. A bank has capital of $200 and a leverage ratio of 5. If the value of the bank's assets declines by 10 percent, then its capital will be reduced to
   a. $100.
   b. $150.
   c. $180.
   d. $185.

Answers are at the end of the chapter.

30-4 The Fed's Tools of Monetary Control

The Federal Reserve is responsible for controlling the supply of money, and it carries out this job in subtle and evolving ways. Because banks create money in a system of fractional-reserve banking, the Fed's control of the money supply is indirect. When the Fed decides to change the money supply, it must consider how its actions will work through the banking system.

The Fed has a variety of tools in its monetary toolbox, and, over time, its use of them changes. We can group the tools into two groups: those that influence the quantity of reserves and those that influence the reserve ratio and thereby the money multiplier.

30-4a How the Fed Influences the Quantity of Reserves

The first way the Fed can alter the money supply is by changing the quantity of reserves. The Fed does this by conducting open-market operations or by lending to banks.

open-market operations
the purchase and sale of U.S. government bonds by the Fed

Open-Market Operations As noted earlier, the Fed conducts **open-market operations** when it buys or sells government bonds. To increase the money supply, it instructs its bond traders at the New York Fed to buy bonds from the public in the nation's bond markets. By paying for the bonds with newly created dollars, the Fed increases the number of dollars in the economy. Some of these new dollars are held as currency, and some are deposited in banks. Each new dollar held as currency increases the money supply by exactly $1. Each new dollar deposited in a bank increases the money supply by more than a dollar because it increases reserves and thereby increases the amount of money that the banking system can create.

To reduce the money supply, the Fed does just the opposite: It sells government bonds in the nation's bond markets. The public pays for these bonds with its holdings of currency and bank deposits, directly reducing the amount of money in circulation. In addition, as people make withdrawals from banks to buy these bonds from the Fed, banks find themselves with a smaller quantity of reserves. In response, banks reduce the amount of lending, and the process of money creation reverses itself.

Open-market operations are easy to conduct. In fact, the Fed's purchases and sales of government bonds in the nation's bond markets are like the transactions that people might undertake for their own investment portfolios. There is, however, an important difference: When two individuals engage in a trade with each other, money changes hands, but the amount of money in circulation remains the same. By contrast, when the Fed is a party in the transaction, the money supply changes. Open-market operations are the tool of monetary policy that the Fed has historically used most often.

Fed Lending to Banks The Fed can also increase the quantity of reserves by lending reserves to banks. Banks borrow from the Fed when they do not have enough reserves on hand either to satisfy bank regulators, meet depositor withdrawals, make new loans, or for some other business reason.

There are several ways for banks to borrow from the Fed. Traditionally, they borrow from the Fed's **discount window** and pay an interest rate on that loan called the **discount rate**. When the Fed makes such a loan to a bank, the banking system has more reserves than it otherwise would, and these additional reserves allow the banking system to create more money.

discount rate
the interest rate on the loans that the Fed makes to banks

The Fed can alter the money supply by changing the discount rate. A higher discount rate discourages banks from borrowing from the Fed, decreasing the quantity of reserves in the banking system and, in turn, the money supply. Conversely, a lower discount rate encourages banks to borrow from the Fed, increasing the quantity of reserves and the money supply.

At times, the Fed has set up other mechanisms for banks to borrow from it. For example, from 2007 to 2010, under the **Term Auction Facility**, the Fed set a quantity of funds it wanted to lend to banks, and eligible banks then bid to borrow those funds. The loans went to the highest eligible bidders—that is, to the banks that had acceptable collateral and offered to pay the highest interest rate. Unlike at the discount window, where the Fed sets the price of a loan and the banks determine the quantity of borrowing, at the Term Auction Facility, the Fed set the quantity of borrowing, and competitive bidding among banks determined the price. The more funds the Fed made available, the greater the quantity of reserves and the larger the money supply.

The Fed lends to banks not only to control the money supply but also to help financial institutions when they are in trouble. For example, when the stock market crashed by 22 percent on October 19, 1987, many Wall Street brokerage firms found themselves temporarily in need of funds to finance the high volume of stock trading. The next morning, before the stock market opened, Fed Chair Alan Greenspan announced the Fed's "readiness to serve as a source of liquidity to support the economic and financial system." Many economists believe that Greenspan's reaction to the stock crash was an important reason it had few repercussions.

Similarly, in 2008 and 2009, a fall in housing prices throughout the United States led to a sharp rise in the number of homeowners defaulting on their mortgage loans, and many financial institutions holding those mortgages ran into trouble. In an attempt to prevent these events from having broader economic ramifications, the Fed provided many billions of dollars in loans to financial institutions in distress.

Once again, early in 2020, when the stock and bond markets tumbled in response to the coronavirus pandemic, the Fed acted as the lender of last resort and embarked on new programs to add liquidity to the economy. Fed Chair Jerome Powell pledged to support banks and the economy "until we are confident that we are solidly on the road to recovery."

30-4b How the Fed Influences the Reserve Ratio

In addition to influencing the quantity of reserves, the Fed changes the money supply by influencing the reserve ratio and the money multiplier. The Fed can influence the reserve ratio either through regulating the quantity of reserves banks must hold or through the interest rate that the Fed pays banks on their reserves.

Reserve Requirements One way the Fed can influence the reserve ratio is by altering **reserve requirements**, the regulations that set the minimum amount of reserves that banks must hold against their deposits. An increase in reserve requirements means that banks must hold more reserves and can lend out less of each dollar of deposits, reducing how much money the banking system can create. In other words, an increase in reserve requirements raises the reserve ratio, lowers the money multiplier, and reduces the money supply. Conversely, a decrease in reserve requirements lowers the reserve ratio, raises the money multiplier, and increases the money supply.

reserve requirements regulations on the minimum amount of reserves that banks must hold against deposits

This policy tool became less relevant after 2008 because banks began to hold substantial reserves above the required level. In March 2020, the Fed set required reserves to zero, in effect eliminating reserve requirements. It is necessary to be aware of reserve requirements, nonetheless. They have been historically important and could be again in the future.

interest on reserves the interest rate paid to banks on the reserves held in deposit at the Fed

Interest on Reserves Traditionally, bank reserves did not earn interest. In October 2008, however, this changed, and the Fed began paying **interest on reserves**. That is, when banks deposit their reserves at the Fed, the Fed pays them interest on these deposits, much like a bank might pay you interest on your savings account. This change in policy explains why reserve requirements are no longer necessary: Once reserves became an interest-bearing asset, banks were more willing to hold them without being required to do so.

As a result, the Fed has another tool with which to influence the economy. The higher the interest rate on reserves, the more reserves banks choose to hold. An increase in the interest rate on reserves tends to increase the reserve ratio, reduce the money multiplier, and reduce the money supply. Conversely, a decrease in the interest rate on reserves tends to reduce the reserve ratio, increase the money multiplier, and increase the money supply.

Since interest on reserves was introduced in 2008, it has become one of the most important tools of monetary policy. As we will see, recent Fed policy has emphasized a short-term target for the federal funds rate, an interest rate on overnight loans between banks. The interest rate the Fed pays on reserves is an especially useful tool for achieving its target for the federal funds rate.

30-4c Problems in Controlling the Money Supply

The Fed's various tools—open-market operations, bank lending, reserve requirements, and interest on reserves—have powerful effects on the money supply. Yet the Fed's control of the money supply is not precise. The Fed has at times wrestled with two problems, each of which arises because much of the money supply is created through fractional-reserve banking.

The first problem is that the Fed does not control the amount of money that households choose to hold as deposits in banks. The more money households deposit, the more reserves banks have, and the more money the banking system can create. The less money households deposit, the less reserves banks have, and the less money the banking system can create. To see why this is a problem, suppose that one day, people lose confidence in the banking system and withdraw some of their deposits to hold more currency. When this happens, the banking system loses reserves and creates less money. The money supply falls, even without any Fed action.

The second problem of monetary control is that the Fed does not control the amount that bankers choose to lend. When money is deposited in a bank, it creates more money only when the bank loans it out. Because banks can choose to hold excess reserves instead, the Fed cannot be sure how much money the banking system will create. For instance, suppose that one day, bankers become more cautious about economic conditions and decide to make fewer loans and hold greater reserves. In this case, the banking system creates less money than it otherwise would. Because of the bankers' decision, the money supply falls.

That's why, in a system of fractional-reserve banking, the amount of money in the economy depends partly on the behavior of depositors and bankers. Because the Fed cannot control or perfectly predict this behavior, it cannot precisely control the money supply. Yet if the Fed is vigilant, these problems need not be large. The Fed collects data on deposits and reserves from banks every week, so it quickly becomes aware of changes in depositor or banker behavior and can respond accordingly.

Moreover, as we will discuss shortly, the Fed has recently conducted policy by setting a target for interest rates rather than a target for the money supply. This approach automatically alters the quantity of reserves to accommodate the changing attitudes of bankers and bank depositors.

Case Study

Bank Runs and the Money Supply

You may never have witnessed a bank run in real life, but perhaps you have seen one depicted in movies such as *Mary Poppins* or *It's a Wonderful Life*. A bank run is a form of mass panic, which occurs when depositors fear that a bank may be having financial troubles and "run" to the bank to withdraw their deposits. The United States has not seen a major bank run since the Great Depression of the 1930s, but in the United Kingdom, a bank called Northern Rock experienced a run in 2007 and was eventually taken over by the government.

Bank runs are a problem for banks under fractional-reserve banking. Because a bank holds only a fraction of its deposits in reserve, it cannot satisfy withdrawal requests from all depositors if they occur simultaneously. Even if the bank is **solvent** (meaning that its assets exceed its liabilities), it will not have enough cash on hand to allow all depositors immediate access to all their money. When a run occurs, the bank is forced to close its doors until some bank loans are repaid or until some lender of last resort (such as the Fed) provides it with the currency it needs to satisfy depositors.

Bank runs complicate the control of the money supply. Consider what happened in the early 1930s. After a wave of bank runs and bank closings, households and bankers became more cautious. Households withdrew their deposits from banks, preferring to hold their money as currency. This decision reversed the process of money creation, as bankers responded to falling reserves by reducing bank loans. At the same time, bankers increased their reserve ratios so they would have enough cash on hand to meet their depositors' demands in future bank runs. The higher reserve ratio reduced the money multiplier and thereby further reduced the money supply. From 1929 to 1933, the money supply fell by 28 percent without the Fed taking any deliberate contractionary action. Many economists point to this massive fall in the money supply to explain the high unemployment and falling prices that prevailed during this period. (Future chapters examine the mechanisms by which changes in the money supply affect employment and prices.)

Today, bank runs are not a major problem for the U.S. banking system, partly because of reforms enacted after the crisis of the 1930s. The federal government now guarantees the safety of deposits at most banks, primarily through the Federal Deposit Insurance Corporation (FDIC). Depositors do not make runs on their banks because they are confident that, even if their bank goes bankrupt, the FDIC will make good on the deposits.

Yet panicky investors sometimes flee other financial institutions that lack government deposit insurance, leading to events that resemble bank runs. For example, in 2008 and again in 2020, people started to doubt the soundness of money market mutual funds, which are technically not banks but have some bank-like characteristics. The Fed was concerned that massive withdrawals from these funds would disrupt their lending, roil financial markets, and depress economic activity. In both cases, the Fed stepped in to add liquidity and restore stability.

The policy of government deposit insurance has been successful in maintaining the public's confidence in the banking system, but it has downsides as well. Bankers whose deposits are guaranteed may have too little incentive to avoid bad credit risks when lending. They

MOVIESTORE COLLECTION LTD/ ALAMY STOCK PHOTO

A not-so-wonderful bank run

may think, "Heads the bank wins, tails the taxpayers lose." That's why regulators keep a close eye on the riskiness of banks' assets.

As for bank runs, most people now see them only in the movies. ●

30-4d The Federal Funds Rate

If you read about U.S. monetary policy in the news, you will find much discussion of the federal funds rate. This raises several questions:

federal funds rate the interest rate at which banks make overnight loans to one another

Q: What is the federal funds rate?

A: The **federal funds rate** is the short-term interest rate that banks charge one another for loans. If a bank finds itself short of reserves, it can borrow reserves from another bank. The loans are temporary—typically overnight. The price of the loan is the federal funds rate.

Q: How does the federal funds rate differ from the interest rate paid on reserves?

In the News — A Trip to Jekyll Island

Here's the story of how the Federal Reserve came into being.

The stranger-than-fiction story of how the Fed was created

By Roger Lowenstein

According to opinion surveys, no institution save the Internal Revenue Service is held in lower regard than the Federal Reserve. It's also a font of conspiracy theories stoked by radical libertarians, who insist the Fed is debauching the currency and will ultimately bankrupt the country.

The Fed's unpopularity would make sense if it had, say, failed to intervene and save the system during the 2008 financial crisis. But, in fact, the Fed did rescue the economy….

Nonetheless, dissatisfaction is alive in Congress, where various bills would strip the Fed's autonomy and subject sensitive monetary decisions to the scrutiny of elected politicians. Some bills would go even further and explore a return to the gold standard.

For central bank watchers, this dynamic—effective policy rewarded with populist scorn—is nothing new. In America, it has always been thus.

At Alexander Hamilton's urging, Congress first chartered a national bank—the ur-Fed—in 1791. However, Thomas Jefferson, who famously mistrusted banks (he thought agriculture more virtuous), and who was fearful of a strong central government, opposed this development. After 20 years, the Jeffersonians won and Congress let the charter expire.

This decision led to disaster: ruinous inflation. So Congress chartered a Second Bank of the United States, which began in 1817, providing the growing country with a better, more uniform currency and improved its public finances. But success couldn't save it. Andrew Jackson despised the Second Bank as a tool of East Coast elites, and it too was abolished.

For most of the 19th century, the U.S., unlike most nations in Europe, did not have a lender of last resort. Frequent panics and credit shortages were the result. Yet some of the very people who could have benefited most from a central bank, such as farmers who were starved for credit, preferred the status quo. Like Jackson and Jefferson before them, they were fearful that a government bank would tyrannize the people, perhaps in cahoots with Wall Street.

After a financial panic in 1907 virtually shut down the banking system, reformers began to press once more for a central bank. But popular mistrust remained so pronounced that they were afraid to go public.

This is the point—105 years ago—when the story seems to have been hijacked by a future Hollywood scriptwriter.

On a November evening in 1910, a powerful senator, Rhode Island Republican Nelson W. Aldrich, boarded his private rail car near New York. A light snow was falling, muting the hushed, conspiratorial tones of his guests, which is exactly how Aldrich wanted it.

The reform-minded banker Paul Warburg, one of his guests, was toting a hunting rifle, but he had no interest in hunting. The party also included a member of the powerful Morgan bank, as well as an assistant U.S. Treasury

A: The interest rate on reserves is the interest rate the Fed pays banks for holding reserves on deposit at the Fed. Lending reserves to another bank in the federal funds market is an alternative to holding reserves at the Fed, and a bank will typically do whichever pays more. As a result, these two interest rates are usually close to each other.

Q: Does the federal funds rate matter only for banks?

A: Not at all. Only banks borrow directly in the federal funds market, but the economic impact of this market is much broader. Because different parts of the financial system are highly interconnected, interest rates on different kinds of loans are strongly correlated with one another. So when the federal funds rate rises or falls, other interest rates typically move in the same direction.

Q: What does the Fed have to do with the federal funds rate?

A: In recent years, the Fed has set a target for the federal funds rate. When the Federal Open Market Committee meets about every six weeks, it decides whether to raise or lower that target.

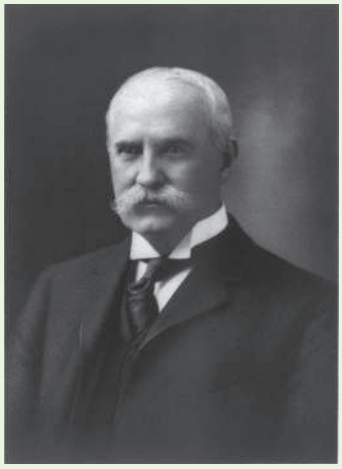
secretary, and Frank Vanderlip, head of the country's largest bank, National City.

"On what sort of errand are we going?" Vanderlip inquired.

"It may be a wild-goose chase; it may the biggest thing you and I ever did," Warburg replied.

Masquerading as duck hunters, they disembarked in Brunswick, Ga., and traveled by launch to Jekyll Island, home of an exclusive club surrounded by pine and palmetto groves. Over the course of a week, Aldrich and his bankers mapped out a draft of what was to become the Federal Reserve Act, changing the U.S. economy forever.

Congress was never told that Aldrich's bill had been drafted by Wall Street moguls. His bill did not pass, but it was the basis of a successor bill, the Federal Reserve Act, which Woodrow Wilson signed in 1913. Years later, when the Jekyll trip was revealed to the public, extremists seized on this stranger-than-fiction episode to bolster their claim that the Fed was a bankers' plot against the American people. For conspiracy theorists, the bankers' conclave on Jekyll became a metaphor for the Fed itself. The obvious irony is that, fearing Americans' irrational suspicion of central banking, Aldrich and his crew resorted to a plot that, ultimately, deepened the country's paranoia.



LIBRARY OF CONGRESS, PRINTS & PHOTOGRAPHS DIVISION, REPRODUCTION NUMBER LC-USZ62-30740 (B&W FILM COPY NEG.)

Senator Nelson Aldrich

Despite their clandestine tactics, the financiers' motives were actually patriotic. Aldrich had visited Europe and studied its central banks. He wanted expert help to draft an American equivalent. And in between sumptuous meals featuring wild turkey and freshly scalloped oysters, his group of wealthy bankers earnestly wrestled with issues that still provoke us today: How should power over the economy be apportioned between Washington and localities? How should the central bank set interest rates and the money supply?

The Federal Reserve today is not perfect. But it is more transparent than ever, thanks to reforms instituted by the previous chairman, Ben S. Bernanke, and it is no less necessary than was a central bank in 1791. Americans' paranoia is unjustified, just as it has always been. ■

Questions to Discuss

1. Why do you think Senator Aldrich wanted to keep the meeting on Jekyll Island a secret? In your view, was this secrecy justified?
2. Most people do not understand what the Federal Reserve does. How do you suppose this lack of understanding affects the job of central bankers?

Mr. Lowenstein is the author of America's Bank: The Epic Struggle to Create the Federal Reserve.

Source: *Los Angeles Times*, November 2, 2015.

Q: How can the Fed make the federal funds rate hit the target it sets?
A: Although the actual federal funds rate is determined by supply and demand in the market for loans among banks, the Fed can use the tools of monetary policy to influence that market. Most directly, if the Fed raises the interest rate it pays on reserves, banks will charge other banks more in the federal funds markets. And if the Fed reduces the interest rate it pays on reserves, banks will be more eager to lend in the federal funds market, driving down the federal funds rate.
Q: But don't these Fed actions affect the money supply?
A: Yes, absolutely. When the Fed announces a target for the federal funds rate, it is committing itself to using its tools to hit that target, and these policy actions affect the supply of money. Changes in the federal funds target and changes in the money supply are two sides of the same coin. Other things being equal, a decrease in the target for the federal funds rate means an expansion in the money supply, and an increase in the target for the federal funds rate means a contraction in the money supply.

QuickQuiz

8. Which of the following actions by the Fed would tend to increase the money supply?
   a. an open-market sale of government bonds
   b. a decrease in reserve requirements
   c. an increase in the interest rate paid on reserves
   d. an increase in the discount rate on Fed lending
9. If the Fed raises the interest rate it pays on reserves, it will _____ the money supply by increasing _____.
   a. decrease; the money multiplier
   b. decrease; excess reserves
   c. increase; the money multiplier
   d. increase; excess reserves
10. In a system of fractional-reserve banking, even without any action by the central bank, the money supply declines if households choose to hold _____ currency or if banks choose to hold _____ reserves.
   a. more; more
   b. more; less
   c. less; more
   d. less; less

Answers are at the end of the chapter.

30-5 Conclusion

Some years ago, a book made the best-seller list with the title *Secrets of the Temple: How the Federal Reserve Runs the Country*. Though no doubt an exaggeration, this title highlighted the important role of the monetary system in our daily lives. Whenever we buy or sell anything, we are relying on the extraordinarily useful social convention called "money." Now that we know what money is and what determines its supply, we can examine how changes in the quantity of money affect the economy. The next chapter begins to address that topic.

Chapter in a Nutshell

- The term **money** refers to assets that people regularly use to buy goods and services.
- Money serves three functions. As a medium of exchange, it is the item used to make transactions. As a unit of account, it provides a measure with which to record prices and other economic values. As a store of value, it offers a way to transfer purchasing power from the present to the future.
- Commodity money, such as gold, is money that has intrinsic value: It would be valued even if it were not used as money. Fiat money, such as paper dollars, is money without intrinsic value: It would be worthless if it were not used as money.
- In the U.S. economy, money takes the form of currency and various types of bank deposits, such as checking accounts.
- The Federal Reserve, the central bank of the United States, is responsible for overseeing the U.S. monetary system. The Fed chair is appointed by the president and confirmed by the Senate. The chair is the head of the Federal Open Market Committee, which sets monetary policy.
- When people deposit money in banks and banks use a fraction of these deposits to make loans to the public, the quantity of money in the economy increases. Because the banking system affects the money supply in this way, the Fed's control of the money supply is imperfect.
- Bank owners provide the resources necessary to start a bank, called bank capital. Because of leverage (the use of borrowed funds for investment), a small change in the value of a bank's assets can lead to a large change in the value of its capital. To protect depositors, regulators require banks to hold a certain minimum amount of capital.
- The Fed has several tools it can use to influence the money supply. It can expand the money supply by buying bonds in open-market operations, reducing the discount rate, increasing its lending to banks, lowering reserve requirements, or decreasing the interest rate on reserves. It can contract the money supply by selling bonds in open-market operations, increasing the discount rate, reducing its lending to banks, raising reserve requirements, or increasing the interest rate on reserves. Historically, open-market operations were the Fed's primary tool, but since 2008, it has relied more on the interest rate it pays on reserves.
- In recent years, the Fed has conducted monetary policy by setting a target for the federal funds rate, a short-term interest rate at which banks lend to one another. As the Fed pursues its target, it adjusts the money supply.

Key Concepts

money, p. 616
medium of exchange, p. 616
unit of account, p. 617
store of value, p. 617
liquidity, p. 617
commodity money, p. 617
fiat money, p. 617
currency, p. 619
demand deposits, p. 619
Federal Reserve (Fed), p. 620
central bank, p. 620
money supply, p. 621
monetary policy, p. 621
reserves, p. 622
fractional-reserve banking, p. 623
reserve ratio, p. 623
money multiplier, p. 625
bank capital, p. 625
leverage, p. 626
leverage ratio, p. 626
capital requirement, p. 627
open-market operations, p. 628
discount rate, p. 628
reserve requirements, p. 629
interest on reserves, p. 630
federal funds rate, p. 632

Questions for Review

1. What distinguishes money from other assets in the economy?
2. What is commodity money? What is fiat money? Which kind do we use?
3. What are demand deposits, and why should they be included in the stock of money?
4. Who is responsible for setting monetary policy in the United States? How is this group chosen?

5. If the Fed wants to increase the money supply with open-market operations, what does it do?
6. Why don't banks hold 100-percent reserves? How is the amount of reserves banks hold related to the amount of money that exists in the economy?
7. Bank A has a leverage ratio of 10, while Bank B has a leverage ratio of 20. Similar losses on bank loans at the two banks cause the value of their assets to fall by 7 percent. Which bank shows a larger change in bank capital? Does either bank remain solvent? Explain.
8. What is the discount rate? What happens to the money supply when the Fed raises the discount rate?
9. What are reserve requirements? What happens to the money supply when the Fed raises reserve requirements?
10. Why can't the Fed control the money supply perfectly?

Problems and Applications

1. Which of the following are considered money in the U.S. economy? Which are not? Explain your answers by discussing each of the three functions of money.
   a. a U.S. penny
   b. a Mexican peso
   c. a Picasso painting
   d. a plastic credit card
2. Explain whether each of the following events increases or decreases the money supply.
   a. The Fed buys bonds in open-market operations.
   b. The Fed reduces the reserve requirement.
   c. The Fed increases the interest rate it pays on reserves.
   d. Citibank repays a loan it had previously taken from the Fed.
   e. After a rash of pickpocketing, people decide to hold less currency.
   f. Fearful of bank runs, bankers decide to hold more reserves.
   g. The FOMC increases its target for the federal funds rate.
3. Your uncle repays a $100 loan from Tenth National Bank (TNB) by writing a $100 check from his TNB checking account. Use T-accounts to show the effect of this transaction on your uncle and on TNB. Has your uncle's wealth changed? Explain.
4. Beleaguered State Bank (BSB) holds $250 million in deposits and maintains a reserve ratio of 10 percent.
   a. Show a T-account for BSB.
   b. Now suppose that Daddy Warbucks, BSB's largest depositor, withdraws $10 million in cash from his account and that BSB decides to restore its reserve ratio by reducing the amount of loans outstanding. Show its new T-account.
   c. Explain what effect BSB's action will have on other banks.
   d. Why might it be difficult for BSB to take the action described in part (b)? Discuss another way for BSB to return to its original reserve ratio.
5. You take $100 you had kept under your mattress and deposit it in your bank account. If this $100 stays in the banking system as reserves and if banks hold reserves equal to 10 percent of deposits, by how much does the total amount of deposits in the banking system increase? By how much does the money supply increase?
6. Happy Bank starts with $200 in bank capital. It then accepts $800 in deposits. It keeps 12.5 percent (1/8th) of deposits in reserve. It uses the rest of its assets to make bank loans.
   a. Show the balance sheet of Happy Bank.
   b. What is Happy Bank's leverage ratio?
   c. Suppose some borrowers from Happy Bank default so the value of the bank loans declines by 10 percent. Show the bank's new balance sheet.
   d. By what percentage do the bank's total assets decline? By what percentage does the bank's capital decline? Which change is larger? Why?
7. In an open-market operation, the Fed buys $10 million of government bonds from individual investors. If the required reserve ratio is 10 percent, what are the largest and smallest possible increases in the money supply that could result? Explain.
8. Assume that the reserve requirement is 5 percent. All other things being equal, will the money supply expand more if the Fed buys $2,000 worth of bonds or if someone deposits in a bank $2,000 that had been hidden in a cookie jar? If one of these actions creates more money than the other, how much more does it create? Support your thinking.

9. Suppose that the reserve requirement for checking deposits is 10 percent and that banks do not hold any excess reserves.
   a. If the Fed sells $1 million of government bonds, what is the effect on the economy's reserves and money supply?
   b. Now suppose that the Fed lowers the reserve requirement to 5 percent but that banks choose to hold another 5 percent of deposits as excess reserves. Why might banks do so? What is the overall change in the money multiplier and the money supply as a result of these actions?
10. Assume that the banking system has total reserves of $100 billion. Assume also that required reserves are 10 percent of checking deposits, banks hold no excess reserves, and households hold no currency.
    a. What is the money multiplier? What is the money supply?
    b. If the Fed now raises required reserves to 20 percent of deposits, what are the change in reserves and the change in the money supply?
11. Assume that the reserve requirement is 20 percent. Also assume that banks do not hold excess reserves and that the public does not hold any cash. The Fed decides that it wants to expand the money supply by $40 million.
    a. If the Fed is using open-market operations, will it buy or sell bonds?
    b. What quantity of bonds does the Fed need to buy or sell to accomplish the goal? Explain your reasoning.
12. The economy of Elmendyn contains 2,000 $1 bills.
    a. If people hold all money as currency, what is the quantity of money?
    b. If people hold all money as demand deposits and banks maintain 100 percent reserves, what is the quantity of money?
    c. If people hold equal amounts of currency and demand deposits and banks maintain 100 percent reserves, what is the quantity of money?
    d. If people hold all money as demand deposits and banks maintain a reserve ratio of 10 percent, what is the quantity of money?
    e. If people hold equal amounts of currency and demand deposits and banks maintain a reserve ratio of 10 percent, what is the quantity of money?

QuickQuiz Answers

1. **b** 2. **c** 3. **a** 4. **c** 5. **c** 6. **d** 7. **a** 8. **b** 9. **b** 10. **a**

Chapter

31

Money Growth and Inflation

Today, if you want to buy an ice-cream cone, you need at least a couple of dollars. But that has not always been the case. In the 1930s, my grandmother ran a sweet shop in Trenton, New Jersey, where she sold ice-cream cones in two sizes. A cone with a small scoop of ice cream cost 3 cents. Hungry customers could buy a large scoop for a nickel.

This increase in the price of ice cream, and of most things sold in the 1930s, is typical of modern economies, where prices tend to rise over time. Such an increase in the overall level of prices is called **inflation**. An earlier chapter discussed how economists measure the inflation rate as the percentage change in the consumer price index (CPI), the GDP deflator, or some other index of the overall price level. The CPI shows that in the United States from 1935 to 2021, prices rose, on average, 3.5 percent per year. Accumulated over so many years, a 3.5 percent annual inflation rate amounts to a nearly twenty-fold increase in the price level.

Inflation may seem natural to a person living in the United States during recent decades, but in fact, it is not inevitable. There were long periods in the 19th century during which most prices

ROMAN SAMBORSKYI/SHUTTERSTOCK.COM; ARNO VAN ENGELAND/EYEEM/GETTY IMAGES

fell—a phenomenon called **deflation**. The average level of prices in the U.S. economy was 23 percent lower in 1896 than in 1880, and this deflation was a major issue in the presidential election of 1896. Farmers, who had accumulated large debts, suffered when declines in crop prices reduced their incomes and their ability to pay off their debts. They advocated government policies to reverse the deflation.

Although inflation has been the norm in more recent U.S. history, there has been substantial variation in the rate at which prices rise. From 1970 to 1980, prices rose by 7.8 percent per year, which meant that the price level more than doubled over the decade. By contrast, from 2010 to 2020, inflation averaged only 1.7 percent per year. But in early 2022, as the nation struggled to recover from the coronavirus pandemic, the inflation rate rose above 7 percent to the highest rate in four decades, and observers wondered whether this surge would be transitory or more persistent.

International data show a broad range of inflation experiences. In 2020, while the inflation rate was 1.2 percent in the United States, it was zero in Japan, 3.4 percent in Mexico, 11 percent in Nigeria, and 12 percent in Turkey. And even the high inflation rates in Nigeria and Turkey were moderate by some standards. In 2018, inflation in Venezuela reached about 1 million percent per year, equivalent to an increase in prices of about 2.5 percent **per day**. Such an extraordinarily high rate of inflation is called **hyperinflation**.

What determines whether an economy experiences inflation and, if so, how much? Many forces can affect the price level in the short run, as later chapters will discuss. But to explain substantial or persistent inflation, economists most often turn to the **quantity theory of money**, the main topic of this chapter. One of the **Ten Principles of Economics** in Chapter 1 summarizes the basic insight: Prices rise when the government prints too much money. The quantity theory can explain moderate inflations, such as those experienced in the United States, as well as hyperinflations.

This chapter also addresses a related question: Why is inflation a problem? At first glance, the answer may seem obvious: It's a problem because people don't like it. In the 1970s, when the United States experienced relatively high inflation, opinion polls placed inflation as the most important issue facing the nation. President Ford echoed this sentiment in 1974 when he called inflation "public enemy number one" and wore a "WIN" button on his lapel—for Whip Inflation Now. And in 2021, when inflation surged in the late stages of the Covid-19 pandemic, President Biden said it was "one of the most pressing economic concerns of the American people."

But what, exactly, are the costs that inflation imposes on a society? The answer may surprise you. Identifying the costs of inflation is not as straightforward as it first appears. All economists decry hyperinflation, but some argue that the costs of moderate inflation are not nearly as large as much of the public believes.

31-1 The Classical Theory of Inflation

The quantity theory of money is **classical** in the sense that it was developed by some of the earliest economic thinkers. According to some historians, the theory originated in the 16th century with the Renaissance polymath Nicolaus Copernicus, who is most famous for his sun-centered model of the solar system. Proponents of the quantity theory include many great economists: David Hume in the 18th century, John Stuart Mill in the 19th century, and Irving Fisher and Milton Friedman in the 20th century. Today, most economists rely on the quantity theory to explain the long-run determinants of the price level and inflation rate.

31-1a The Level of Prices and the Value of Money

Suppose the price of an ice-cream cone rises from a nickel to a dollar over some period. What conclusion should we draw from the willingness of people to give up so much more money in exchange for a cone? It's possible that they have come to enjoy ice cream more (perhaps because someone has come up with a miraculous new flavor). But it is more likely that people's enjoyment of ice cream has stayed roughly the same and that, over time, the money used to buy it has become less valuable. In most cases, inflation is more about the value of money than about the value of goods.

This insight points the way toward a theory of inflation. When the consumer price index and other measures of the price level rise, it is tempting to look at the individual prices that make up these price indexes. A typical news outlet might report, "The CPI rose by 3 percent last month, led by a 20 percent increase in the price of fresh fruits and vegetables and a 30 percent increase in the price of motor fuel." This approach contains some useful information about what's happening, but it misses a key point: Inflation is an economy-wide phenomenon that concerns, first and foremost, the value of the economy's medium of exchange.

The economy's overall price level can be seen in two ways. So far, we have viewed it as the price of a basket of goods and services. When the price level rises, people must pay more for what they buy. Alternatively, we can view the price level as a measure of money's value. A rise in the price level means a lower value of money because each dollar in your wallet or checking account buys fewer goods and services.

It may help to express these ideas mathematically. Suppose P is the price level as measured by the consumer price index or the GDP deflator. Then P measures the number of dollars needed to buy a basket of goods and services. Now turn this idea around: The quantity of goods and services that can be bought with \$1 equals $1/P$. In other words, if P is the price of goods and services measured in terms of money, $1/P$ is the value of money measured in terms of goods and services.

This math is simplest to grasp for an economy with one good—say, ice-cream cones. In that case, P is the price of a cone. When the price of a cone (P) is \$2, the value of a dollar ($1/P$) is half a cone. When the price (P) rises to \$3, the value of a dollar ($1/P$) falls to a third of a cone. The actual economy produces thousands of goods and services, so in practice, we use a price index rather than the price of a single good. But the logic is the same: When the price level rises, the value of money falls.

FRANK MODELL/ THE NEW YORKER COLLECTION /WWW.CARTOONBANK.COM

"So what's it going to be? The same size as last year or the same price as last year?"

31-1b Money Supply, Money Demand, and Monetary Equilibrium

What determines the value of money? The answer is supply and demand, as it is for so many other questions in economics. Just as the supply and demand for bananas determines the price of bananas, the supply and demand for money determines the value of money. But what factors influence money supply and money demand?

First, consider money supply. The preceding chapter discussed how the Federal Reserve, together with the banking system, influences the supply of money. To do so, the Fed uses a variety of tools. It can increase the money supply by buying government bonds in open-market operations or by reducing the interest rate it pays banks for the reserves they hold. It can contract the money supply by selling bonds or by increasing the interest rate on reserves. These policy actions affect the

money supply by working through the system of fractional-reserve banking. For the purposes of this chapter, however, these details about the banking system and money creation can be put aside. We simplify matters by taking the quantity of money supplied as a policy variable that the Fed controls.

Now consider money demand. Most fundamentally, the demand for money reflects how much wealth people want to hold in liquid form. Many factors influence the quantity of money demanded. The amount of currency that people hold in their wallets, for instance, depends on how much they rely on credit cards and how easily they can find an automatic teller machine. And as we will emphasize in Chapter 35, the quantity of money demanded depends on the interest rate that a person could earn by using the money to buy an interest-bearing bond rather than leaving it in his wallet or low-interest checking account.

Although many variables affect the demand for money, one is particularly important: the average level of prices in the economy. People hold money because it is the medium of exchange. Unlike other assets, such as bonds or stocks, people can use money to buy the goods and services on their shopping lists. How much money they choose to hold for this purpose depends on the prices of those goods and services. The higher the prices are, the more money the typical transaction requires, and the more money people will choose to hold in their wallets and checking accounts. That is, a higher price level (a lower value of money) increases the quantity of money demanded.

What ensures that the quantity of money the Fed supplies balances the quantity of money people demand? The answer depends on the time horizon being considered. Later in this book, we examine the short-run answer and learn that interest rates play a key role. The long-run answer, however, is much simpler. **In the long run, money supply and money demand are brought into equilibrium by the overall level of prices.** If the price level is above the equilibrium level, people will want to hold more money than the Fed has created, so the price level must fall to balance supply and demand. If the price level is below the equilibrium level, people will want to hold less money than the Fed has created, and the price level must rise to balance supply and demand. At the equilibrium price level, the quantity of money that people want to hold exactly balances the quantity of money supplied by the Fed.

Figure 1 illustrates these ideas. The horizontal axis of this graph shows the quantity of money. The left vertical axis shows the value of money $1/P$, and the right vertical axis shows the price level P. Notice that the price-level axis on the right is inverted: A low price level is shown near the top of this axis, and a high price level is shown near the bottom. This inverted axis illustrates that when the value of money is high (as shown near the top of the left axis), the price level is low (as shown near the top of the right axis).

The two curves in this figure are the supply and demand curves for money. The supply curve is vertical because the Fed has fixed the quantity of money available. The demand curve for money slopes downward, indicating that when the value of money is low (and the price level is high), people demand a larger quantity of money to buy goods and services. At the equilibrium, shown in the figure as point A, the quantity of money demanded balances the quantity of money supplied. This equilibrium of money supply and money demand determines the value of money and the price level.

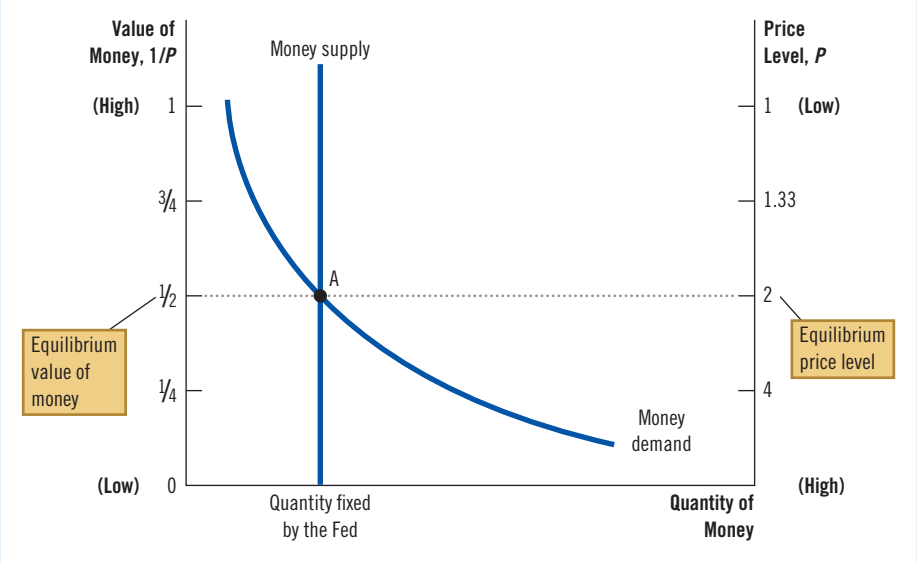
31-1c The Effects of a Monetary Injection

Consider the effects of a change in monetary policy. To do so, imagine that the economy is in equilibrium and then, suddenly, the Fed doubles the supply of money by printing dollar bills and dropping them around the country from helicopters.

Figure 1

How the Supply and Demand for Money Determine the Equilibrium Price Level

The horizontal axis shows the quantity of money. The left vertical axis shows the value of money, and the right vertical axis shows the price level. The supply curve for money is vertical because the quantity of money supplied is fixed by the Fed. The demand curve for money slopes downward because people want to hold a larger quantity of money when each dollar buys less. At the equilibrium, point A, the value of money (on the left axis) and the price level (on the right axis) have adjusted to bring the quantity of money supplied and the quantity of money demanded into balance.



(More realistically, the Fed alters the money supply using the tools discussed in the previous chapter, but the helicopter metaphor is simpler and more vivid.) What happens after such a monetary injection? How does the new equilibrium compare with the old one?

Figure 2 shows what happens. The monetary injection shifts the supply curve to the right from MS_1 to MS_2, and the equilibrium moves from point A to point B. As a result, the value of money (shown on the left axis) decreases from ½ to ¼, and the equilibrium price level (shown on the right axis) increases from 2 to 4. In other words, when an increase in the money supply makes dollars more plentiful, the result is an increase in the price level and a decrease in the value of each dollar.

This explanation of how the price level is determined and why it might change over time is called the **quantity theory of money**. According to the quantity theory, the quantity of money available in an economy determines the value of money, and growth in the quantity of money is the primary cause of inflation. As the economist Milton Friedman once put it, "Inflation is always and everywhere a monetary phenomenon."

quantity theory of money
a theory asserting that the quantity of money available determines the price level and that the growth rate in the quantity of money available determines the inflation rate

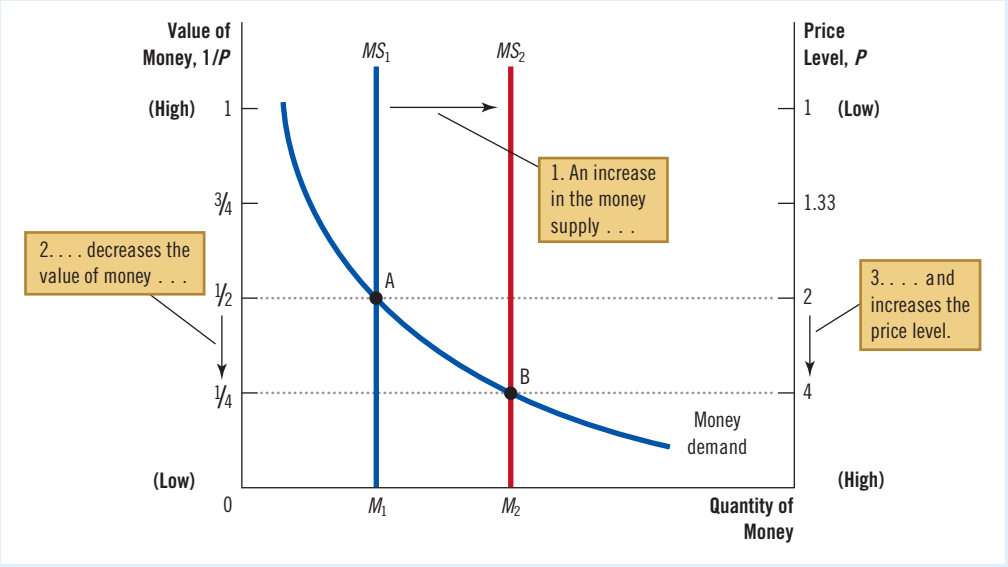
31-1d A Brief Look at the Adjustment Process

So far, we have compared the old and new equilibrium after an injection of money. How does the economy move from one to the other? A complete answer requires an understanding of short-run economic fluctuations, which we examine later in

Figure 2

An Increase in the Money Supply

When the Fed increases the supply of money, the money supply curve shifts from MS_1 to MS_2. The value of money (on the left axis) and the price level (on the right axis) adjust to bring supply and demand back into balance. The equilibrium moves from point A to point B. Thus, when an increase in the money supply makes dollars more plentiful, the price level increases, making each dollar less valuable.



this book. Here, let's briefly consider the adjustment process that occurs after a change in the money supply.

The immediate effect of a monetary injection is to create an excess supply of money. Before the injection, the economy was in equilibrium (point A in Figure 2). At the prevailing price level, people were holding exactly as much of the medium of exchange as they wanted. But after the helicopters drop the new money and people pick it up off the streets, their wallets are bulging with more dollars than they need for their planned transactions. At the prevailing price level, the quantity of money supplied now exceeds the quantity demanded.

People try to get rid of this excess supply of money in various ways. They might use it to buy more goods and services. Or they might use this excess money to make loans to others by buying bonds or by depositing the money in a bank savings account. These loans allow other people to buy more goods and services. In either case, the injection of money increases the demand for goods and services.

The economy's ability to supply goods and services, however, has not changed. As we saw in the chapter on production and growth, the economy's output of goods and services is determined by the available labor, physical capital, human capital, natural resources, and technological knowledge. The injection of money alters none of these.

That's why the greater demand for goods and services causes the prices of goods and services to rise. The increase in the price level, in turn, increases the quantity of money demanded because people use more dollars for every transaction. Eventually, the economy reaches a new equilibrium (point B in Figure 2) at which the quantity

of money demanded again equals the quantity of money supplied. In this way, the overall price level adjusts to bring money supply and money demand back into balance.

31-1e The Classical Dichotomy and Monetary Neutrality

We have seen how changes in the money supply lead to changes in the average level of prices of goods and services. How do monetary changes affect other variables, such as production, employment, wages, and interest rates? This question has long intrigued economists.

Economic variables can be divided into two groups. The first group consists of **nominal variables**—variables measured in monetary units. The second group consists of **real variables**—variables measured in physical units. For example, the income of corn farmers is a nominal variable because it is expressed in dollars, while the quantity of corn they produce is a real variable because it is expressed in bushels. Nominal GDP measures the dollar value of the economy's output of goods and services; real GDP measures the total quantity of goods and services and is not influenced by the current prices of those goods and services. The separation of real and nominal variables is called the **classical dichotomy**. A **dichotomy** is a division into two groups, and **classical** refers to the early economic thinkers who first proposed this division.

nominal variables
variables measured in monetary units

real variables
variables measured in physical units

classical dichotomy
the theoretical separation of nominal variables and real variables

Applying the classical dichotomy is tricky for prices. Most prices are quoted in units of money and, therefore, are nominal variables. When we say that the price of corn is $2 a bushel or that the price of wheat is $1 a bushel, both prices are nominal variables. But what about a **relative price**—the price of one thing in terms of another? In this example, we could say that the price of a bushel of corn is 2 bushels of wheat, a relative price that is not measured in terms of money. When comparing the prices of any two goods, the dollar signs cancel, and the resulting number is expressed in physical units. Thus, while dollar prices are nominal variables, relative prices are real variables.

This has many applications. For instance, the real wage (the dollar wage adjusted for inflation) is a real variable because it measures the rate at which people exchange a unit of labor for goods and services. Similarly, the real interest rate (the nominal interest rate adjusted for inflation) is a real variable because it measures the rate at which people exchange goods and services today for goods and services in the future. One clue, of course, is the powerful word "real." When it appears before a variable, that variable has been adjusted for inflation and is, therefore, measured in physical units.

Why separate variables into these groups? The classical dichotomy is useful because different forces influence real and nominal variables. According to classical analysis, nominal variables are influenced by developments in the economy's monetary system, but real variables are not.

This idea was implicit in our discussion of the real economy in the long run. Previous chapters examined the determinants of production, saving, investment, real interest rates, and unemployment without mentioning the effects of money. In that analysis, the economy's output of goods and services depends on technology and factor supplies, the real interest rate balances the supply and demand for loanable funds, the real wage balances the supply and demand for labor, and unemployment results when the real wage is above the equilibrium level. These conclusions have nothing to do with the quantity of money supplied.

Changes in the supply of money, according to classical theory, affect nominal variables but not real ones. When the central bank doubles the money supply, the

price level doubles, the dollar wage doubles, and all other dollar values double. Real variables, such as production, employment, real wages, and real interest rates, are unchanged. The irrelevance of monetary changes to real variables is called **monetary neutrality**.

monetary neutrality the proposition that changes in the money supply do not affect real variables

An analogy helps explain monetary neutrality. As the unit of account, money is the yardstick used to measure economic transactions. When a central bank doubles the money supply, all prices double, and the value of the unit of account falls by half. A similar change would occur if the government were to reduce the length of the yard from 36 to 18 inches: With the new, shorter yardstick, all **measured** distances (nominal variables) would double, but the **actual** distances (real variables) would remain the same. The dollar, like the yard, is merely a unit of measurement, so a change in its value should not have real effects.

Is monetary neutrality realistic? Not completely. A change in the length of the yard from 36 to 18 inches might not matter in the long run, but in the short run, it would lead to confusion and mistakes. Similarly, most economists today believe that over short periods of time—within the span of a year or two—monetary changes affect real variables. (The classical economists themselves, most notably David Hume, also doubted that monetary neutrality would apply in the short run.) We will study short-run non-neutrality later in the book, and this topic will help explain why the Fed adjusts the money supply over time.

Yet classical analysis may be fundamentally right about the economy in the long run. Over the course of a decade, monetary changes have significant effects on nominal variables (such as the price level) but appear to have negligible effects on real variables (such as real GDP). When studying long-run changes in the economy, the neutrality of money offers a reasonably good description of how the world works.

31-1f Velocity and the Quantity Equation

For another perspective on the quantity theory of money, consider the following question: How many times per year is the typical dollar bill used to pay for a newly produced good or service? The answer is given by a variable called the **velocity of money**. In physics, the term **velocity** refers to the speed at which an object travels. In economics, the velocity of money refers to the speed at which the typical dollar travels around the economy from person to person.

velocity of money the rate at which money changes hands

The velocity of money is calculated by dividing the nominal value of output (nominal GDP) by the quantity of money. If P is the price level (the GDP deflator), Y is the quantity of output (real GDP), and M is the quantity of money, then velocity is

$$V = (P \times Y)/M.$$

To see why this makes sense, imagine a simple economy that produces only pizza. Suppose that the economy produces 100 pizzas in a year, a pizza sells for $10, and the quantity of money in the economy is $50. Then the velocity of money is

$$\begin{aligned} V &= (\$10 \times 100)/\$50 \\ &= 20. \end{aligned}$$

In this economy, people spend a total of $1,000 per year on pizza. For this $1,000 of spending to take place with only $50 of money, each dollar must change hands, on average, 20 times per year.

With slight algebraic rearrangement, this equation can be rewritten as

$$M \times V = P \times Y.$$

This equation states that the quantity of money (M) times the velocity of money (V) equals the price of output (P) times the amount of output (Y). It is called the **quantity equation** because it relates the quantity of money (M) to the nominal value of output ($P \times Y$). The quantity equation shows that an increase in the quantity of money in an economy must be reflected in one of the other three variables: The price level must rise, the quantity of output must rise, or the velocity of money must fall.

quantity equation the equation $M \times V = P \times Y$, which relates the quantity of money, the velocity of money, and the dollar value of the economy's output of goods and services

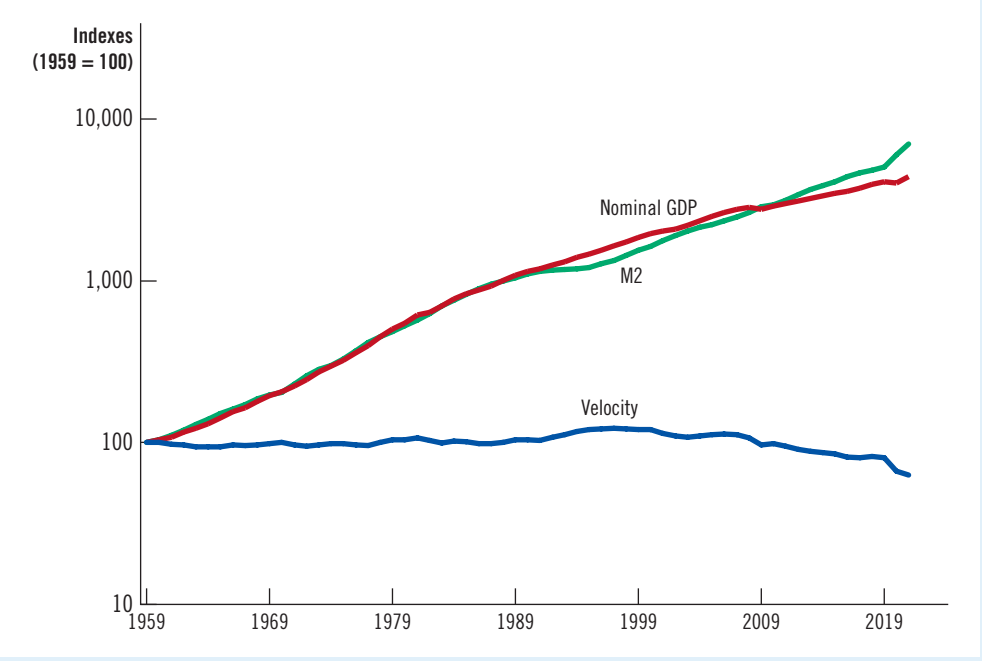
In many cases, it turns out that the velocity of money is relatively stable, at least compared with other economic variables. For example, Figure 3 shows nominal GDP, the quantity of money (as measured by M2), and the velocity of money for the U.S. economy since 1959. During this period, the money supply and nominal GDP both increased more than fortyfold. By contrast, the velocity of money, although not exactly constant, has not changed nearly as much. For some purposes, the assumption of constant velocity is a good approximation.

Figure 3

Nominal GDP, the Quantity of Money, and the Velocity of Money

This figure shows the nominal value of output as measured by nominal GDP, the quantity of money as measured by M2, and the velocity of money as measured by their ratio. For comparability, all three series have been scaled to equal 100 in 1959. Notice that nominal GDP and the quantity of money have grown substantially over this period, while velocity has been relatively stable.

Source: U.S. Department of Commerce; Federal Reserve Board.



We now have all the elements necessary to explain the equilibrium price level and inflation rate. They are as follows:

1. The velocity of money is relatively stable over time.
2. Because velocity is stable, when the central bank changes the quantity of money (M), it causes proportionate changes in the nominal value of output ($P \times Y$).
3. The economy's output of goods and services (Y) is determined by factor supplies (labor, physical capital, human capital, and natural resources) and the available production technology. In particular, since money is neutral, money does not affect output.
4. Because output (Y) is fixed by factor supplies and technology, when the central bank alters the money supply (M) and induces a proportional change in the nominal value of output ($P \times Y$), this change is reflected in a change in the price level (P).
5. Therefore, when the central bank increases the money supply rapidly, the result is a high rate of inflation.

These five points are the essence of the quantity theory of money. In short, the theory says that inflation results from "too much money chasing too few goods."

Case Study

Money and Prices during Four Hyperinflations

Earthquakes can wreak havoc on a society, but they provide much useful data for seismologists. These data can shed light on alternative theories and help society predict and deal with future threats. Similarly, hyperinflations are terrible to live through (more on this later), but they offer economists a natural experiment to study the effects of money on the economy.

Hyperinflations are interesting for a simple reason: The changes in the money supply and price level are enormous. Hyperinflation is generally defined as inflation that exceeds 50 percent **per month**, which amounts to more than a hundredfold increase in the price level over the course of a year. Returning to our ice-cream example, this rate of inflation would mean that a $2 cone you bought in the summer of 2023 would cost $260 in the summer of 2024.

Four classic hyperinflations occurred during the 1920s in Austria, Hungary, Germany, and Poland in the aftermath of World War I. In Austria, for example, the price level roughly doubled from July 1922 to August 1922 and then doubled again the next month. "One cannot even talk about prices," a visitor to Vienna said. "Before one's sentence has ended, they have increased again."

The data on these hyperinflations show a clear link between the quantity of money and the price level. Figure 4 shows the quantity of money in the economy and an index of the price level for each of these economies. The slope of the money line represents the rate at which the quantity of money was growing, and the slope of the price line represents the inflation rate. The steeper the lines, the higher the rates of money growth or inflation.

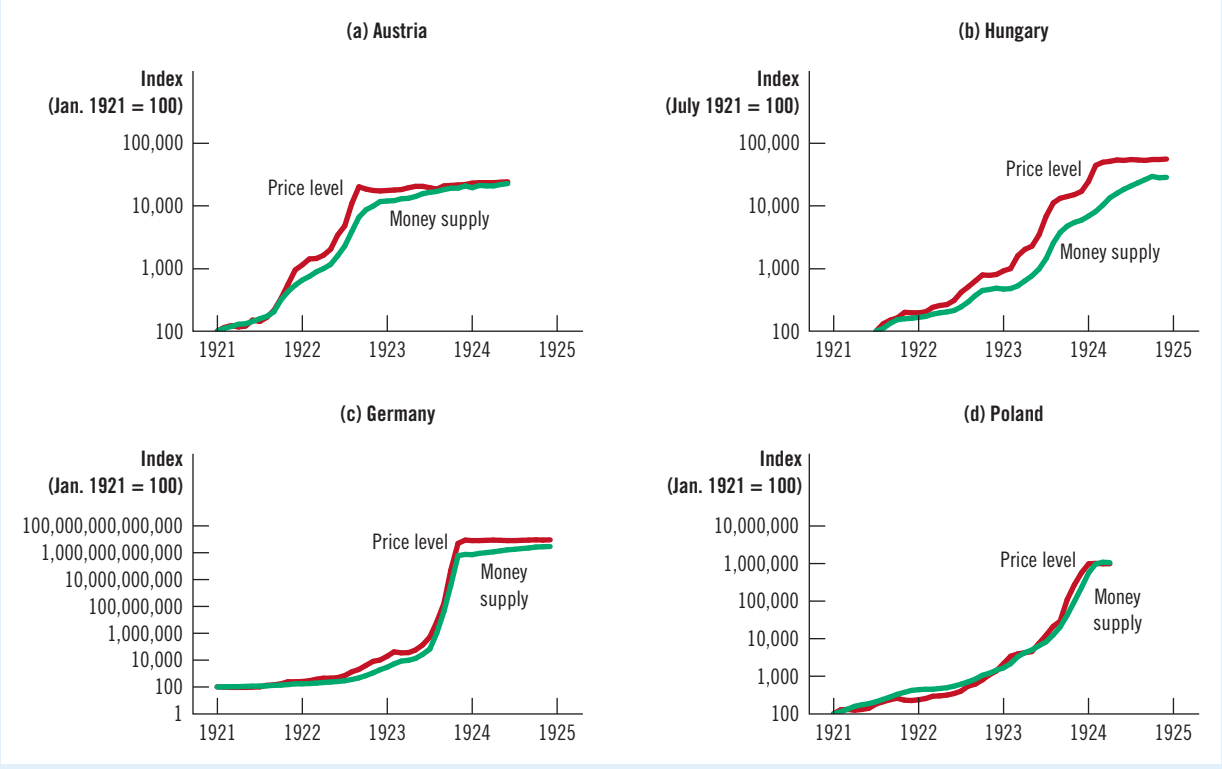
Notice that in each graph, the quantity of money and the price level are almost parallel. In each instance, growth in the quantity of money is moderate at first, and so is inflation. But over time, the quantity of money in the economy starts growing faster and faster. At about the same time, inflation also takes off. Then, when the quantity of money stabilizes, the price level stabilizes as well. These episodes illustrate one of the **Ten Principles of Economics**: Prices rise when the government prints too much money. ●

Figure 4

Money and Prices during Four Hyperinflations

This figure shows the quantity of money and the price level during four hyperinflations. (Note that because these variables are graphed on **logarithmic** scales, equal vertical distances on the graph represent equal **percentage** changes in the variable.) In each case, the quantity of money and the price level move closely together. The strong association between these two variables is consistent with the quantity theory of money, which states that growth in the money supply is the primary cause of inflation.

Source: Adapted from Thomas J. Sargent, "The End of Four Big Inflations," in Robert Hall, ed., *Inflation* (Chicago: University of Chicago Press, 1983), pp. 41–93. Each series is normalized to equal 100 for the initial observation.



31-1g The Inflation Tax

If inflation is so easy to explain, why do countries experience hyperinflation? That is, why do the central banks of these countries print so much money that its value is certain to fall rapidly?

The answer is that the governments of these countries are using money creation to pay for their spending. When governments want to build roads or schools, pay the salaries of public employees, or give transfer payments to the disadvantaged or politically powerful, they first must raise the necessary funds. Normally, they do this by levying taxes, such as income and sales taxes, and by borrowing from the public by selling government bonds. Yet in an economy with a system of fiat money, a government can also pay for spending simply by printing the money it needs.

When the government raises revenue by printing money, it is said to levy an **inflation tax**. The inflation tax is not exactly like other taxes, however, because the government doesn't send anyone a bill for this tax. Instead, the inflation tax is subtler. When the

inflation tax
the revenue the government raises by creating money

government prints money, the price level rises, and the dollars in your wallet become less valuable. **Thus, the inflation tax is like a tax on everyone who holds money.**

The importance of the inflation tax varies from country to country and over time. In the United States, in recent years, the inflation tax has been a minor source of revenue: It has accounted for less than 3 percent of federal receipts. During the 1770s, however, the Continental Congress of the fledgling United States relied heavily on the inflation tax to pay for military spending. Because the new government had a limited ability to raise funds through regular taxes or borrowing, printing dollars was the easiest way to compensate the soldiers fighting for independence. As the quantity theory predicts, the result was a high rate of inflation: Prices measured in terms of the continental dollar rose more than a hundredfold over a few years.

Almost all hyperinflations follow the same pattern as the hyperinflation during the American Revolution. The government has high spending, inadequate tax revenue, and limited capacity to borrow. To pay for its spending, it turns to the printing press. The massive increases in the quantity of money then lead to massive inflation. The hyperinflation ends when the government institutes fiscal reforms—such as cuts in government spending—that eliminate the need for the inflation tax.

31-1h The Fisher Effect

According to the principle of monetary neutrality, an increase in the rate of money growth raises the rate of inflation but does not affect any real variable. An important application concerns the effect of money on interest rates. Macroeconomists are especially concerned about interest rates because they link the economy of the present and the economy of the future through their effects on saving and investment.

To understand the relationship among money, inflation, and interest rates, recall the distinction between the nominal interest rate and the real interest rate. The **nominal interest rate** is the interest rate you hear about at your bank. If you have a savings account, for instance, the nominal interest rate tells you how fast the number of dollars in your account will rise over time. The **real interest rate** corrects the nominal interest rate for the effect of inflation to tell you how fast the purchasing power of your savings account will rise over time. The real interest rate is the nominal interest rate minus the inflation rate:

$$\text{Real interest rate} = \text{Nominal interest rate} - \text{Inflation rate}.$$

For example, if the bank posts a nominal interest rate of 7 percent per year and the inflation rate is 3 percent per year, then the real value of the deposits grows by 4 percent per year.

We can rewrite this equation to show that the nominal interest rate is the sum of the real interest rate and the inflation rate:

$$\text{Nominal interest rate} = \text{Real interest rate} + \text{Inflation rate}.$$

This way of looking at the nominal interest rate is useful because different forces determine each of the two terms on the right side of this equation. As an earlier chapter discussed, the supply and demand for loanable funds determine the real interest rate. And according to the quantity theory of money, growth in the money supply determines the inflation rate.

So how does growth in the money supply affect interest rates? In the long run, over which money is neutral, a change in money growth should not affect the real interest rate. The real interest rate is, after all, a real variable. For the real interest rate to remain unchanged, a change in the inflation rate must result in a one-for-one change

in the nominal interest rate. **Thus, when a central bank increases the rate of money growth, the long-run result is an equal increase in the inflation rate and the nominal interest rate.** This adjustment of the nominal interest rate to the inflation rate is called the **Fisher effect**, after Irving Fisher (1867–1947), the economist who first studied it.

Fisher effect
the one-for-one adjustment of the nominal interest rate to the inflation rate

Keep in mind that this analysis of the Fisher effect has maintained a long-run perspective. The Fisher effect need not hold in the short run because inflation may be unanticipated. A nominal interest rate is a payment on a loan, and it is typically set when the loan is first made. If a jump in inflation catches the borrower and lender by surprise, the nominal interest rate they agreed on will fail to reflect the higher inflation. But if inflation remains high, people will eventually come to expect it, and the nominal interest rates set in loan agreements will reflect this expectation. To be precise, therefore, the Fisher effect states that the nominal interest rate adjusts to expected inflation. Expected inflation moves with actual inflation in the long run but not necessarily in the short run.

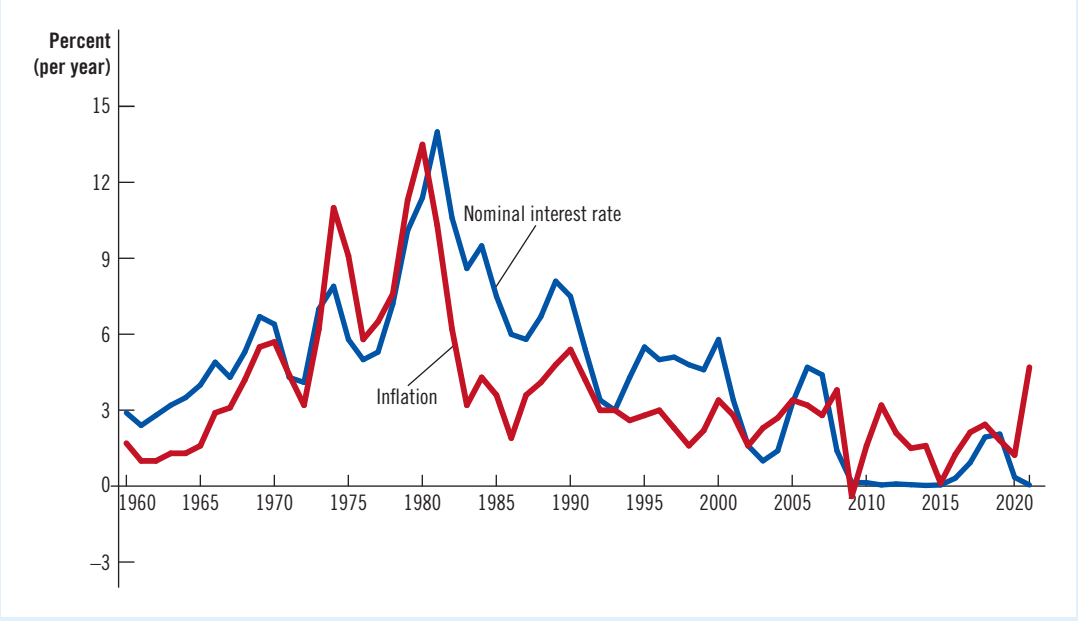
The Fisher effect is crucial for understanding changes in the nominal interest rate over time. Figure 5 shows the nominal interest rate and the inflation rate in the U.S. economy since 1960. The close association between these two variables is clear. The nominal interest rate rose from the early 1960s through the 1970s because inflation was also rising during this time. Similarly, the nominal interest rate fell from the early 1980s through the 1990s because the Fed got inflation under control. In the 2010s, both the nominal interest rate and the inflation rate were low by historical standards.

Figure 5

The Nominal Interest Rate and the Inflation Rate

This figure uses annual data since 1960 to show the nominal interest rate on three-month Treasury bills and the inflation rate as measured by the consumer price index. The close association between these two variables provides evidence for the Fisher effect: When the inflation rate rises, so does the nominal interest rate.

Source: U.S. Department of Treasury; U.S. Department of Labor.



QuickQuiz

1. The classical principle of monetary neutrality states that changes in the money supply do not influence _____ variables, and it is thought to be most applicable in the _____ run.
   a. nominal; short
   b. nominal; long
   c. real; short
   d. real; long
2. If nominal GDP is $400, real GDP is $200, and the money supply is $100, then
   a. the price level is ½, and velocity is 2.
   b. the price level is ½, and velocity is 4.
   c. the price level is 2, and velocity is 2.
   d. the price level is 2, and velocity is 4.
3. According to the quantity theory of money, which variable in the quantity equation is most stable over long periods of time?
   a. money
   b. velocity
   c. price level
   d. output
4. Hyperinflation occurs when the government runs a large budget _____, which the central bank finances with a substantial monetary _____.
   a. deficit; contraction
   b. deficit; expansion
   c. surplus; contraction
   d. surplus; expansion
5. According to the quantity theory of money and the Fisher effect, if the central bank increases the rate of money growth, then
   a. inflation and the nominal interest rate both increase.
   b. inflation and the real interest rate both increase.
   c. the nominal interest rate and the real interest rate both increase.
   d. inflation, the real interest rate, and the nominal interest rate all increase.

Answers are at the end of the chapter.

31-2 The Costs of Inflation

In the late 1970s, when the U.S. inflation rate reached about 10 percent per year, inflation dominated debates over economic policy. And even when inflation is low, it remains a closely watched macroeconomic variable. One study found that **inflation** is the economic term mentioned most often in U.S. newspapers (ahead of second-place finisher **unemployment** and third-place finisher **productivity**).

Inflation is closely watched and widely discussed because it is thought to be a serious economic problem. But is that true? And if so, why?

31-2a A Fall in Purchasing Power? The Inflation Fallacy

If you ask most people why inflation is bad, they will tell you that the answer is obvious: Inflation robs them of the purchasing power of their hard-earned dollars. When prices rise, each dollar of income buys fewer goods and services. So it might seem that inflation directly lowers living standards.

Yet further thought reveals a fallacy in this answer. When prices rise, buyers of goods and services pay more for what they buy. At the same time, however, sellers of goods and services get more for what they sell. Because most people earn their incomes by selling their services, such as their labor, inflation in incomes goes hand in hand with inflation in prices. **Inflation does not in itself reduce people's real purchasing power.**

People believe the inflation fallacy because they do not appreciate the principle of monetary neutrality. Workers who receive an annual raise of 10 percent tend to view that raise as a reward for their talent and effort. When an inflation rate of 6 percent reduces the real value of that raise to only 4 percent, the workers might feel cheated. In fact, as the chapter on production and growth discussed, real incomes are determined by real variables, such as physical capital, human capital, natural resources, and the available production technology. Nominal incomes are determined by a combination of those factors and the overall price level. If the Fed lowered the inflation rate from 6 percent to zero, the workers' annual raise would fall from 10 percent to 4 percent. They might feel less robbed by inflation, but their real income would not rise more quickly.

If nominal incomes tend to keep pace with rising prices in the long run, why then is inflation a problem? It turns out that there is no single answer. Instead, economists have identified several costs of inflation. Each of these costs shows some way in which persistent growth in the money supply does, in fact, have some adverse effect on real variables.

31-2b Shoeleather Costs

As we have seen, inflation is like a tax on the holders of money. The tax itself is not a cost to society: It is only a transfer of resources from households to the government. Yet most taxes give people an incentive to alter their behavior to avoid paying the tax, and this distortion of incentives causes deadweight losses. This is true for the inflation tax as well. It causes deadweight losses because people waste scarce resources trying to escape it.

How can a person avoid paying the inflation tax? Because inflation erodes the real value of money, you can avoid the inflation tax by holding less money. One way to do this is to go to the bank more often. For example, rather than withdrawing $200 every four weeks, you might withdraw $50 once a week. By making more frequent trips to the bank, you can keep more of your wealth in your interest-bearing savings account and less in your wallet, where inflation erodes its value.

The cost of reducing your money holdings is called the **shoeleather cost** of inflation because making more frequent trips to the bank causes your shoes to wear out more quickly. Of course, this term is not to be taken literally: The actual cost of reducing your money holdings is not the wear and tear on your shoes but the time and convenience you must sacrifice to keep less money on hand than you would if there were no inflation.

shoeleather costs the resources wasted when inflation encourages people to reduce their money holdings

The shoeleather costs of inflation have been trivial in the U.S. economy, which usually has only moderate inflation. But this cost is magnified in countries with hyperinflation. Here is a description of one person's experience in Bolivia during its hyperinflation (as reported in the August 13, 1985, issue of *The Wall Street Journal*):

> When Edgar Miranda gets his monthly teacher's pay of 25 million pesos, he hasn't a moment to lose. Every hour, pesos drop in value. So, while his wife rushes to market to lay in a month's supply of rice and noodles, he is off with the rest of the pesos to change them into black-market dollars.
>
> Mr. Miranda is practicing the First Rule of Survival amid the most out-of-control inflation in the world today. Bolivia is a case study of how runaway inflation undermines a society. Price increases are so huge that the figures build up

> almost beyond comprehension. In one six-month period, for example, prices soared at an annual rate of 38,000 percent. By official count, however, last year's inflation reached 2,000 percent, and this year's is expected to hit 8,000 percent—though other estimates range many times higher. In any event, Bolivia's rate dwarfs Israel's 370 percent and Argentina's 1,100 percent—two other cases of severe inflation.
>
> It is easier to comprehend what happens to the thirty-eight-year-old Mr. Miranda's pay if he doesn't quickly change it into dollars. The day he was paid 25 million pesos, a dollar cost 500,000 pesos. So he received $50. Just days later, with the rate at 900,000 pesos, he would have received $27.

As this story shows, the shoeleather costs of inflation can be large. With the high inflation rate, Mr. Miranda does not have the luxury of holding the local money as a store of value. Instead, he is forced to convert his pesos quickly into goods or into U.S. dollars, which offer a more stable store of value. The time and effort that Mr. Miranda expends to reduce his money holdings are wasted resources. If the monetary authority pursued a low-inflation policy, Mr. Miranda would be happy to hold pesos, and he could put his time and effort to more productive use. In fact, shortly after this article was written, the inflation rate in Bolivia fell substantially as a result of a more restrictive monetary policy.

31-2c Menu Costs

Most firms do not change the prices of their products every day. Instead, firms often announce prices and leave them unchanged for weeks, months, or even years. One survey found that the typical U.S. firm changes its prices about once a year.

menu costs
the costs of changing prices

Firms change prices infrequently because there are costs to changing prices. Costs of price adjustment are called **menu costs**, a term derived from a restaurant's cost of printing a new menu. Menu costs include the costs of deciding on new prices, printing new price lists and catalogs, sending these new price lists and catalogs to dealers and customers, advertising the new prices, and even dealing with customer annoyance over price changes.

Inflation increases the menu costs that firms must bear. In the low-inflation environment that has usually prevailed in the modern U.S. economy, annual price adjustment is an appropriate business strategy for many firms. But when high inflation makes firms' costs rise rapidly, annual price adjustment is impractical. During hyperinflations, for example, firms must change their prices daily or even more often just to keep up with all the other prices in the economy.

31-2d Relative-Price Variability and the Misallocation of Resources

Suppose that the Eatabit Eatery prints a new menu with new prices every January and then leaves its prices unchanged for the rest of the year. If there is no inflation, Eatabit's relative prices—the prices of its meals compared with other prices in the economy—would be constant over the course of the year. But if the inflation rate is 12 percent per year, Eatabit's relative prices will automatically fall by 1 percent each month. The restaurant's relative prices will be highest in the early months of the year, just after it has printed a new menu, and lowest in the later months. And the higher the inflation rate, the greater this swing in relative prices will be. Thus, because prices change only once in a while, inflation causes relative prices to vary more than they otherwise would.

Why does this matter? The reason is that market economies rely on relative prices to allocate scarce resources. Consumers decide what to buy by comparing the quality and prices of various goods and services. Through these decisions, they determine how the scarce factors of production are allocated among industries and firms. When inflation distorts relative prices, consumer decisions are distorted, and markets are less able to allocate resources to their best use.

31-2e Inflation-Induced Tax Distortions

Most taxes distort incentives, cause people to alter their behavior, and lead to a less efficient allocation of the economy's resources. In the presence of inflation, many taxes become even more problematic. The reason is that lawmakers often fail to take inflation into account when writing the tax laws. Economists who have studied the tax code conclude that inflation tends to raise the tax burden on income earned from savings.

One example of how inflation discourages saving is the tax treatment of **capital gains**—the profits made by selling an asset for more than its purchase price. Suppose that, in 1975, you used some of your savings to buy one share of IBM stock for $10 and that you sold it for $110 in 2020. According to the tax law, you have earned a capital gain of $100, which you must include in your income when figuring out how much income tax you owe. But because the overall price level increased fivefold from 1975 to 2020, the $10 you invested in 1975 is equivalent (in terms of purchasing power) to $50 in 2020. So when you sell your stock for $110, you have a real gain (an increase in purchasing power) of only $60 ($110 – $50). The tax code, however, ignores inflation and taxes you on a gain of $100. In this way, inflation exaggerates the size of capital gains and inadvertently increases the tax burden on this type of income.

Another example is the tax treatment of interest income. The income tax treats the **nominal** interest earned on savings as income, even though part of the nominal interest rate merely compensates for inflation. To see the effects, consider the numerical example in Table 1. The table compares two economies, both of which tax interest income at a rate of 25 percent. In Economy A, inflation is zero, and

Table 1

How Inflation Raises the Tax Burden on Saving

In the presence of zero inflation, a 25 percent tax on interest income reduces the real interest rate from 4 percent to 3 percent. In the presence of 8 percent inflation, the same tax reduces the real interest rate from 4 percent to 1 percent.

	Economy A (zero inflation)	Economy B (high inflation)
Real interest rate	4%	4%
Inflation rate	0	8
Nominal interest rate (real interest rate + inflation rate)	4	12
Reduced interest due to 25 percent tax (0.25 × nominal interest rate)	1	3
After-tax nominal interest rate (0.75 × nominal interest rate)	3	9
After-tax real interest rate (after-tax nominal interest rate − inflation rate)	3	1

the nominal and real interest rates are both 4 percent. In this case, the 25 percent tax on interest income reduces the real interest rate from 4 percent to 3 percent. In Economy B, the real interest rate is again 4 percent, but the inflation rate is 8 percent. As a result of the Fisher effect, the nominal interest rate is 12 percent. Because the income tax treats this entire 12 percent interest as income, the 25 percent tax leaves an after-tax nominal interest rate of only 9 percent and, after correcting for 8 percent inflation, an after-tax real interest rate of only 1 percent. In this case, the 25 percent tax on interest income reduces the real interest rate from 4 percent to 1 percent. Because the after-tax real interest rate provides the incentive to save, saving is much less attractive in the economy with inflation (Economy B) than in the economy with stable prices (Economy A).

The taxes on nominal capital gains and on nominal interest income are two examples of how the tax code interacts with inflation. There are many others. Because of these inflation-induced tax changes, higher inflation tends to discourage people from saving. Recall that the economy's saving provides the resources for investment, which, in turn, is important to long-run economic growth. When inflation raises the tax burden on saving, it tends to depress the economy's long-run growth rate. There is, however, no consensus among economists about the size of this effect.

One solution to this problem, other than eliminating inflation, is to index the tax system. That is, the tax laws could be revised to account for the effects of inflation. In the case of capital gains, for example, the tax code could adjust the purchase price using a price index and assess the tax only on the real gain. In the case of interest income, the government could tax only real interest income by excluding that portion of the interest income that merely compensates for inflation. To some extent, the tax laws have moved in the direction of indexation. For example, the income levels at which income tax rates change are adjusted automatically each year based on changes in the consumer price index. Yet many other aspects of the tax laws—such as the tax treatment of capital gains and interest income—are not indexed.

In an ideal world, the tax laws would be written so that inflation would not alter anyone's real tax liability. In the real world, however, tax laws are far from perfect. More complete indexation would probably be desirable, but it would further complicate a tax code that many people already consider onerous.

31-2f Confusion and Inconvenience

Imagine that we took a poll and asked the following question: "This year, a yard is 36 inches. How long do you think it should be next year?" If we could get people to take us seriously, they would tell us that a yard should stay the same length—36 inches. Anything else would just complicate life needlessly.

What does this have to do with inflation? Recall that money, as the economy's unit of account, is what we use to quote prices and record debts. Money is the yardstick with which we measure economic transactions. The Federal Reserve is a bit like the Bureau of Standards: Its job is to ensure the reliability of a commonly used unit of measurement. When the Fed creates inflation, it erodes the real value of the unit of account.

It is difficult to judge the costs of the confusion and inconvenience that arise from inflation. As noted earlier, the tax code incorrectly measures real incomes in the presence of inflation. Similarly, accountants have difficulty measuring firms' earnings when prices are rising over time. Because inflation causes dollars at different times

to have different real values, computing a firm's profit—the difference between its revenue and costs—is more complicated in an economy with rapid inflation. Therefore, to some extent, inflation can make it harder to sort successful from unsuccessful firms, impeding the financial system as it aims to allocate the economy's capital to its best uses.

31-2g A Special Cost of Unexpected Inflation: Arbitrary Redistributions of Wealth

The costs of inflation discussed so far occur even if the change in the price level is steady and predictable. Inflation has another cost, however, when it comes as a surprise. Unexpected inflation redistributes wealth in a way that has nothing to do with either merit or need. These redistributions occur because many loans in the economy are specified in terms of the unit of account—money.

Consider an example. Suppose that Sophie Student takes out a $50,000 loan at a 7 percent interest rate from Bigbank to attend college. In 10 years, the loan will come due. After her debt has compounded for 10 years at 7 percent, Sophie will owe Bigbank $100,000. But the real value of this debt will depend on inflation over the decade. If Sophie is lucky, the economy will have a hyperinflation. In this case, wages and prices will rise so high that Sophie will be able to pay the $100,000 debt out of pocket change. (The hyperinflation may, however, hurt Sophie in other ways.) By contrast, if the economy goes through a major deflation, then wages and prices will fall, and Sophie will find the $100,000 debt a greater burden than anticipated.

This example shows that unexpected changes in prices redistribute wealth among debtors and creditors. High inflation enriches Sophie at the expense of Bigbank because it diminishes the real value of the debt; Sophie can repay the loan in dollars that are less valuable than she expected. Deflation enriches Bigbank at Sophie's expense because it increases the real value of the debt; in this case, Sophie must repay the loan in dollars that are more valuable. If inflation were predictable, then Bigbank and Sophie could take inflation into account when setting the nominal interest rate on the loan. (Recall the Fisher effect.) But if inflation is hard to predict, it imposes risk on Sophie and Bigbank that both would prefer to avoid.

This cost of unexpected inflation is important to consider together with another tendency: Inflation is especially volatile and uncertain when the average rate of inflation is high. This is evident in comparisons among different countries. Countries with low average inflation, such as Germany in the late 20th century, tend to have stable inflation. Countries with high average inflation, such as many countries in Latin America, tend to have unstable inflation. There are no known examples of economies with high, stable inflation. This relationship between the level and volatility of inflation points to another cost of inflation. If a country pursues a high-inflation monetary policy, it will have to bear not only the costs of high expected inflation but also the arbitrary redistributions of wealth associated with unexpected inflation.

31-2h Inflation Is Bad, but Deflation May Be Worse

In recent U.S. history, inflation has been the norm. But the level of prices has fallen at times, such as during the late 19th century and early 1930s. From 1998 to 2012, Japan experienced a 4-percent decline in its overall price level. So as we conclude our discussion of the costs of inflation, the costs of deflation are worth considering as well.

Some economists have suggested that a small and predictable amount of deflation may be desirable. Milton Friedman pointed out that deflation would lower the nominal interest rate (via the Fisher effect) and that a lower nominal interest rate would reduce the cost of holding money. The shoeleather costs of holding money would, he argued, be minimized by a nominal interest rate close to zero, which in turn would require deflation equal to the real interest rate. This prescription for moderate deflation is called the **Friedman rule**.

Yet there are also costs of deflation. Some of these mirror the costs of inflation. For example, just as a rising price level induces menu costs and relative-price variability, so does a falling price level. Moreover, in practice, deflation is rarely as steady and predictable as Friedman recommended. More often, it comes as a surprise, resulting in the redistribution of wealth toward creditors and away from debtors. Because debtors tend to be poorer, these redistributions in wealth are especially painful.

Perhaps most important, deflation often arises from broader macroeconomic difficulties. As we will see in future chapters, falling prices result when some event, such as a monetary contraction, reduces the overall demand for goods and services. This fall in aggregate demand can lead to falling incomes and rising unemployment. In other words, deflation is often a symptom of deeper economic problems.

The Wizard of Oz and the Free-Silver Debate

The movie *The Wizard of Oz*, based on a children's book written in 1900, has been a staple of television and video streaming for decades. The movie and book tell the familiar story of a young girl, Dorothy, who finds herself lost in a strange land far from home. You may not know, though, that some scholars believe that the story is an allegory about U.S. monetary policy in the late 19th century, a period of severe deflation.

From 1880 to 1896, the price level in the U.S. economy fell by 23 percent. Because this event was unanticipated, it led to an acute redistribution of wealth. Most farmers in the west were debtors. Their creditors were the bankers in the east. When the price level fell, the real value of these debts rose, enriching the bankers at the expense of the farmers.

According to Populist politicians of the time, the solution to the farmers' problem was the free coinage of silver. During this period, the United States was operating with a gold standard. The quantity of gold determined the money supply and the price level. Free-silver advocates wanted silver, in addition to gold, to be used as money. This would have increased the money supply sharply, pushing up the price level and reducing the burden of farmers' debts.

The debate over silver was heated, and it was central to the politics of the 1890s. A Populist election slogan was "We Are Mortgaged. All but Our Votes." One prominent advocate of free silver was William Jennings Bryan, the Democratic nominee for president in 1896. He is remembered in part for a spellbinding speech at the Democratic Party's nominating convention in which he said, "You shall not press down upon the brow of labor this crown of thorns. You shall not crucify mankind upon a cross of gold." Rarely since then have politicians waxed so poetic about alternative approaches to monetary policy. Nonetheless, Bryan lost the election to Republican William McKinley, and the United States remained on the gold standard.

L. Frank Baum, author of the book *The Wonderful Wizard of Oz*, was a Midwestern journalist. When he sat down to write a story that was ostensibly for children, he

made the characters represent protagonists in the major political battle of his time. Here is how the economic historian Hugh Rockoff, writing in the *Journal of Political Economy* in 1990, interprets the story:

Dorothy:	Traditional American values
Toto:	Prohibitionist party, also called the Teetotalers
Scarecrow:	Farmers
Tin Woodsman:	Industrial workers
Cowardly Lion:	William Jennings Bryan
Munchkins:	Citizens of the East
Wicked Witch of the East:	Grover Cleveland
Wicked Witch of the West:	William McKinley
Wizard:	Marcus Alonzo Hanna, chairman of the Republican Party
Oz:	Abbreviation for ounce of gold
Yellow Brick Road:	Gold standard

At the end of Baum's story, Dorothy finds her way home, but it is not by just following the yellow brick road. After a long and perilous journey, she learns that the wizard is incapable of helping her. Instead, Dorothy discovers the magical power of her **silver** slippers. (When the book became a movie in 1939, Dorothy's slippers were changed from silver to ruby. The Hollywood filmmakers were more interested in showing off the new technology of Technicolor than in telling a story about 19th-century monetary policy.)

The Populists didn't achieve the free coinage of silver, but they eventually got the monetary expansion and inflation they wanted. In 1898, prospectors discovered gold near the Klondike River in the Canadian Yukon. Increased gold supplies also arrived from the mines of South Africa. As a result, the money supply and the price level started to rise in the United States and in other countries operating on the gold standard. Within 15 years, prices in the United States returned to the levels that had prevailed in the 1880s, and farmers were better able to handle their debts. ●

AF ARCHIVE/ALAMY STOCK PHOTO

An early debate over monetary policy

QuickQuiz

6. Ongoing inflation does not automatically reduce most people's incomes because
   a. the tax code is fully indexed for inflation.
   b. people respond to inflation by holding less money.
   c. wage inflation goes together with price inflation.
   d. higher inflation lowers real interest rates.
7. If an economy always has inflation of 10 percent per year, which of the following costs of inflation will it NOT suffer?
   a. shoeleather costs from reduced holdings of money
   b. menu costs from more frequent price adjustment
   c. distortions from the taxation of nominal capital gains
   d. arbitrary redistributions between debtors and creditors
8. Because most loans are written in _____ terms, an unexpected increase in inflation hurts _____.
   a. real; creditors
   b. real; debtors
   c. nominal; creditors
   d. nominal; debtors

Answers are at the end of the chapter.

In the News

Life during Hyperinflation

The costs of inflation are most apparent when inflation becomes extreme.

What 52,000 Percent Inflation Can Do to a Country

By Brook Larmer

I walked into the empty restaurant in Managua carrying a backpack stuffed with cash, thick stacks of Nicaraguan córdobas bound by rubber bands. The waiter, as expected, asked me to hand over the entire stash. It may have looked like an illicit transaction. But this was Nicaragua, in 1990, at the end of its war with the American-trained contra rebels, and I was only trying to buy a meal before my money lost its value. A decade of guerrilla war and deficit spending had whipped up a maelstrom of hyperinflation and shortages. Only two items on the menu were available, and prices had doubled in a matter of weeks. With inflation surging past 13,000 percent annually, the restaurant now demanded payment upfront—to ensure that the staff had enough time to tally it. As I ate my rice and beans, two waiters at another table counted every bill. I finished before they did, even though the meal—and all those millions of córdobas—added up to less than $10.

Hyperinflation is a mercurial phenomenon, a rupture that occurs when a government persistently spends (or prints) money that it doesn't have, and the public loses confidence in the process. The distortions that emerge—like the backpack full of soon-to-be-worthless cash—can seem absurd, even laughable. Yet there is nothing amusing about the damage that hyperinflation can inflict on the lives of people and nations. "If you can't trust the money the government issues, then you can't trust anything," says Steve Hanke, a professor of applied economics at Johns Hopkins University and a leading expert on hyperinflation (which he has defined as 50 percent monthly inflation sustained for at least 30 days). Hanke has studied the 58 cases of hyperinflation that have been recorded, from Germany's Weimar Republic to the episode I witnessed in Nicaragua, each one an earthquake that caused people to lose faith in the very foundation—the value of money—on which their lives depended.

The newest addition to the ignominious list, and a cause for alarm in Washington, is the crisis in Venezuela. Even with the world's most-abundant oil fields, Venezuela has mismanaged its way to economic disaster. Hyperinflation and its common companion, chronic shortages of food and medicine, have impoverished almost all of the country's 31 million people. Nine out of 10 Venezuelans do not earn enough money to buy sufficient food, according to a recent survey. Over all, Venezuelans have lost an average of 24 pounds each. Malaria is on the rise, as is crime. Those who can are getting out: More than 2.3 million Venezuelans have fled the country, including more than half of the nation's doctors.

The situation is still out of control. Venezuela's economy shrank by 35 percent between 2013 and 2017, and economists forecast another 18 percent drop in 2018. Oil production, crippled by the lack of maintenance and investment, fell in July to its lowest point in nearly seven decades. According to Hanke, the rate of inflation over the last 12 months was 52,000 percent. The chaos poses a risk for the entire region. "Venezuela has sparked the most serious economic, humanitarian and political crisis in the Americas in decades," says the Brazilian economist Monica de Bolle, the director of Latin American studies at the Johns Hopkins University School for Advanced International Studies. "There has never been a crisis quite like this in the region, and we've had plenty."

Nearly a century ago, Vladimir Lenin was quoted in *The New York Times* saying that hyperinflation was "the simplest way to exterminate the very spirit of capitalism." If a country were flooded with high face-value notes untethered to anything of real value, he reasoned, "men will cease to covet and hoard [money] so soon as they discover it will not buy anything, and

31-3 Conclusion

This chapter discussed the causes and costs of inflation. The primary cause of substantial or persistent inflation is growth in the quantity of money. When the central bank creates money in large quantities, the value of money falls quickly. To maintain stable prices, the central bank must limit growth in the money supply.

The costs of inflation are more subtle. They include shoeleather costs, menu costs, increased variability of relative prices, unintended changes in tax liabilities, confusion and inconvenience, and arbitrary redistributions of wealth. Are these costs, in total, large or small? All economists agree that they become huge during

the great illusion of the value and power of money, on which the capitalist state is based, will have been definitely destroyed."

Lenin's dark musings seemed almost prophetic in the jittery aftermath of World War I. Weimar Germany had gambled, badly, in financing its losing war effort with borrowed funds. Buried in debt and forced, in 1921, to pay reparations to the victorious Allies, Germany printed bank notes and ignited the most infamous bout of hyperinflation. By late 1923, prices were doubling roughly every three and a half days, and at one point a single American dollar was worth 6.7 trillion German marks. An even more severe hyperinflation followed the end of World War II, when Hungary printed notes of ever-higher value to finance its recovery. The fastest-ever recorded hyperinflation resulted: At its peak in July 1946, prices doubled every 15 hours.

War has often played a catalytic role in hyperinflation, but it rarely acts alone. In the early 1990s, the phenomenon stalked countries in Eastern Europe (Yugoslavia, Bosnia-Herzegovina, Armenia) that were confronting wars and the fall of the Soviet Union. A decade later in Zimbabwe, despite a long slide in agricultural output, Robert Mugabe's regime printed money to pay the bloated bureaucracy and to line its own pockets. By the time Mugabe declared inflation illegal in 2007, people had lost belief in their currency. Within a year, inflation shot up to 79.6 billion percent, so high that even the government's $100 trillion bills became useless souvenirs soon after they were printed.

Hyperinflation is not, as some might assume, just inflation gone bad. It's a different beast altogether, driven by politics and psychology as much as economics. A government's decision to continue spending (or printing money) far beyond its means is political, whether done to finance war, win an election or assuage its populace. Such monetary incontinence, unchecked, leads to a spiral of food shortages, price hikes and currency devaluations. Those hit hardest are not the rich (whose wealth is in property, stocks and commodities) but the middle class, which depends on local-currency salaries, savings and pensions whose value is siphoned off by hyperinflation.

No conflict or natural calamity can be blamed for Venezuela's descent into chaos. Its leaders did this on their own. With proven oil reserves of 300 billion barrels—surpassing even Saudi Arabia's—Venezuela should be rich. But the country's early oil boom, led largely by foreign companies, yielded only spotty development. When Hugo Chávez won the presidency in 1998, he vowed to give power and wealth to the people. Buoyed by a sustained rise in oil prices, he nationalized companies and funneled oil revenues into welfare programs and food imports. Poverty and unemployment rates fell by half. When oil prices cratered in 2008, Chávez kept spending as if nothing had changed. Since his death in 2013, his successor, Nicolás Maduro, has doubled down on Chávez's policies, even as he has violently repressed the opposition. . . .

A new era isn't likely to begin as long as Maduro remains in power. He has shown no interest in taking steps that might restore economic balance, like cutting spending and tying the bolívar to a solid foreign currency. Washington murmurs about regime change. But the biggest threat to Maduro now may be a series of civil cases in American courts against Citgo. The Venezuela-owned company is the regime's biggest generator of hard currency, the only asset creditors can go after. If these cases succeed in claiming damages for being nationalized by the Chávez regime, Maduro's main lifeline could be cut off. "If the money disappears," de Bolle says, "so does his support, and the regime crumbles." Only then, it seems, will Venezuelans be able to escape a nightmare in which they can't trust the money in their hands. ■

Questions to Discuss

1. The article mentions that Venezuela's hyperinflation coincided with a shrinking economy. How does declining output contribute to hyperinflation? How does hyperinflation contribute to declining output?
2. Why do you think politicians pursue policies that lead to hyperinflation?

Source: *New York Times*, November 4, 2018.

hyperinflation. But during periods of moderate inflation—when prices rise by less than 10 percent per year—the size of these costs is more open to debate.

This chapter presented many of the most important lessons about inflation, but the analysis is incomplete. When the central bank reduces the rate of money growth, prices rise less rapidly, as the quantity theory suggests. Yet as the economy makes the transition to the lower inflation rate, the change in monetary policy will likely disrupt production and employment. That is, even though monetary policy is neutral in the long run, it has profound effects on real variables in the short run. Later in this book, we will examine the reasons for short-run monetary non-neutrality to enhance our understanding of the causes and effects of inflation.

Chapter in a Nutshell

- The overall level of prices in an economy adjusts to bring money supply and money demand into balance. When the central bank increases the supply of money, it causes the price level to rise. Persistent growth in the quantity of money leads to continuing inflation.
- The principle of monetary neutrality asserts that changes in the quantity of money influence nominal variables but not real variables. Most economists believe that monetary neutrality approximately describes the behavior of the economy in the long run.
- A government can pay for some of its spending simply by printing money. When countries rely heavily on this "inflation tax," the result is hyperinflation.
- One application of the principle of monetary neutrality is the Fisher effect: When the expected inflation rate increases, the nominal interest rate increases by the same amount so the real interest rate remains the same.
- Many people think that inflation makes them poorer by raising the cost of what they buy. This view is a fallacy because inflation also raises nominal incomes.
- Economists have identified six costs of inflation: shoeleather costs associated with reduced money holdings, menu costs associated with more frequent adjustment of prices, increased variability of relative prices, unintended changes in tax liabilities because of nonindexation of the tax code, confusion and inconvenience resulting from a changing unit of account, and arbitrary redistributions of wealth between debtors and creditors. Many of these costs are large during hyperinflation, but the size of these costs for moderate inflation is less clear.

Key Concepts

quantity theory of money, p. 643
nominal variables, p. 645
real variables, p. 645
classical dichotomy, p. 645
monetary neutrality, p. 646
velocity of money, p. 646
quantity equation, p. 647
inflation tax, p. 649
Fisher effect, p. 651
shoeleather costs, p. 653
menu costs, p. 654

Questions for Review

1. Explain how an increase in the price level affects the real value of money.
2. According to the quantity theory of money, what is the effect of an increase in the quantity of money?
3. Explain the difference between nominal variables and real variables and give two examples of each. According to the principle of monetary neutrality, which variables are affected by changes in the quantity of money?
4. In what sense is inflation like a tax? How does thinking about inflation as a tax help explain hyperinflation?
5. According to the Fisher effect, how does an increase in the inflation rate affect the real interest rate and the nominal interest rate?
6. What are the costs of inflation? Which of these costs do you think are most important for the U.S. economy?
7. If inflation is less than expected, who benefits—debtors or creditors? Explain.

Problems and Applications

1. Suppose that this year's money supply is $500 billion, nominal GDP is $10 trillion, and real GDP is $5 trillion.
   a. What is the price level? What is the velocity of money?
   b. Suppose that velocity is constant, and the economy's output of goods and services rises by 5 percent each year. What will happen to nominal GDP and the price level next year if the Fed keeps the money supply constant?
   c. What money supply should the Fed set next year if it wants to keep the price level stable?
   d. What money supply should the Fed set next year if it wants inflation of 10 percent?

2. Suppose that changes in bank regulations expand the availability of credit cards so that people can hold less cash.
   a. How does this event affect the demand for money?
   b. If the Fed does not respond to this event, what will happen to the price level?
   c. If the Fed wants to keep the price level stable, what should it do?
3. It is sometimes suggested that the Fed should try to achieve zero inflation. If we assume that velocity is constant, does this zero-inflation goal require that the rate of money growth equal zero? If yes, explain why. If no, explain what the rate of money growth should equal.
4. Suppose that a country's inflation rate increases sharply. What happens to the inflation tax on the holders of money? Why is wealth held in savings accounts **not** subject to a change in the inflation tax? Can you think of any way in which holders of savings accounts are hurt by the increase in inflation?
5. Consider the effects of inflation in an economy composed of only two people: Bob, a bean farmer, and Rita, a rice farmer. Bob and Rita both always consume equal amounts of rice and beans. In 2022, the price of beans was \$1, and the price of rice was \$3.
   a. Suppose that in 2023, the price of beans was \$2, and the price of rice was \$6. What was inflation? Did the price changes leave Bob better off, worse off, or unaffected? What about Rita?
   b. Now suppose that in 2023, the price of beans was \$2, and the price of rice was \$4. What was inflation? Did the price changes leave Bob better off, worse off, or unaffected? What about Rita?
   c. Finally, suppose that in 2023, the price of beans was \$2, and the price of rice was \$1.50. What was inflation? Did the price changes leave Bob better off, worse off, or unaffected? What about Rita?
   d. What matters more to Bob and Rita—the overall inflation rate or the relative price of rice and beans?
6. Assuming a tax rate of 40 percent, compute the before-tax real interest rate and the after-tax real interest rate for each of the following cases.
   a. The nominal interest rate is 10 percent, and the inflation rate is 5 percent.
   b. The nominal interest rate is 6 percent, and the inflation rate is 2 percent.
   c. The nominal interest rate is 4 percent, and the inflation rate is 1 percent.
7. Recall that money serves three functions in the economy. What are those functions? How does inflation affect the ability of money to serve each of these functions?
8. Suppose that people expect inflation to be 3 percent but that, in fact, prices rise by 5 percent. Describe how this unexpectedly high inflation would help or hurt the following:
   a. the government
   b. a homeowner with a fixed-rate mortgage
   c. a union worker in the second year of a labor contract
   d. a college that has invested some of its endowment in government bonds
9. Explain whether the following statements are true, false, or uncertain.
   a. "Inflation hurts borrowers and helps lenders because borrowers must pay a higher rate of interest."
   b. "If prices change in a way that leaves the overall price level unchanged, then no one is made better or worse off."
   c. "Inflation does not reduce the purchasing power of most workers."

QuickQuiz Answers

1. **d** 2. **d** 3. **b** 4. **b** 5. **a** 6. **c** 7. **d** 8. **c**

Chapter

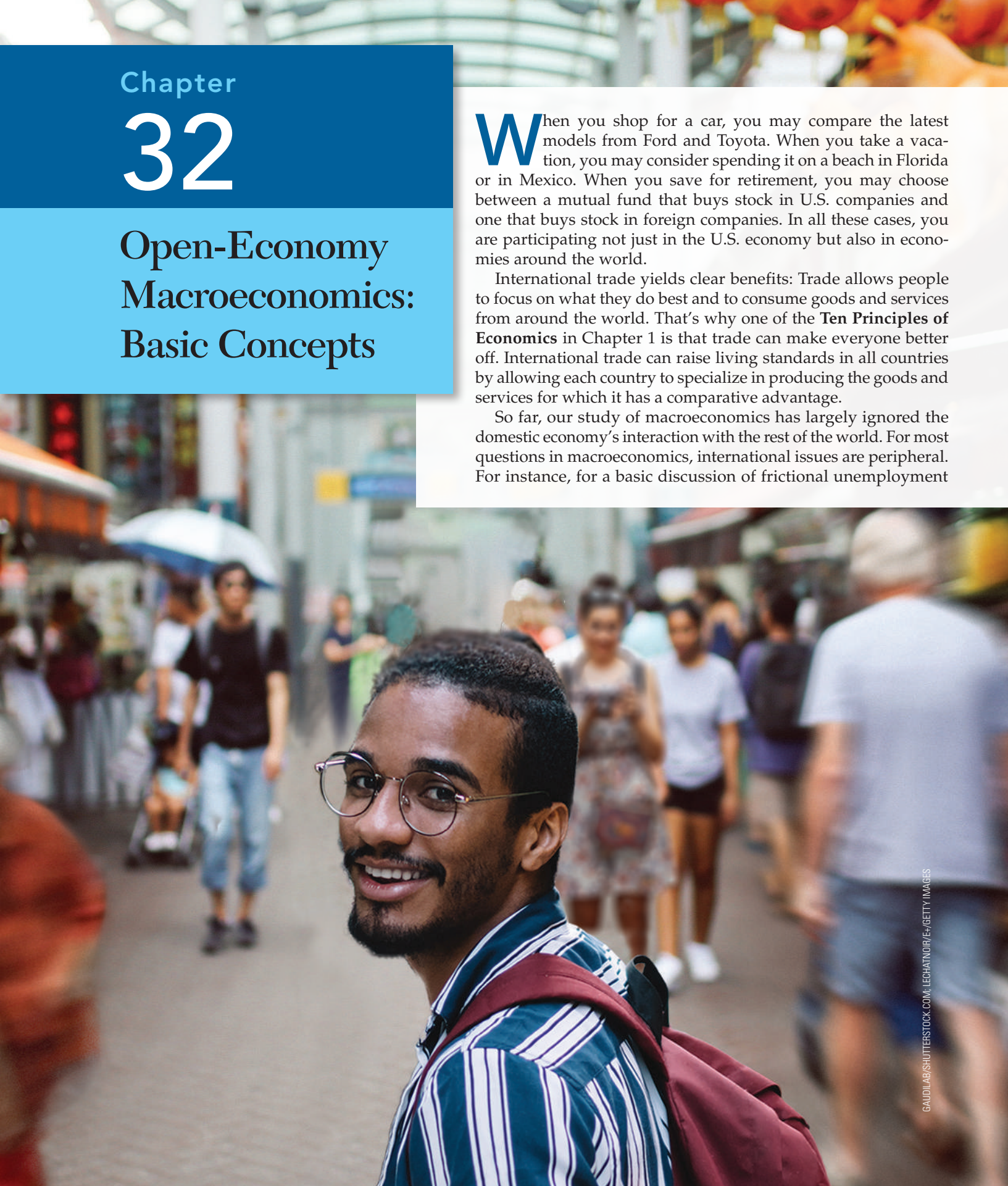
32

Open-Economy Macroeconomics: Basic Concepts

When you shop for a car, you may compare the latest models from Ford and Toyota. When you take a vacation, you may consider spending it on a beach in Florida or in Mexico. When you save for retirement, you may choose between a mutual fund that buys stock in U.S. companies and one that buys stock in foreign companies. In all these cases, you are participating not just in the U.S. economy but also in economies around the world.

International trade yields clear benefits: Trade allows people to focus on what they do best and to consume goods and services from around the world. That's why one of the **Ten Principles of Economics** in Chapter 1 is that trade can make everyone better off. International trade can raise living standards in all countries by allowing each country to specialize in producing the goods and services for which it has a comparative advantage.

So far, our study of macroeconomics has largely ignored the domestic economy's interaction with the rest of the world. For most questions in macroeconomics, international issues are peripheral. For instance, for a basic discussion of frictional unemployment



GAUDILAB/SHUTTERSTOCK.COM; LECHATNOIR/E+/GETTY IMAGES

or the quantity theory of money, the effects of international trade can safely be ignored. Indeed, to simplify their models, macroeconomists often assume a **closed economy**—an economy that does not interact with other economies.

closed economy
an economy that does not interact with other economies in the world

Yet when macroeconomists study an **open economy**—one that interacts freely with other economies around the world—they encounter a whole set of new issues. This chapter and the next introduce open-economy macroeconomics. We first discuss the key macroeconomic variables that describe an open economy's interactions in world markets. You may have heard of these variables—exports, imports, the trade balance, and exchange rates—from the news. This chapter explains what these data mean. The next chapter develops a model of how these variables are determined and how government policies affect them.

open economy
an economy that interacts freely with other economies around the world

32-1 The International Flows of Goods and Capital

An open economy interacts with other economies in two ways: It buys and sells goods and services in world product markets, and it buys and sells capital assets such as stocks and bonds in world financial markets. Here we discuss these two activities and the close relationship between them.

32-1a The Flow of Goods: Exports, Imports, and Net Exports

Exports are goods and services that are produced domestically and sold abroad, and **imports** are goods and services that are produced abroad and sold domestically. When Boeing, the U.S. aircraft manufacturer, builds a plane and sells it to Air France, the sale is an export for the United States and an import for France. When Fiat, the Italian car company, makes a car and sells it to a U.S. resident, the sale is an import for the United States and an export for Italy.

exports
goods and services produced domestically and sold abroad

imports
goods and services produced abroad and sold domestically

The **net exports** of any country are the difference between the value of its exports and the value of its imports:

$$\text{Net exports} = \text{Value of country's exports} - \text{Value of country's imports}.$$

net exports
the value of a nation's exports minus the value of its imports; also called the trade balance

The Boeing sale raises U.S. net exports, and the Fiat sale reduces them. Because net exports tell us whether a country is, in sum, a seller or a buyer in world markets for goods and services, net exports are also called the **trade balance**. If a country's net exports are positive, its exports exceed its imports, indicating that the country sells more goods and services abroad than it buys from other countries. In this case, the country is said to run a **trade surplus**. If a country's net exports are negative, its exports are less than its imports, indicating that it sells fewer goods and services abroad than it buys from other countries. In this case, the country is said to run a **trade deficit**. If a country's net exports are zero, its exports and imports are equal, and the country is said to have **balanced trade**.

trade balance
the value of a nation's exports minus the value of its imports; also called net exports

trade surplus
an excess of exports over imports

trade deficit
an excess of imports over exports

balanced trade
a situation in which exports equal imports

The next chapter develops a theory that explains an economy's trade balance, but even at this early stage, it is easy to think of many factors that might influence a country's exports, imports, and net exports. They include:

- Consumer tastes for domestic and foreign goods
- The prices of goods at home and abroad
- The exchange rates at which people can use domestic currency to buy foreign currencies

- The incomes of consumers at home and abroad
- The cost of transporting goods from country to country
- Government policies toward international trade

As these factors change, so does the amount of international trade.

MORT GERBERG/THE NEW YORKER COLLECTION/WWW.CARTOONBANK.COM

"But we're not just talking about buying a car—we're talking about confronting this country's trade deficit with Japan."

Case Study

The Increasing Openness of the U.S. Economy

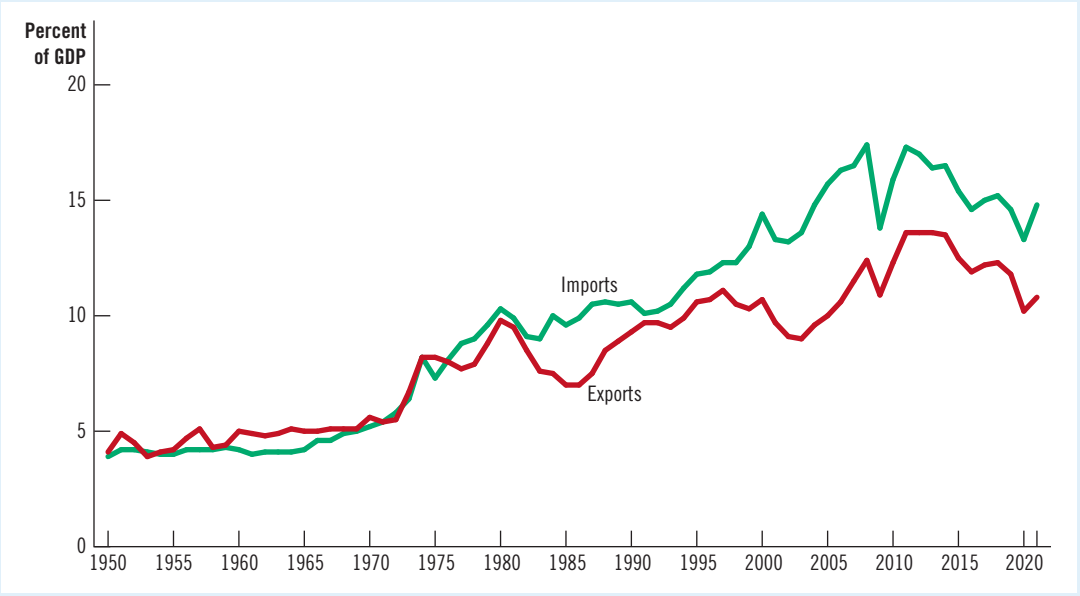
A significant change in the U.S. economy over the past seven decades has been an increase in the importance of international trade and finance. It is illustrated in Figure 1, which shows the total value of goods and services exported and imported expressed as a percentage of GDP. Despite some short-term drops, such as during the pandemic of 2020, the general trend is upward. In the 1950s, imports and exports were typically between 4 and 5 percent of GDP. In recent years, they have been about two or three times that level. The trading partners of the United States include a diverse group of countries. As of 2021, the largest trading partners, as measured by the sum of imports and exports, were Mexico, Canada, and China, each of which represented about 14 percent of total U.S. trade, followed by Japan, Germany, South Korea, and the United Kingdom.

Figure 1

The Internationalization of the U.S. Economy

This figure shows exports and imports of the U.S. economy as a percentage of U.S. GDP since 1950. The substantial increases over time show the increasing importance of international trade and finance.

Source: U.S. Department of Commerce.



The tendency for international trade to increase during the past several decades is partly due to improvements in transportation. In 1950, the average merchant ship carried less than 10,000 tons of cargo; today, many ships carry more than 100,000 tons. The long-distance jet was introduced in 1958 and the wide-body jet in 1967, making air transport far cheaper than it had been previously. These developments have allowed goods that once had to be produced locally to be traded around the world. Cut flowers grown in Israel are flown to the United States to be sold. Fresh fruits and vegetables that can grow in the United States only in the summer can now be consumed there in the winter because they can be shipped from the Southern Hemisphere.

Advances in telecommunications have also facilitated the increase in international trade by allowing businesses to reach overseas customers more easily. For example, the first transatlantic telephone cable was not laid until 1956. As recently as 1966, the technology allowed only 138 simultaneous conversations between North America and Europe. Today, thanks to e-mail and videoconferencing, it is often as easy to communicate with a customer across the world as it is to communicate with one across town.

Technological progress has also fostered trade by changing the kinds of goods that economies produce. When bulky raw materials (such as steel) and perishable goods (such as foodstuffs) were a large part of the world's output, transporting goods was often costly and sometimes impossible. By contrast, goods produced with modern technology are often light and easy to transport. Consumer electronics, for instance, have low weight for every dollar of value, making them easy to produce in one country and sell in another. An even more extreme example is the film industry. Once a studio in Hollywood makes a movie, it can send the film to audiences abroad at almost zero cost. And indeed, movies are a major export of the United States.

For most of this period, governments' trade policies were also a factor in increasing international trade. As we learned earlier in this book, economists have long believed that free trade between countries is mutually beneficial. Over time, most policymakers around the world have come to accept these conclusions. International agreements have gradually lowered tariffs, import quotas, and other trade barriers. These pacts include the General Agreement on Tariffs and Trade (GATT); the North American Free Trade Agreement (NAFTA), along with its successor, the United States–Mexico–Canada Agreement (USMCA); and many agreements between the United States and individual countries, such as Australia, Chile, Colombia, Panama, Singapore, and South Korea. Thus, the pattern of increasing trade shown in Figure 1 is a phenomenon that most policymakers have endorsed and encouraged.

During his presidency, Donald Trump challenged these trends. He said that the United States failed to benefit from past trade agreements, and he used his authority to impose tariffs on many foreign goods. As this book was going to press, the intentions of the Biden administration were not yet entirely clear, though some reductions in the tariffs imposed by President Trump were underway. ●

32-1b The Flow of Financial Resources: Net Capital Outflow

Residents of open economies participate not only in world markets for goods and services but also in world financial markets. A U.S. resident with $30,000 could use that money to buy a car from Toyota or stock in the Toyota Corporation. The first transaction would be a flow of goods, while the second would be a flow of capital.

The term **net capital outflow** refers to the difference between the purchase of foreign assets by domestic residents and the purchase of domestic assets by foreigners:

$$\begin{aligned}\text{Net capital outflow} = {} & \text{Purchase of foreign assets by domestic residents} \\ & - \text{Purchase of domestic assets by foreigners.}\end{aligned}$$

net capital outflow the purchase of foreign assets by domestic residents minus the purchase of domestic assets by foreigners

When a U.S. resident buys stock in Petrobras, the Brazilian energy company, the purchase increases the first term on the right side of this equation and, therefore, increases U.S. net capital outflow. When a resident of Japan buys a bond issued by the U.S. government, the purchase increases the second term on the right side of this equation and, therefore, decreases U.S. net capital outflow.

The flow of capital between the U.S. economy and the rest of the world takes two forms: **foreign direct investment** and **foreign portfolio investment**. An example of foreign direct investment is McDonald's opening a fast-food outlet in Russia. An example of foreign portfolio investment is an American buying stock in a Russian corporation. In the first case, the American owner (McDonald's Corporation) actively manages the investment, while in the second case, the American owner (the stockholder) has a more passive role. In both cases, U.S. residents are buying assets located in another country, so both purchases increase U.S. net capital outflow.

The net capital outflow (sometimes called **net foreign investment**) can be either positive or negative. When it is positive, domestic residents are buying more foreign assets than foreigners are buying domestic assets, and capital is said to be flowing out of the country. When the net capital outflow is negative, domestic residents are buying fewer foreign assets than foreigners are buying domestic assets, and capital is said to be flowing into the country. That is, when net capital outflow is negative, a country is experiencing a capital inflow.

The next chapter develops a theory to explain net capital outflow. In the meantime, let's briefly consider some of the variables that influence it:

- The real interest rates paid on foreign assets
- The real interest rates paid on domestic assets
- The perceived economic and political risks of holding assets abroad
- The government policies that affect foreign ownership of domestic assets

For example, consider U.S. investors deciding whether to buy Mexican government bonds or U.S. government bonds. (Recall that a bond is, in essence, an IOU of the issuer.) To make this decision, U.S. investors compare the real interest rates offered on the two bonds. The higher a bond's real interest rate, the more attractive it is. While making this comparison, however, U.S. investors must also take into account the risk that one of these governments might default on its debt (that is, not pay interest or principal when it is due), as well as any restrictions that the Mexican government has imposed, or might impose in the future, on foreign investors in Mexico.

32-1c The Equality of Net Exports and Net Capital Outflow

We have seen that an open economy interacts with the rest of the world in two ways—in markets for goods and services and in financial markets. Net exports and net capital outflow each measure a type of imbalance in these markets. Net exports measure an imbalance between a country's exports and its imports. Net capital outflow measures an imbalance between the amount of foreign assets bought by domestic residents and the amount of domestic assets bought by foreigners.

An important but subtle fact of accounting states that, for an economy as a whole, net capital outflow (*NCO*) equals net exports (*NX*):

$$NCO = NX.$$

This equation holds because every transaction that affects one side of this equation affects the other side by the same amount. This equation is an **identity**—an equation that must hold because of how the variables in the equation are defined and measured.

To see why this accounting identity is true, consider an example. Imagine that you are a computer programmer in the United States. One day, you write some software and sell it to a Japanese consumer for 10,000 yen. The sale of software is an export of the United States, so it increases U.S. net exports. What else happens to ensure that this identity holds? The answer depends on what you do with the 10,000 yen you are paid.

First, suppose that you simply stuff the yen in your mattress. (We might say you have a yen for yen.) In this case, you are using some of your income to invest in the Japanese economy. That is, a domestic resident (you) has acquired a foreign asset (the Japanese currency). The increase in U.S. net exports is matched by an increase in the U.S. net capital outflow.

But if you want to invest in the Japanese economy, you probably won't do so by holding on to Japanese currency. More likely, you will use the 10,000 yen to buy stock in a Japanese corporation, or you might buy a Japanese government bond. Yet, from the standpoint of economic accounting, the result of your decision is much the same: A domestic resident ends up acquiring a foreign asset. The increase in U.S. net capital outflow (the purchase of the Japanese stock or bond) equals the increase in U.S. net exports (the sale of software).

Now suppose that instead of using the 10,000 yen to buy a Japanese asset, you use it to buy a good made in Japan, such as a Sony TV. Because of your TV purchase, U.S. imports increase. Together, the software export and the TV import represent balanced trade. Because exports and imports increase by the same amount, net exports are unchanged. In this case, no American ends up acquiring a foreign asset, and no foreigner ends up acquiring a U.S. asset, so there is also no impact on U.S. net capital outflow.

A final possibility is that you go to a local bank to exchange your 10,000 yen for U.S. dollars. But this decision doesn't change the situation because the bank must now do something with the 10,000 yen. It can buy Japanese assets (a U.S. net capital outflow), it can buy a Japanese good (a U.S. import), or it can sell the yen to another American who wants to make such a transaction. In the end, U.S. net exports must equal U.S. net capital outflow.

This example started with a U.S. programmer selling some software abroad, but the story is much the same when Americans buy goods and services from other countries. For example, if Walmart buys $50 million of clothing from China and sells it to American consumers, something must happen to that $50 million. China could invest the $50 million in the U.S. economy. This capital inflow from China might take the form of Chinese purchases of U.S. government bonds. In this case, the purchase of the clothing reduces U.S. net exports, and the sale of bonds reduces U.S. net capital outflow. Alternatively, China could use the $50 million to buy a plane from Boeing, the U.S. aircraft manufacturer. In this case, the U.S. import of clothing balances the U.S. export of aircraft, so net exports and net capital outflow are both unchanged. In all cases, the transactions have the same effect on net exports and net capital outflow.

Let's recap:

- When a nation is running a trade surplus ($NX > 0$), it is selling more goods and services to foreigners than it is buying from them. What is it doing with the foreign currency it receives from the net sale of goods and services abroad? It must be using it to buy foreign assets. Capital is flowing out of the country ($NCO > 0$).
- When a nation is running a trade deficit ($NX < 0$), it is buying more goods and services from foreigners than it is selling to them. How is it financing the net purchase of these goods and services in world markets? It must be selling assets abroad. Capital is flowing into the country ($NCO < 0$).

The international flow of goods and services and the international flow of capital are two sides of the same coin.

32-1d Saving, Investment, and Their Relationship to International Flows

A nation's saving and investment are crucial to its long-run economic growth. As an earlier chapter explained, in a closed economy, saving equals investment. But matters are not as simple in an open economy. We now consider how saving and investment are related to the international flows of goods and capital as measured by net exports and net capital outflow.

As you may recall, the term **net exports** appeared earlier in the book when we introduced gross domestic product. The economy's GDP (denoted Y) is divided among four components: consumption (C), investment (I), government purchases (G), and net exports (NX). We write this as

$$Y = C + I + G + NX.$$

The total expenditure on the economy's output of goods and services is the sum of expenditure on consumption, investment, government purchases, and net exports. Because each dollar of expenditure is placed into one of these four components, this equation is an accounting identity: It must be true because of the way the variables are defined and measured.

Recall that national saving is the income of the nation that is left after paying for current consumption and government purchases. National saving (S) equals $Y - C - G$. If we rearrange the equation to reflect this fact, we obtain

$$\begin{aligned} Y - C - G &= I + NX \\ S &= I + NX. \end{aligned}$$

Because net exports (NX) also equal net capital outflow (NCO), we can write this equation as

$$\begin{aligned} S &= I + NCO \\ \text{Saving} &= \text{Domestic investment} + \text{Net capital outflow}. \end{aligned}$$

This equation states that a nation's saving must equal its domestic investment plus its net capital outflow. In other words, when a U.S. household saves a dollar of its

income for the future, that dollar can be used to finance either the accumulation of domestic capital or the purchase of foreign capital.

This equation may look familiar. Earlier in the book, when analyzing the financial system, we saw this identity for the special case of a closed economy. In a closed economy, net capital outflow is zero ($NCO = 0$), so saving equals investment ($S = I$). By contrast, an open economy has two uses for its saving: domestic investment and net capital outflow.

As before, the financial system can be seen as standing between the two sides of this identity. For example, suppose the Garcia family decides to save some of its income for retirement. This decision contributes to national saving, the left side of the equation. If the Garcias deposit their saving in a mutual fund, the fund may use some of the deposit to buy stock issued by General Motors, which uses the proceeds to build a factory in Ohio. In addition, the mutual fund may use some of the Garcias' deposit to buy stock issued by Toyota, which uses the proceeds to build a factory in Osaka. These transactions show up on the right side of the equation. From the standpoint of U.S. accounting, the General Motors expenditure on the factory is domestic investment, and the purchase of Toyota stock by a U.S. resident is net capital outflow. Thus, all saving in the U.S. economy shows up as investment in the U.S. economy or as U.S. net capital outflow.

The bottom line is that saving, investment, and international capital flows are inextricably linked. When a nation's saving exceeds its domestic investment, its net capital outflow is positive, indicating that the nation is using some of its saving to buy assets abroad. When a nation's domestic investment exceeds its saving, its net capital outflow is negative, indicating that foreigners are financing some of this investment by purchasing domestic assets.

32-1e Summing Up

Table 1 summarizes many of these ideas. It describes the three possibilities for an open economy: a country with a trade deficit, a country with balanced trade, and a country with a trade surplus.

Consider first a country with a trade surplus. By definition, a trade surplus means that the value of exports exceeds the value of imports. Because net exports are exports minus imports, net exports, NX, are positive. As a result, income, $Y = C + I + G + NX$, exceeds domestic spending, $C + I + G$. But if income, Y, is more than spending, $C + I + G$, then saving, $S = Y - C - G$, must be more than investment, I. Because the country is saving more than it is investing, it must be sending some of its saving abroad. That is, the net capital outflow must be positive.

Table 1

International Flows of Goods and Capital: Summary

This table shows the three possible outcomes for an open economy.

Trade Deficit	Balanced Trade	Trade Surplus
Exports < Imports	Exports = Imports	Exports > Imports
Net Exports < 0	Net Exports = 0	Net Exports > 0
$Y < C + I + G$	$Y = C + I + G$	$Y > C + I + G$
Saving < Investment	Saving = Investment	Saving > Investment
Net Capital Outflow < 0	Net Capital Outflow = 0	Net Capital Outflow > 0

Similar logic applies to a country with a trade deficit (such as the U.S. economy in recent years). By definition, a trade deficit means that the value of exports is less than the value of imports. Because net exports are exports minus imports, net exports, NX, are negative. Thus, income, $Y = C + I + G + NX$, must be less than domestic spending, $C + I + G$. But if income, Y, is less than spending, $C + I + G$, then saving, $S = Y - C - G$, must be less than investment, I. Because the country is investing more than it is saving, it must be financing some domestic investment by selling assets abroad. That is, the net capital outflow must be negative.

A country with balanced trade falls between these cases. Exports equal imports, so net exports are zero. Income equals domestic spending, and saving equals investment. The net capital outflow equals zero.

Case Study

Is the U.S. Trade Deficit a National Problem?

You may have heard the United States called "the world's largest debtor." It earned that description by borrowing heavily in world financial markets during the past four decades to finance large trade deficits. Why did the United States do this, and should this practice give Americans reason to worry?

To answer these questions, let's see what the macroeconomic accounting identities tell us about the U.S. economy. Panel (a) of Figure 2 shows national saving and domestic investment as a percentage of GDP since 1960. Panel (b) shows net capital outflow (that is, the trade balance) as a percentage of GDP. Notice that, as the identities require, net capital outflow always equals national saving minus domestic investment. The figure shows that both national saving and domestic investment as a percentage of GDP fluctuate substantially over time. Before 1980, they tended to fluctuate together, so the net capital outflow was typically small—between −1 and 1 percent of GDP. Since 1980, national saving has often fallen well below domestic investment, leading to sizable trade deficits and substantial inflows of capital. That is, in recent decades, the net capital outflow has often been a large negative number.

The fluctuations in Figure 2 may seem bewildering. They have been the source of much controversy in American politics. To reach a deeper understanding, we need to go beyond these data and discuss the policies and events that influence national saving and domestic investment. History shows that there is no single cause of trade deficits. Rather, they can arise under a variety of circumstances. Here are four historical episodes.

Unbalanced fiscal policy: From 1980 to 1987, the flow of capital into the United States went from 0.5 to 2.9 percent of GDP. This 2.4 percentage point change is largely attributable to a fall in national saving of 2.7 percentage points. The decline in national saving, in turn, is often explained by the reduction in public saving—that is, the increase in the government budget deficit. These budget deficits arose because President Ronald Reagan cut taxes and increased defense spending without enacting his proposed cuts in nondefense spending.

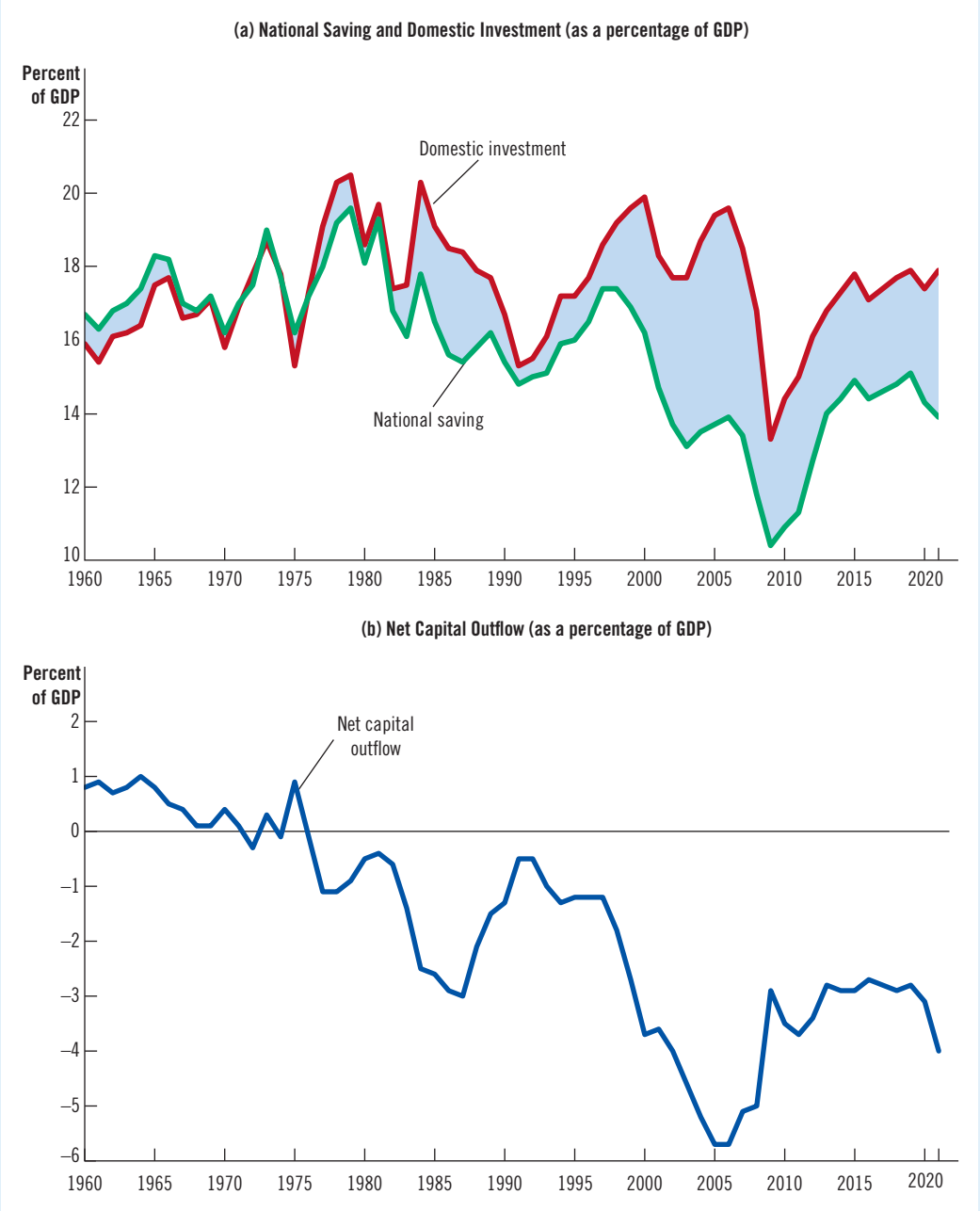
An investment boom: A different story explains the trade deficits that arose during the next decade. From 1991 to 2000, the capital flow into the United States went from 0.5 to 3.7 percent of GDP. None of this 3.2 percentage point change is attributable to a decline in saving; in fact, saving increased over this time as the government's budget switched from deficit to surplus. But investment went from 15.3 to 19.9 percent of GDP as information technology saw tremendous advances, and many firms were eager to make high-tech investments.

Figure 2

National Saving, Domestic Investment, and Net Capital Outflow

Panel (a) shows national saving and domestic investment as a percentage of GDP. Panel (b) shows net capital outflow as a percentage of GDP. You can see from the figure that national saving has been lower since 1980 than it was before 1980. This fall in national saving has been reflected primarily in reduced net capital outflow rather than in reduced domestic investment.

Source: U.S. Department of Commerce.



An economic downturn and recovery: From 2000 to 2019, the capital flow into the United States remained large. The consistency of this variable, however, stands in stark contrast to the remarkable changes in saving and investment. From 2000 to 2009, both fell by about 6 percentage points. Investment dropped because tough economic times made capital accumulation less profitable, while national saving declined because the government began running extraordinarily large budget deficits

in response to the recession of 2008–2009. From 2009 to 2019, as the economy recovered, these forces reversed themselves, and both saving and investment increased by more than 4 percentage points.

Pandemic spending: From 2019 to 2021, the U.S. trade deficit and capital inflow increased by 1.2 percentage points. This change is entirely attributable to a decline in national saving. During this time, the federal government vastly increased government spending in response to the Covid-19 pandemic, leading to large budget deficits and lower national saving.

Are these trade deficits and international capital flows a problem for the U.S. economy? There is no easy answer. One must evaluate the circumstances and possible alternatives.

Consider first a trade deficit induced by a fall in saving, as occurred during the 1980s. Lower saving means that the nation is putting away less income to provide for the future. Once national saving has fallen, however, there is no reason to deplore the resulting trade deficits. If national saving fell without inducing a trade deficit, investment in the United States would have to fall. This fall in investment, in turn, would adversely affect the growth in the capital stock, labor productivity, and real wages. In other words, once U.S. saving has declined, it is better to have foreigners invest in the U.S. economy than no one at all.

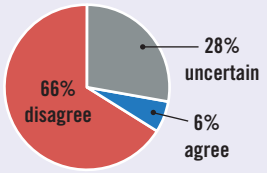
Now consider a trade deficit induced by an investment boom, like the trade deficits of the 1990s. In this case, the economy is borrowing from abroad to finance the purchase of new capital goods. If this additional capital provides a good return in the form of higher production of goods and services, then the economy should be able to handle the debts it has accumulated. On the other hand, if the investment projects fail to yield the expected returns, the debts will look less desirable, at least with the benefit of hindsight.

Just as an individual can go into debt in either a prudent or a profligate manner, so can a nation. A trade deficit is not a problem in itself, but it can sometimes be a symptom of a problem. ●

Ask the Experts: Trade Balances and Trade Negotiations

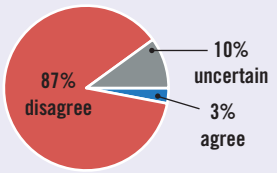
"A typical country can increase its citizens' welfare by enacting policies that would increase its trade surplus (or decrease its trade deficit)."

What do economists say?



"An important reason why many workers in Michigan and Ohio have lost jobs in recent years is because U.S. presidential administrations over the past 30 years have not been tough enough in trade negotiations."

What do economists say?



Source: IGM Economic Experts Panel, December 9, 2014 and March 22, 2016.

QuickQuiz

1. As a percentage of U.S. GDP, exports are now _____ and imports are now _____ than they were in 1950.
   a. higher; higher
   b. lower; lower
   c. higher; lower
   d. lower; higher
2. In an open economy, national saving equals domestic investment
   a. plus the government's budget deficit.
   b. minus the net exports of goods and services.
   c. plus the net outflow of capital.
   d. minus foreign portfolio investment.
3. If the value of a nation's imports exceeds the value of its exports, which of the following is NOT true?
   a. Net exports are negative.
   b. GDP is less than the sum of consumption, investment, and government purchases.
   c. Domestic investment exceeds national saving.
   d. The nation is experiencing a net outflow of capital.

Answers are at the end of the chapter.

32-2 The Prices for International Transactions: Real and Nominal Exchange Rates

So far, we have discussed measures of the flow of goods and services and the flow of capital across a nation's border. In addition to these quantity variables, macroeconomists also study the prices at which these international transactions take place. Just as the price in any market coordinates buyers and sellers, international prices coordinate consumers, producers, and investors as they interact in world markets. Two of the most important of these international prices are the nominal and real exchange rates.

32-2a Nominal Exchange Rates

nominal exchange rate
the rate at which a person can trade the currency of one country for the currency of another

The **nominal exchange rate** tells you how much you will receive if you trade the currency of one country for the currency of another. For example, when you go to a bank, you might see a posted exchange rate of 80 yen per dollar. If you give the bank 1 U.S. dollar, you will receive 80 Japanese yen in return, and if you give the bank 80 Japanese yen, you will receive 1 U.S. dollar. (In actuality, the bank posts slightly different prices for buying and selling yen. The difference compensates the bank for offering this service. For our purposes, we can ignore this detail.)

An exchange rate can always be expressed in two ways. If the exchange rate is 80 yen per dollar, it is also 1/80 (= 0.0125) dollar per yen. This book expresses the nominal exchange rate as units of foreign currency per U.S. dollar, such as 80 yen per dollar.

appreciation
an increase in the value of a currency as measured by the amount of foreign currency it can buy

depreciation
a decrease in the value of a currency as measured by the amount of foreign currency it can buy

If the exchange rate changes so a dollar buys more foreign currency, that change is called an **appreciation** of the dollar. If the exchange rate changes so a dollar buys less foreign currency, that is a **depreciation** of the dollar. For example, when the exchange rate rises from 80 to 90 yen per dollar, the dollar is said to appreciate. At the same time, because a Japanese yen now buys less of the U.S. currency, the yen is said to depreciate. When the exchange rate falls from 80 to 70 yen per dollar, the dollar is said to depreciate, and the yen is said to appreciate.

At times, you may have heard that the dollar is either "strong" or "weak." These descriptions usually refer to recent changes in the nominal exchange rate. When a currency appreciates, it is said to **strengthen** because it can then buy more foreign currency. Similarly, when a currency depreciates, it is said to **weaken**. Perhaps because the word "strong" is more stirring than the word "weak," government officials are prone to say that they favor a strong currency. From an economist's standpoint, however, exchange rates are just prices, and there is nothing inherently better about them being high or low.

For any country, there are many nominal exchange rates. The U.S. dollar can be used to buy Japanese yen, British pounds, Mexican pesos, and so on. When economists study changes in the exchange rate, they often use indexes that average these rates. Just as the consumer price index turns the vast number of prices in the economy into a single measure of the price level, an exchange-rate index turns many exchange rates into a single measure of a currency's international value. So when economists talk about the dollar appreciating or depreciating, they often are referring to an exchange-rate index that includes many individual exchange rates.

FYI

The Euro

You may have once heard of, or perhaps even seen, a French franc, a German mark, or an Italian lira. These former national currencies no longer function as money. During the 1990s, many European nations decided to give up their national currencies and instead use a common currency called the **euro**. The euro started circulating on January 1, 2002, when 12 nations began using it as their official money. As of 2021, the euro is used in about two dozen nations, encompassing 340 million people. A number of other European countries, however, have kept their own currencies, including the United Kingdom, Norway, Sweden, Denmark, Bulgaria, and the Czech Republic.

PETER STONE/ALAMY STOCK PHOTO

Monetary policy for the euro area is set by the European Central Bank (ECB), which brings together representatives from all the participating countries. The ECB issues the euro and controls the supply of this money, much as the Federal Reserve controls the supply of dollars in the U.S. economy.

Why did these countries adopt a common currency? One benefit is that it makes trade easier. Imagine if each of the 50 U.S. states had a different currency. Every time you crossed a state border, you would need to change your money. And when comparing prices in different states, you would have to perform the kind of exchange-rate calculations discussed in the text. This would be inconvenient, and it might deter you from buying goods and services outside your own state. The countries of Europe decided that as their economies became more integrated, it would be better to avoid this inconvenience.

To some extent, the adoption of a common currency in Europe was a political decision that went beyond economic concerns. Some advocates of the euro wanted to reduce nationalistic feelings and to make Europeans appreciate more fully their shared history and destiny. A single money for most of the continent, they argued, would help achieve this goal.

There are, however, costs to adopting a common currency. When nations use the same money, they are committing to sharing a single monetary policy. If they disagree about what monetary policy is best, they will have to compromise, rather than each country going its own way. Because a common currency has both benefits and costs, economists differ on whether Europe's adoption of the euro has been a good decision.

From 2010 to 2015, the question of the euro as a pan-European currency heated up as several European nations dealt with economic difficulties. Greece, in particular, had run up a large government debt and found itself facing possible default. As a result, it had to raise taxes and cut back government spending substantially. Some observers suggested that dealing with these problems would have been easier if the Greek government had had an additional tool—a national monetary policy. The possibility of Greece leaving the euro area and reintroducing its own currency was discussed, but in the end, that outcome did not occur. ■

32-2b Real Exchange Rates

The **real exchange rate** is the rate at which you can trade the goods and services of one country for those of another. For example, if you go shopping and find that a pound of Swiss cheese is twice as expensive as a pound of American cheese, the real exchange rate is ½ pound of Swiss cheese per pound of American cheese. Notice that, like the nominal exchange rate, the real exchange rate is expressed as units of the foreign item per unit of the domestic item. But in this instance, the item is a good rather than a currency.

real exchange rate the rate at which a person can trade the goods and services of one country for the goods and services of another

Real and nominal exchange rates are closely related. For example, suppose that a bushel of American rice sells for $100, and a bushel of Japanese rice sells for 16,000 yen. What is the real exchange rate between American and Japanese rice? To answer this question, we must first use the nominal exchange rate to convert the prices into a common currency. If the nominal exchange rate is 80 yen per dollar,

then the price for American rice of $100 per bushel is equivalent to 8,000 yen per bushel. American rice is half as expensive as Japanese rice. The real exchange rate is ½ bushel of Japanese rice per bushel of American rice.

We can summarize this calculation for the real exchange rate with the following formula:

$$\text{Real exchange rate} = \frac{\text{Nominal exchange rate} \times \text{Domestic price}}{\text{Foreign price}}.$$

Using the numbers in our example, the formula applies as follows:

$$\begin{aligned}\text{Real exchange rate} &= \frac{(80 \text{ yen/dollar}) \times (\$100/\text{bushel of American rice})}{16{,}000 \text{ yen/bushel of Japanese rice}} \\ &= \frac{8{,}000 \text{ yen/bushel of American rice}}{16{,}000 \text{ yen/bushel of Japanese rice}} \\ &= \tfrac{1}{2} \text{ bushel of Japanese rice/bushel of American rice.}\end{aligned}$$

Thus, the real exchange rate depends on the nominal exchange rate and on the prices of goods in the two countries measured in the local currencies.

The real exchange rate is a key determinant of how much a country exports and imports. When Ben's Original is deciding whether to buy U.S. rice or Japanese rice to put into its boxes, it will ask which rice is cheaper. The real exchange rate gives the answer. As another example, imagine that you are deciding whether to take a seaside vacation in Miami, Florida, or in Cancún, Mexico. You might ask your travel agent the price of a hotel room in Miami (measured in dollars), the price of a hotel room in Cancún (measured in pesos), and the exchange rate between pesos and dollars. If you decide where to vacation by comparing costs, you are basing your decision on the real exchange rate.

When studying an economy as a whole, macroeconomists focus on overall prices rather than the prices of individual items. That is, to measure the real exchange rate, they use price indexes, such as the consumer price index, which measure the price of a basket of goods and services. By using a price index for a U.S. basket (P), a price index for a foreign basket (P^*), and the nominal exchange rate between the U.S. dollar and foreign currencies (e), we can compute the overall real exchange rate between the United States and other countries as follows:

$$\text{Real exchange rate} = (e \times P)/P^*.$$

This real exchange rate measures the price of a basket of goods and services available domestically relative to a basket of goods and services available abroad.

As the next chapter will discuss more fully, a country's real exchange rate is a key determinant of its net exports of goods and services. A depreciation (fall) in the U.S. real exchange rate means that U.S. goods have become cheaper compared with foreign goods. This change encourages consumers both at home and abroad to buy more from the United States and less from other countries. As a result, U.S. exports rise, and U.S. imports fall. Both of these changes increase U.S. net exports. Conversely, an appreciation (rise) in the U.S. real exchange rate means that U.S. goods have become more expensive compared with foreign goods, so U.S. net exports decline.

QuickQuiz

4. If a nation's currency doubles in value on foreign exchange markets, the currency is said to _____, reflecting a change in the _____ exchange rate.
   a. appreciate; nominal
   b. appreciate; real
   c. depreciate; nominal
   d. depreciate; real
5. If the U.S. dollar appreciates and prices remain the same at home and abroad, foreign goods become _____ expensive relative to American goods, pushing the U.S. trade balance toward _____.
   a. more; surplus
   b. more; deficit
   c. less; surplus
   d. less; deficit
6. The dollar-yen exchange rate falls from 100 to 80 yen per dollar. At the same time, the price level in the United States rises from 180 to 200, and the price level in Japan remains the same. As a result,
   a. American goods have become more expensive relative to Japanese goods.
   b. American goods have become less expensive relative to Japanese goods.
   c. the relative price of American and Japanese goods has not changed.
   d. both American and Japanese goods have become relatively less expensive.

Answers are at the end of the chapter.

32-3 A First Theory of Exchange-Rate Determination: Purchasing-Power Parity

Exchange rates vary substantially over time. In 1970, a U.S. dollar could buy 3.65 German marks or 627 Italian lira. In 1998, as both Germany and Italy were getting ready to adopt the euro as their common currency, a U.S. dollar could buy 1.76 German marks or 1,737 Italian lira. In other words, over this period, the value of the dollar fell by more than half compared with the mark, while it more than doubled compared with the lira.

What explains these large and opposite changes? Economists have developed many models to explain how exchange rates are determined, each emphasizing some of the forces at work. Here we consider a simple theory, called **purchasing-power parity**, which is generally believed to describe the forces that influence exchange rates in the long run. It states that a unit of any given currency should be able to buy the same quantity of goods in all countries. Let's consider the theory's logic as well as its implications and limitations.

purchasing-power parity
a theory of exchange rates that says a unit of any given currency should be able to buy the same quantity of goods in all countries

32-3a The Basic Logic of Purchasing-Power Parity

The theory of purchasing-power parity is based on a principle called the **law of one price**. This law asserts that a good must sell for the same price in all locations. Otherwise, there would be opportunities for profit left unexploited. For example, suppose that coffee beans sold for less in Seattle than in Dallas. A person could buy coffee in Seattle for, say, $4 a pound and then sell it in Dallas for $5 a pound, making a profit of $1 per pound from the difference in price. The process of taking advantage of price differences for the same item in different markets is called **arbitrage**. In this example, as people took advantage of the arbitrage opportunity, they would increase the demand for coffee in Seattle and increase the supply in Dallas. The price of coffee would rise in Seattle (in response to greater demand) and fall in Dallas (in response to greater supply). This process would continue until, eventually, the prices were the same in the two markets.

Now consider how the law of one price applies to the international marketplace. If a dollar (or any other currency) could buy more coffee in the United States than in Japan, international traders could profit by buying coffee in the United States and selling it in Japan. This export of coffee from the United States to Japan would drive up the U.S. price of coffee and drive down the Japanese price. Conversely, if a dollar could buy more coffee in Japan than in the United States, traders could buy coffee in Japan and sell it in the United States. This import of coffee into the United States from Japan would drive down the U.S. price of coffee and drive up the Japanese price. In the end, the law of one price tells us that a dollar must buy the same amount of coffee in all countries.

This logic leads to the theory of purchasing-power parity. According to this theory, a currency must have the same purchasing power in all countries. That is, a U.S. dollar must buy the same quantity of goods in the United States and Japan, and a Japanese yen must buy the same quantity of goods in Japan and the United States. The name of this theory describes it well. **Parity** means equality, and **purchasing power** refers to the value of money in terms of the quantity of goods it can buy. **Purchasing-power parity** states that a unit of a currency must have the same real value in every country.

32-3b Implications of Purchasing-Power Parity

What does the theory of purchasing-power parity say about exchange rates? It tells us that the nominal exchange rate between the currencies of two countries depends on the price levels in those countries. If a dollar buys the same quantity of goods in the United States (where prices are measured in dollars) as in Japan (where prices are measured in yen), then the number of yen per dollar must reflect the prices of goods in the United States and Japan. For example, if a pound of coffee costs 500 yen in Japan and \$5 in the United States, then the nominal exchange rate must be 100 yen per dollar (500 yen/\$5 = 100 yen per dollar). Otherwise, the purchasing power of the dollar would not be the same in the two countries.

To see more fully how this works, it is helpful to use just a bit of mathematics. Suppose that P is the price of a basket of goods in the United States (measured in dollars), P^* is the price of a basket of goods in Japan (measured in yen), and e is the nominal exchange rate (the number of yen a dollar can buy). Now consider the quantity of goods a dollar can buy at home and abroad. At home, the price level is P, so the purchasing power of \$1 at home is $1/P$. That is, a dollar can buy $1/P$ units of goods. Abroad, a dollar can be exchanged into e units of foreign currency, which in turn have purchasing power e/P^*. For the purchasing power of a dollar to be the same in the two countries, it must be the case that

$$1/P = e/P^*.$$

With rearrangement, this equation becomes

$$1 = eP/P^*.$$

Notice that the left side of this equation is a constant, and the right side is the real exchange rate. **If the purchasing power of the dollar is always the same at home and abroad, then the real exchange rate—the relative price of domestic and foreign goods—cannot change.**

To see the implication of this analysis for the nominal exchange rate, we can rearrange the last equation to solve for the nominal exchange rate:

$$e = P^*/P.$$

That is, the nominal exchange rate equals the ratio of the foreign price level (measured in units of the foreign currency) to the domestic price level (measured in units of the domestic currency). **According to the theory of purchasing-power parity, the nominal exchange rate between the currencies of two countries must reflect the price levels in those countries.**

A key implication of purchasing-power parity is that nominal exchange rates change when price levels change. As the preceding chapter explained, the price level in any country adjusts to bring the quantity of money supplied and the quantity of money demanded into balance. Because the nominal exchange rate depends on the price levels, it also depends on the money supply and demand in each country. When a central bank in any country increases the money supply and causes the price level to rise, it also causes that country's currency to depreciate relative to other currencies in the world. **In other words, when the central bank prints large quantities of money, that money loses value both in terms of the goods and services it can buy and in terms of the amount of other currencies it can buy.**

We can now answer the question that began this section: Why did the U.S. dollar lose value compared with the German mark and gain value compared with the Italian lira? The answer is that Germany pursued a less inflationary monetary policy than the United States, and Italy pursued a more inflationary monetary policy. From 1970 to 1998, inflation in the United States was 5.3 percent per year. By contrast, inflation was 3.5 percent in Germany and 9.6 percent in Italy. As U.S. prices rose relative to German prices, the value of the dollar fell relative to the mark. Similarly, as U.S. prices fell relative to Italian prices, the value of the dollar rose relative to the lira.

Germany and Italy now have a common currency—the euro. Sharing a currency means that the two countries share a single monetary policy and, as a result, have similar inflation rates. But the historical lessons of the lira and the mark will apply to the euro as well. Whether the U.S. dollar buys more or fewer euros 20 years from now than it does today depends on whether the European Central Bank generates more or less inflation in Europe than the Federal Reserve does in the United States.

The Nominal Exchange Rate during a Hyperinflation

Macroeconomists can only rarely conduct controlled experiments. Most often, they must glean what they can from the natural experiments that history gives them. One natural experiment is hyperinflation—the high inflation that arises when a government turns to the printing press to pay for large amounts of government spending. Because hyperinflations are so extreme, they illustrate some basic economic principles with clarity.

Consider the German hyperinflation of the early 1920s. Figure 3 shows the German money supply, the German price level, and the nominal exchange rate (measured as U.S. cents per German mark) for that period. Notice that these series move closely together. When the supply of money starts growing quickly, the price level also takes off, and the German mark depreciates. When the money supply stabilizes, so do the price level and the exchange rate.

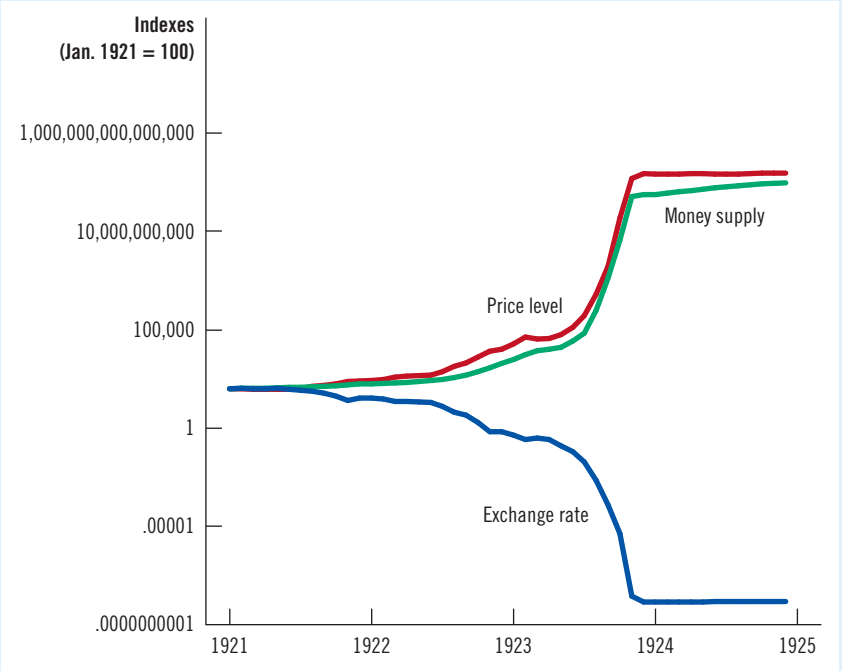
The pattern shown in this figure appears during every hyperinflation. It leaves no doubt that there is a fundamental link among money, prices, and the nominal exchange rate. The quantity theory of money discussed in the previous chapter explains how the money supply affects the price level. The theory of purchasing-power parity discussed here explains how the price level affects the nominal exchange rate. ●

Figure 3

Money, Prices, and the Nominal Exchange Rate during the German Hyperinflation

This figure shows the money supply, the price level, and the nominal exchange rate (measured as U.S. cents per mark) for the German hyperinflation from January 1921 to December 1924. Notice how similarly these three variables move. When the quantity of money started growing quickly, the price level followed, and the mark depreciated relative to the dollar. When the German central bank stabilized the money supply, the price level and exchange rate stabilized as well.

Source: Adapted from Thomas J. Sargent, "The End of Four Big Inflations," in Robert Hall, ed., *Inflation* (Chicago: University of Chicago Press, 1983), pp. 41–93.



32-3c Limitations of Purchasing-Power Parity

Purchasing-power parity provides a simple model of how exchange rates are determined. For understanding many economic phenomena, the theory works well. In particular, it can explain some important long-term trends, such as the depreciation of the U.S. dollar against the German mark and the appreciation of the U.S. dollar against the Italian lira. It can also explain the major changes in exchange rates that occur during hyperinflations.

Yet the theory of purchasing-power parity is not completely accurate. That is, exchange rates do not always move to ensure that a dollar has the same real value in all countries all the time. There are two reasons the theory of purchasing-power parity does not always hold in practice.

The first is that many goods are not easily traded. Imagine, for instance, that haircuts are more expensive in Paris than in New York. International travelers might avoid getting their haircuts in Paris, and some haircutters might move from New York to Paris. Yet such arbitrage would be too limited to eliminate the differences in prices. The deviation from purchasing-power parity might persist, and a dollar (or euro) would continue to buy less of a haircut in Paris than in New York.

The second reason that purchasing-power parity does not always hold is that even tradable goods are not always perfect substitutes when they are produced in different countries. For example, some consumers prefer German cars, and others prefer American cars. Moreover, consumer tastes can change over time. If German cars suddenly become more popular, the increase in demand will drive up the price of German cars relative to American cars. Despite this difference in prices in the two markets, there might be no opportunity for profitable arbitrage because consumers do not view the two country's cars as equivalent.

Both because some goods are not tradable and because some tradable goods are not perfect substitutes for their foreign counterparts, purchasing-power parity is not a perfect theory of exchange-rate determination. For these reasons, real exchange rates fluctuate over time. Nonetheless, the theory of purchasing-power parity does provide a useful first step in understanding exchange rates. The basic logic is persuasive: As the real exchange rate drifts from the level predicted by purchasing-power parity, people have a greater incentive to move goods across national borders. Even if the forces of purchasing-power parity do not completely fix the real exchange rate, they provide a reason to expect that changes in the real exchange rate are most often small or temporary. As a result, large and persistent movements in nominal exchange rates typically reflect changes in price levels at home and abroad.

Case Study

The Hamburger Standard

When economists apply the theory of purchasing-power parity to explain exchange rates, they need data on the prices of a basket of goods available in different countries. One analysis of this sort is conducted by *The Economist*, an international news magazine. The magazine occasionally collects data on a basket of goods consisting of "two all-beef patties, special sauce, lettuce, cheese, pickles, onions, on a sesame seed bun." It's called the "Big Mac" and is sold by McDonald's around the world.

Once we have the prices of Big Macs in two countries denominated in the local currencies, we can compute the exchange rate predicted by the theory of purchasing-power parity. The predicted exchange rate is the one that makes the cost of a Big Mac the same in the two countries. For instance, if the price of a Big Mac is $5 in the United States and 500 yen in Japan, purchasing-power parity would predict an exchange rate of 100 yen per dollar.

How well does purchasing-power parity work when applied using Big Mac prices? Here are some examples from July 2021, when the price of a Big Mac was $5.65 in the United States:

Country	Price of a Big Mac	Predicted Exchange Rate	Actual Exchange Rate
Indonesia	34,000 rupiah	6,018 rupiah/$	14,517 rupiah/$
South Korea	4,600 won	814 won/$	1,150 won/$
Japan	390 yen	69 yen/$	110 yen/$
Mexico	64 pesos	11.3 pesos/$	20.1 pesos/$
Sweden	54 krona	9.6 krona/$	8.7 krona/$
China	22.4 renminbi	4.0 renminbi/$	6.5 renminbi/$
Euro area	4.27 euros	0.76 euros/$	0.85 euros/$
Britain	3.49 pounds	0.62 pounds/$	0.73 pounds/$

You can see that the predicted and actual exchange rates are not exactly the same. After all, international arbitrage in Big Macs is not easy. Yet the two exchange rates are often in the same ballpark. The theory predicts, for example, that a dollar can buy the most rupiahs and the fewest pounds, and that is the case. Purchasing-power parity is far from a precise theory of exchange rates, but it can provide a reasonable first approximation. ●

TESTING/SHUTTERSTOCK.COM

You can find a Big Mac almost anywhere you look.

QuickQuiz

7. If a cup of coffee costs 2 euros in Paris and $6 in New York and purchasing-power parity holds, what is the exchange rate?
   a. 1/4 euro per dollar
   b. 1/3 euro per dollar
   c. 3 euros per dollar
   d. 4 euros per dollar

8. The theory of purchasing-power parity says that higher inflation in a nation causes the nation's currency to _____, leaving the _____ exchange rate unchanged.
   a. appreciate; nominal
   b. appreciate; real
   c. depreciate; nominal
   d. depreciate; real

Answers are at the end of the chapter.

32-4 Conclusion

This chapter introduced some basic concepts that macroeconomists use to study open economies. You should now understand what a nation's trade balance represents and how, in an open economy, domestic investment can differ from national saving. You should understand why a nation with a trade surplus must be sending capital abroad and why a nation with a trade deficit must be experiencing a capital inflow. You should also understand the meaning of nominal and real exchange rates as well as the implications and limitations of purchasing-power parity as a theory of exchange rates.

The macroeconomic variables introduced here provide a starting point for analyzing an open economy's interactions with the rest of the world. The next chapter develops a model to explain what determines these variables. We can then discuss how events and policies affect an economy's trade balance and the rates at which nations make exchanges in world markets.

Chapter in a Nutshell

- Net exports are the value of domestic goods and services sold abroad (exports) minus the value of foreign goods and services sold domestically (imports). Net capital outflow is the acquisition of foreign assets by domestic residents (capital outflow) minus the acquisition of domestic assets by foreigners (capital inflow). Because every export is balanced either with an import or with the acquisition of a capital asset, an economy's net exports always equal its net capital outflow.
- An economy's saving can be used either to finance investment at home or to buy assets abroad. Thus, national saving equals domestic investment plus net capital outflow.
- The nominal exchange rate is the relative price of the currency of two countries, and the real exchange rate is the relative price of the goods and services of two countries. When the nominal exchange rate changes so each dollar buys more foreign currency, the dollar is said to **appreciate** or **strengthen**. When the nominal exchange rate changes so each dollar buys less foreign currency, the dollar is said to **depreciate** or **weaken**.
- According to the theory of purchasing-power parity, a dollar (or a unit of any other currency) should be able to buy the same quantity of goods in all countries. This theory implies that the nominal exchange rate between the currencies of two countries should reflect the price levels in those countries. As a result, countries with relatively high inflation should have depreciating currencies, and countries with relatively low inflation should have appreciating currencies.

Key Concepts

closed economy, p. 666
open economy, p. 666
exports, p. 666
imports, p. 666
net exports, p. 666
trade balance, p. 666
trade surplus, p. 666
trade deficit, p. 666
balanced trade, p. 666
net capital outflow, p. 669
nominal exchange rate, p. 676
appreciation, p. 676
depreciation, p. 676
real exchange rate, p. 677
purchasing-power parity, p. 679

Questions for Review

1. Define **net exports** and **net capital outflow**. Explain how and why they are related.
2. Explain the relationship among saving, investment, and net capital outflow.
3. If a Japanese car costs 1,500,000 yen, a similar American car costs $30,000, and a dollar can buy 100 yen, what are the nominal and real exchange rates?
4. Describe the economic logic behind the theory of purchasing-power parity.
5. If the Fed started printing large quantities of U.S. dollars, what would happen to the number of Japanese yen a dollar could buy? Why?

Problems and Applications

1. How would the following transactions affect U.S. exports, imports, and net exports?
   a. An American art professor spends the summer touring museums in Europe.
   b. Students in Paris flock to see the latest movie from Hollywood.
   c. Your uncle buys a new Fiat.
   d. The student bookstore at Oxford University in England sells a copy of this textbook.
   e. A Canadian citizen shops at a store in northern Vermont to avoid Canadian sales taxes.
2. Would each of the following transactions be included in U.S. net exports or in U.S. net capital outflow? Indicate whether it would represent an increase or a decrease in that variable.
   a. An American buys a Sony TV.
   b. An American buys a share of Sony stock.
   c. The Sony pension fund buys a bond from the U.S. Treasury.
   d. A worker at a Sony plant in Japan buys some Georgia peaches from an American farmer.
3. Describe the difference between foreign direct investment and foreign portfolio investment. Who is more likely to engage in foreign direct investment—a corporation or an individual investor? Who is more likely to engage in foreign portfolio investment?
4. Explain how the following transactions would affect U.S. net capital outflow. For each transaction, state whether it represents direct investment or portfolio investment.
   a. An American cellular phone company establishes an office in the Czech Republic.
   b. Harrods of London sells stock to the General Motors pension fund.
   c. Honda expands its factory in Marysville, Ohio.
   d. A Fidelity mutual fund sells its Toyota stock to a French investor.
5. Would each of the following groups be happy or unhappy if the U.S. dollar appreciated? Explain.
   a. Dutch pension funds holding U.S. government bonds
   b. U.S. manufacturing industries
   c. Australian tourists planning a trip to the United States
   d. an American firm trying to purchase property overseas
6. What is happening to the U.S. real exchange rate in each of the following situations? Explain.
   a. The U.S. nominal exchange rate is unchanged, but prices rise faster in the United States than abroad.
   b. The U.S. nominal exchange rate is unchanged, but prices rise faster abroad than in the United States.
   c. The U.S. nominal exchange rate declines, and prices are unchanged in the United States and abroad.
   d. The U.S. nominal exchange rate declines, and prices rise faster abroad than in the United States.

7. A can of soda costs $1.25 in the United States and 25 pesos in Mexico. What is the peso–dollar exchange rate (measured in pesos per dollar) if purchasing-power parity holds? If a monetary expansion causes all prices in Mexico to double so a soda now costs 50 pesos, what happens to the peso–dollar exchange rate?
8. A case study in the chapter analyzed purchasing-power parity using the prices of Big Macs in several countries. Here are data for a few more countries:

Country	Price of a Big Mac	Predicted Exchange Rate	Actual Exchange Rate
Chile	2,990 pesos	_____ pesos/$	759 pesos/$
Hungary	900 forints	_____ forints/$	305 forints/$
Czech Republic	89 korunas	_____ korunas/$	21.8 korunas/$
Brazil	22.9 real	_____ real/$	5.25 real/$
Canada	6.77 C$	_____ C$/$	1.27 C$/$

   a. For each country, compute the predicted exchange rate in terms of the local currency per U.S. dollar. (Recall that the U.S. price of a Big Mac was $5.65.)
   b. According to purchasing-power parity, what is the predicted exchange rate between the Chilean peso and the Canadian dollar? What is the actual exchange rate?
   c. How well does the theory of purchasing-power parity explain exchange rates?
9. Purchasing-power parity holds between the nations of Ectenia and Wiknam, where the only commodity is Spam.
   a. Right now, a can of Spam costs 4 dollars in Ectenia and 24 pesos in Wiknam. What is the exchange rate between Ectenian dollars and Wiknamian pesos?
   b. Over the next 20 years, inflation is expected to be 3.5 percent per year in Ectenia and 7 percent per year in Wiknam. If this inflation comes to pass, what will the price of Spam and the exchange rate be in 20 years? (Hint: Recall the rule of 70 from Chapter 28.)
   c. Which of these two nations will likely have a higher nominal interest rate? Why?
   d. A friend of yours suggests a get-rich-quick scheme: Borrow from the nation with the lower nominal interest rate, invest in the nation with the higher nominal interest rate, and profit from the interest-rate differential. Do you see any potential problems with this idea? Explain.

Quick**Quiz** Answers

1. **a** 2. **c** 3. **d** 4. **a** 5. **d** 6. **b** 7. **b** 8. **d**

Chapter

33

A Macroeconomic Theory of the Open Economy

Over the past few decades, the United States has consistently imported more goods and services than it has exported. That is, U.S. net exports have been negative. While economists debate whether these trade deficits are a problem for the U.S. economy, the nation's business community often has a strong opinion. Many business leaders claim that the trade deficits reflect unfair competition: Foreign firms are allowed to sell their products in U.S. markets, they contend, while foreign governments impede U.S. firms from selling U.S. products abroad.

Imagine that you are the president, and you want to end these trade deficits. What should you do? Should you limit imports, perhaps by imposing a quota on European steel or a tariff on Chinese smartphones? Or should you address the nation's trade deficit in some other way?

To understand what determines a country's trade balance and how government policies can affect it, we need a macroeconomic theory that explains how an open economy works. The preceding chapter introduced the key macroeconomic variables that describe an economy's relationship with other economies, including net

BLACKCAT/E+/GETTY IMAGES; GOWITHSTOCK/SHUTTERSTOCK.COM

exports, net capital outflow, and the real and nominal exchange rates. This chapter develops a model that identifies the forces that influence these variables and explains how these variables are related to one another.

To develop this macroeconomic model of an open economy, we build on our previous analysis in two ways. First, the model takes the economy's GDP as given. We assume that the economy's output of goods and services, as measured by real GDP, is determined by the quantities of the factors of production and by the available production technology that turns these inputs into output. Second, the model takes the economy's price level as given. We assume that the price level adjusts to balance the supply and demand for money. In other words, this chapter starts with the lessons from previous chapters about the economy's output and price level and focuses on the forces that determine the economy's trade balance and exchange rate.

In one sense, the model is simple: It applies the tools of supply and demand to an open economy. Yet it is also more complex than others we have seen because it involves the simultaneous equilibrium in two related markets: the market for loanable funds and the market for foreign-currency exchange. After developing the model, we use it to examine how events and policies affect an economy's trade balance and exchange rate. We can then say which government policies are likely to reverse the trade deficits that the U.S. economy has experienced in recent years.

33-1 Supply and Demand for Loanable Funds and for Foreign-Currency Exchange

To understand the forces at work in an open economy, we focus on supply and demand in two markets. The first is the market for loanable funds, which coordinates national saving, domestic investment, and the flow of loanable funds to be invested abroad (called the net capital outflow). The second is the market for foreign-currency exchange, which coordinates people who want to trade domestic and foreign monies. This section discusses supply and demand in each of these markets separately. The next section puts them together to explain an open economy's overall equilibrium.

33-1a The Market for Loanable Funds

Our model of the open economy makes the simplifying assumption that the financial system has only one market, the **market for loanable funds**. All savers go to this market to deposit their saving, and all borrowers go there to get their loans. In this market, there is one interest rate, which is both the return to saving and the cost of borrowing.

To understand the market for loanable funds in an open economy, let's start with the identity from the preceding chapter:

$$S = I + NCO$$

$$\text{Saving} = \text{Domestic investment} + \text{Net capital outflow}.$$

When a nation saves a dollar of its income, it can use that dollar to finance the purchase of domestic capital or to finance the purchase of an asset abroad. The two sides of this identity represent the two sides of the market for loanable funds. The supply of loanable funds comes from national saving (S), and the demand for loanable funds comes from domestic investment (I) and net capital outflow (NCO).

Loanable funds in this model represent the domestically generated flow of resources available for capital accumulation. The purchase of a capital asset adds to the demand for loanable funds, regardless of whether that asset is located at home (I) or abroad (NCO). Net capital outflow can be positive or negative, so it can add to or subtract from the demand for loanable funds that arises from domestic investment. When $NCO > 0$, the country has a net outflow of capital; the net purchase of capital overseas adds to the demand for domestically generated loanable funds. When $NCO < 0$, the country has a net inflow of capital; the capital resources coming from abroad reduce the demand for domestically generated loanable funds.

Recall that the quantity of loanable funds supplied and the quantity of loanable funds demanded depend on the real interest rate. A higher real interest rate means a higher return to saving, which encourages people to save and therefore raises the quantity of loanable funds supplied. A higher interest rate also means a higher cost of borrowing to finance capital projects, which discourages investment and reduces the quantity of loanable funds demanded.

In addition to influencing national saving and domestic investment, a country's real interest rate affects its net capital outflow. To see why, consider two mutual funds—one in the United States and one in Germany—deciding whether to buy a U.S. bond or a German bond. Each mutual fund manager would make this decision in part by comparing the real interest rates in the two countries. When the U.S. real interest rate rises, the U.S. bond becomes more attractive, discouraging Americans from buying foreign assets and encouraging foreigners to buy U.S. assets. For both reasons, a rise in the U.S. real interest rate reduces U.S. net capital outflow.

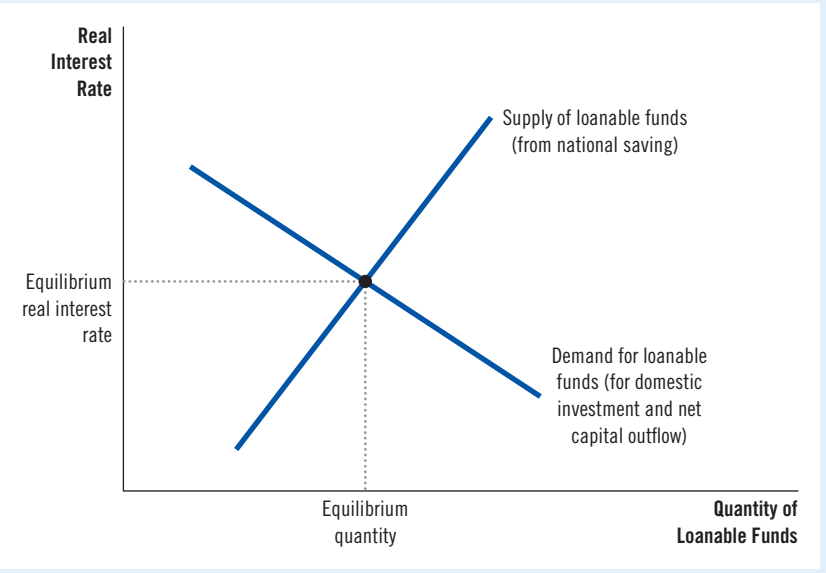
The familiar supply-and-demand diagram in Figure 1 illustrates the market for loanable funds. As in our earlier analysis of the financial system, the supply curve slopes upward because a higher interest rate increases the quantity of loanable funds supplied, and the demand curve slopes downward because a higher interest rate decreases the quantity of loanable funds demanded. Unlike the situation in our previous discussion, however, the demand side of the market now represents both domestic investment and net capital outflow. That is, in an open economy, the demand for loanable funds comes not only from those who want loanable funds to buy domestic capital goods but also from those who want loanable funds to buy foreign assets.

The interest rate adjusts to bring the supply and demand for loanable funds into balance. If the interest rate were below the equilibrium level, the quantity of loanable funds supplied would be less than the quantity demanded. The resulting shortage of loanable funds would push the interest rate upward. Conversely, if the interest rate were above the equilibrium level, the quantity of loanable funds supplied would exceed the quantity demanded. The surplus of loanable funds would drive the interest rate downward. At the equilibrium interest rate, the supply exactly balances the demand. **In other words, at the equilibrium interest rate, the amount that people want to save exactly balances the desired quantities of domestic investment and net capital outflow.**

Figure 1

The Market for Loanable Funds

The interest rate in an open economy, as in a closed economy, is determined by the supply and demand for loanable funds. National saving is the source of the supply of loanable funds. Domestic investment and net capital outflow are the sources of the demand for loanable funds. At the equilibrium interest rate, the amount that people want to save balances the amount that people want to borrow for the purpose of buying domestic capital and foreign assets.



33-1b The Market for Foreign-Currency Exchange

The second market in this model of the open economy is the market for foreign-currency exchange. Participants in the market trade U.S. dollars in exchange for foreign currencies. The analysis starts with another identity from the last chapter:

$$NCO = NX$$

$$\text{Net capital outflow} = \text{Net exports}.$$

This identity states that the imbalance between the purchase and sale of capital assets abroad (NCO) equals the imbalance between exports and imports of goods and services (NX). For example, when the U.S. economy is running a trade surplus ($NX > 0$), foreigners are buying more U.S. goods and services than Americans are buying foreign goods and services. What are Americans doing with the foreign currency they are getting from this net sale of goods and services abroad? They must be buying foreign assets, so U.S. capital is flowing abroad ($NCO > 0$). Conversely, if the United States is running a trade deficit ($NX < 0$), Americans are spending more on foreign goods and services than they are earning from selling goods and services abroad. Some of this spending must be financed by selling American assets abroad, so foreign capital is flowing into the United States ($NCO < 0$).

This model of the open economy treats the two sides of the identity as representing the two sides of the market for foreign-currency exchange. Net capital outflow represents the quantity of dollars supplied for the purpose of buying foreign assets. For example, when a U.S. mutual fund buys a Japanese bond, it must first change dollars into yen, so it supplies dollars in the market for foreign-currency exchange. Net exports represent the quantity of dollars demanded for the purpose of buying

U.S. net exports of goods and services. For example, when a Japanese airline buys a plane from Boeing, it must first change its yen into dollars, so it demands dollars in the market for foreign-currency exchange.

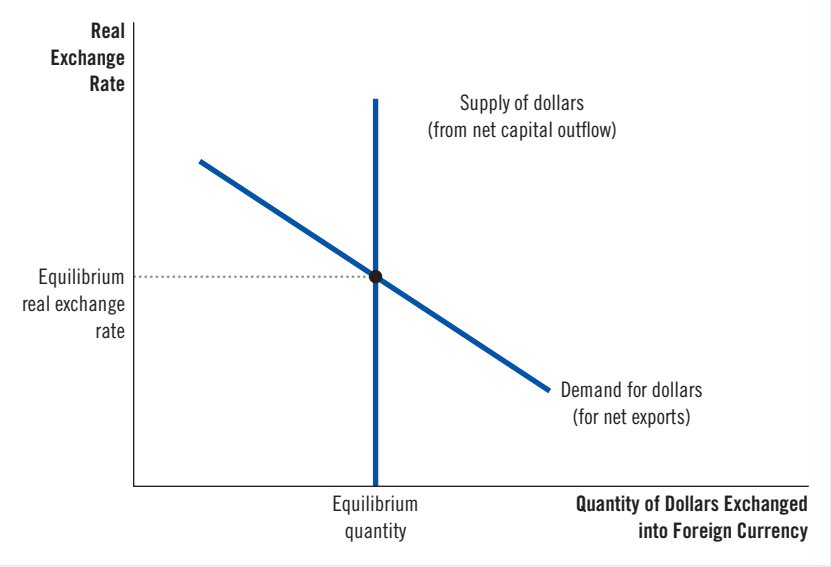
What price balances the supply and demand in the market for foreign-currency exchange? The answer is the real exchange rate. As the preceding chapter discussed, the real exchange rate is the relative price of domestic and foreign goods and a key determinant of net exports. When the U.S. real exchange rate appreciates, U.S. goods become more expensive relative to foreign goods, making U.S. goods less attractive to consumers at home and abroad. As a result, exports from the United States fall, and imports into the United States rise. Both changes reduce net exports. Hence, an appreciation of the real exchange rate reduces the quantity of dollars demanded in the market for foreign-currency exchange.

Figure 2 shows supply and demand in the market for foreign-currency exchange. The demand curve slopes downward for the reason just discussed: A higher real exchange rate makes U.S. goods more expensive and reduces the quantity of dollars demanded to buy those goods. The supply curve is vertical because the quantity of dollars supplied for net capital outflow does not depend on the real exchange rate. (As discussed earlier, net capital outflow depends on the real interest rate. When considering the market for foreign-currency exchange, the real interest rate and net capital outflow are given.)

Figure 2

The Market for Foreign-Currency Exchange

The real exchange rate is determined by the supply and demand for foreign-currency exchange. The supply of dollars to be exchanged into foreign currency comes from net capital outflow. Because net capital outflow does not depend on the real exchange rate, the supply curve is vertical. The demand for dollars comes from net exports. Because a lower real exchange rate stimulates net exports (and increases the quantity of dollars demanded to pay for these net exports), the demand curve slopes downward. At the equilibrium real exchange rate, the number of dollars supplied to buy foreign assets exactly balances the number of dollars demanded to buy net exports.



It might seem strange that, in this model, net capital outflow does not depend on the exchange rate. After all, a higher exchange value of the U.S. dollar not only makes foreign goods less expensive for American buyers but also makes foreign assets less expensive. A stronger dollar would make foreign assets more attractive, or so it might seem. But that is not necessarily the case. An American investor will eventually want to turn the foreign asset, as well as any profits earned on it, back into dollars. For example, an increase in the value of the dollar makes it less expensive for an American to buy stock in a Japanese company. But any dividends that the stock pays will be in yen. As these yen are exchanged for dollars, the higher value of the dollar means that the dividends will buy fewer dollars than before. Changes in the exchange rate influence not only the cost of buying foreign assets but also the benefit of owning them, and these two effects offset each other. For this reason, our model of the open economy posits that net capital outflow does not depend on the real exchange rate, as represented by the vertical supply curve in Figure 2.

The real exchange rate moves to ensure equilibrium in this market. It adjusts to balance the supply and demand for dollars just as the price of any good adjusts to balance supply and demand for it. If the real exchange rate were below the equilibrium level, the quantity of dollars supplied would be less than the quantity demanded. The resulting shortage of dollars would push the value of the dollar upward. Conversely, if the real exchange rate were above the equilibrium level, the quantity of dollars supplied would exceed the quantity demanded. The surplus of dollars would drive the value of the dollar downward. **At the equilibrium real exchange rate, the demand for dollars by foreigners arising from the U.S. net exports of goods and services balances the supply of dollars from Americans arising from U.S. net capital outflow.**

Purchasing-Power Parity as a Special Case

An alert reader of this book might ask: Why are we developing a theory of the exchange rate here? Didn't we already do that in the preceding chapter?

Well, yes, we discussed a theory of the exchange rate called **purchasing-power parity**. It asserts that, because of international arbitrage, a dollar (or any other currency) must buy the same quantity of goods and services in every country. As a result, the real exchange rate is fixed, and all changes in the nominal exchange rate between two currencies reflect changes in the price levels in the two countries.

The model of the exchange rate developed here is related to the theory of purchasing-power parity. According to purchasing-power parity, international trade responds quickly to international price differences. If goods were cheaper in one country than in another, they would be exported from the first country and imported into the second until the price difference disappeared. In other words, the theory of purchasing-power parity assumes that net exports are highly responsive to small changes in the real exchange rate. If net exports were, in fact, so responsive, the demand curve in Figure 2 would be horizontal.

Thus, the theory of purchasing-power parity can be viewed as a special case of the model considered here. In that special case, the demand curve for foreign-currency exchange, instead of sloping downward, is horizontal at the level of the real exchange rate that ensures parity of purchasing power at home and abroad.

This special case is a good starting point for explaining exchange rates, but it's not the last word. In practice, because of transport costs and the imperfect substitutability of foreign and domestic goods, purchasing-power parity sometimes fails to hold. This chapter, therefore, assumes that the demand curve for foreign-currency exchange slopes downward. This assumption allows for the possibility that the real exchange rate changes over time as, in fact, it does in the real world. ■

QuickQuiz

1. Holding other things constant, an increase in a nation's interest rate reduces
   a. national saving and domestic investment.
   b. national saving and the net capital outflow.
   c. domestic investment and the net capital outflow.
   d. national saving only.

2. Holding other things constant, an appreciation of a nation's currency causes
   a. exports to rise and imports to fall.
   b. exports to fall and imports to rise.
   c. both exports and imports to rise.
   d. both exports and imports to fall.

Answers are at the end of the chapter.

33-2 Equilibrium in the Open Economy

So far, we have discussed supply and demand in the market for loanable funds and in the market for foreign-currency exchange. Let's now consider how these markets are connected.

33-2a Net Capital Outflow: The Link between the Two Markets

We begin by recapping what we've learned so far in this chapter. We have been discussing how the economy coordinates four important macroeconomic variables: national saving (S), domestic investment (I), net capital outflow (NCO), and net exports (NX). These are related through the following identities:

$$S = I + NCO$$

and

$$NCO = NX.$$

In the market for loanable funds, supply comes from national saving (S), demand comes from domestic investment (I) and net capital outflow (NCO), and the real interest rate balances supply and demand. In the market for foreign-currency exchange, supply comes from net capital outflow (NCO), demand comes from net exports (NX), and the real exchange rate balances supply and demand.

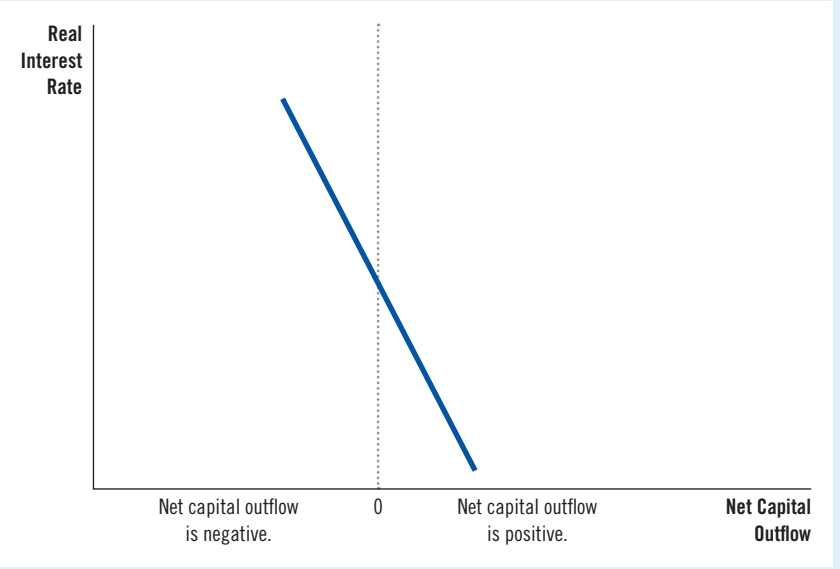
Net capital outflow is the variable that links the two markets. In the market for loanable funds, net capital outflow is part of demand. An American who wants to buy an asset abroad must finance this purchase by obtaining resources in the U.S. market for loanable funds. In the market for foreign-currency exchange, net capital outflow is the source of supply. An American who wants to buy an asset in another country must supply dollars to exchange them for the currency of that country.

The key determinant of net capital outflow, as we have discussed, is the real interest rate. An increase in the U.S. interest rate makes U.S. assets more attractive and lowers the U.S. net capital outflow. Figure 3 shows this negative relationship between the interest rate and net capital outflow. This net-capital-outflow curve is the link between the market for loanable funds and the market for foreign-currency exchange.

Figure 3

How Net Capital Outflow Depends on the Interest Rate

Because a higher domestic real interest rate makes domestic assets more attractive, it reduces net capital outflow. Note the position of zero on the horizontal axis: Net capital outflow can be positive or negative. A negative value of net capital outflow means that the economy has a net inflow of capital.



33-2b Simultaneous Equilibrium in Two Markets

Figure 4 puts all the pieces of our model together. It shows how the market for loanable funds and the market for foreign-currency exchange jointly determine the important macroeconomic variables of an open economy.

Panel (a) of the figure shows the market for loanable funds (taken from Figure 1). As before, national saving is the source of the supply of loanable funds. Domestic investment and net capital outflow are the sources of the demand for loanable funds. The equilibrium real interest rate (r_1) brings the quantity of loanable funds supplied and demanded into balance.

Panel (b) of the figure shows net capital outflow (taken from Figure 3). The interest rate comes from panel (a) and determines net capital outflow. A higher interest rate at home makes domestic assets more attractive, reducing net capital outflow. Therefore, the net-capital-outflow curve in panel (b) slopes downward.

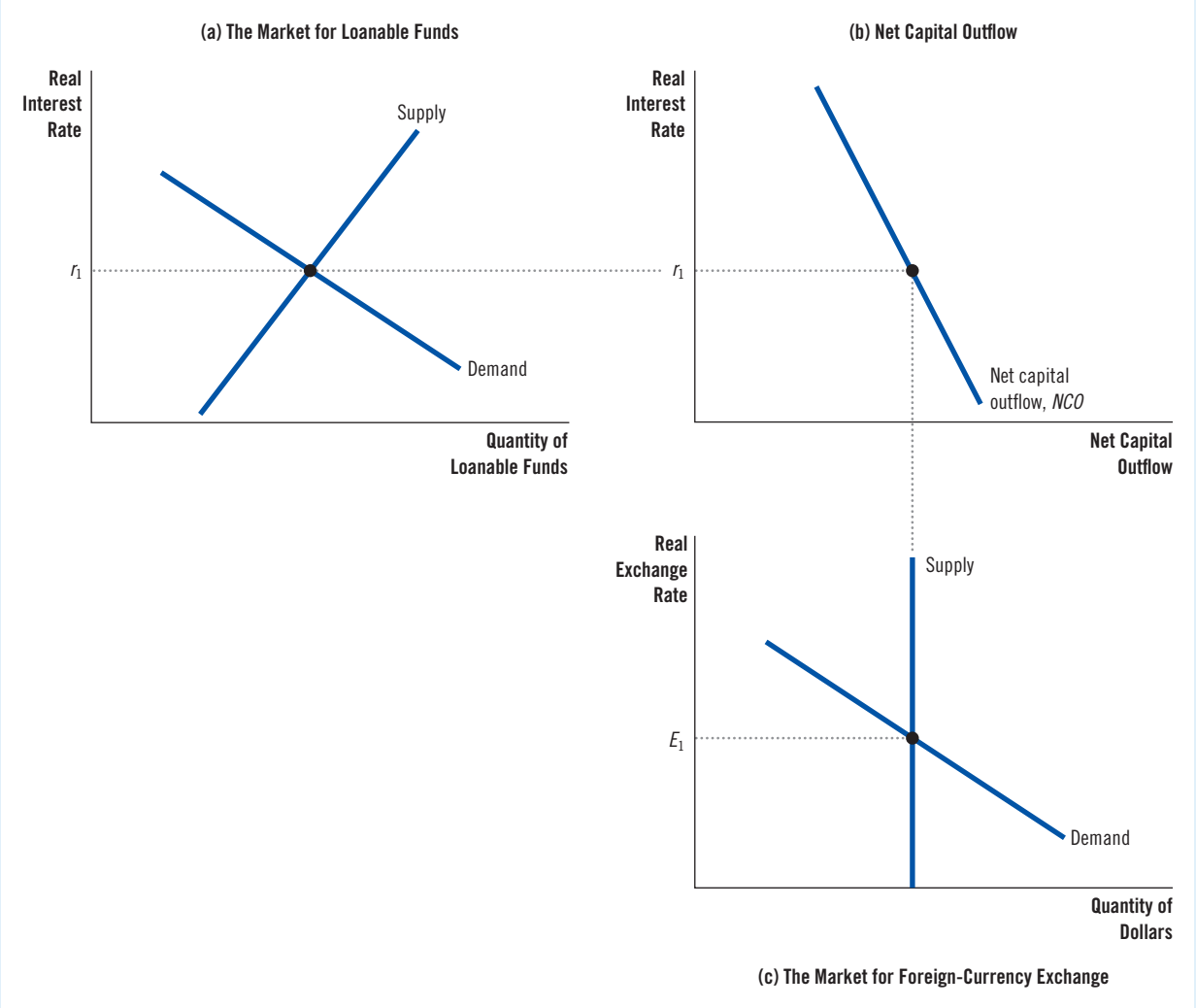
Panel (c) of the figure shows the market for foreign-currency exchange (taken from Figure 2). Because foreign assets must be purchased with foreign currency, the quantity of net capital outflow from panel (b) determines the supply of dollars to be exchanged into foreign currencies. The real exchange rate does not affect net capital outflow, so the supply curve is vertical. The demand for dollars comes from net exports. Because a depreciation of the real exchange rate increases net exports, the demand curve for foreign-currency exchange slopes downward. The equilibrium real exchange rate (E_1) brings into balance the quantity of dollars supplied and the quantity of dollars demanded in the market for foreign-currency exchange.

The two markets shown in Figure 4 determine two relative prices: the real interest rate and the real exchange rate. The real interest rate determined in panel (a) is the price of goods and services in the present relative to goods and

Figure 4

The Real Equilibrium in an Open Economy

In panel (a), the supply and demand for loanable funds determine the real interest rate. In panel (b), the interest rate determines net capital outflow, which provides the supply of dollars in the market for foreign-currency exchange. In panel (c), the supply and demand for dollars in the market for foreign-currency exchange determine the real exchange rate.



services in the future. The real exchange rate determined in panel (c) is the price of domestic goods and services relative to foreign goods and services. These two relative prices adjust simultaneously to balance supply and demand in these two markets. As they do so, they determine national saving, domestic investment, net capital outflow, and net exports. We can use this model to see how all these variables change when some policy or event causes one of these curves to shift.

FYI

Disentangling Supply and Demand

Suppose the owners of an apple orchard decide to consume some of their own apples. Does this decision represent an increase in the demand for apples or a decrease in the supply? Either answer is defensible, and if we are careful in our subsequent analysis, nothing important will hinge on the answer we choose. Sometimes how we divide things between supply and demand is a bit arbitrary.

In the macroeconomic model of the open economy developed in this chapter, the division of transactions between "supply" and "demand" is also a bit arbitrary—both in the market for loanable funds and in the market for foreign-currency exchange.

Consider first the market for loanable funds. The model treats the net capital outflow as part of the demand for loanable funds. Yet instead of writing $S = I + NCO$, we could just as easily have written $S - NCO = I$. When the equation is rewritten in this way, a capital outflow looks like a reduction in the supply of loanable funds. Either way works. The first interpretation ($S = I + NCO$) emphasizes loanable funds generated domestically, whether used at home or abroad. The second interpretation ($S - NCO = I$) emphasizes loanable funds available for domestic investment, whether generated at home or abroad. The difference is more semantic than substantive.

Similarly, consider the market for foreign-currency exchange. In our model, net exports are the source of the demand for dollars, and net capital outflow is the source of the supply. When a U.S. resident imports a car made in Japan, our model treats that transaction as a decrease in the quantity of dollars demanded (because net exports fall) rather than an increase in the quantity of dollars supplied. Similarly, when a Japanese citizen buys a U.S. government bond, our model treats that transaction as a decrease in the quantity of dollars supplied (because net capital outflow falls) rather than an increase in the quantity of dollars demanded. This definition of terms may seem somewhat unnatural at first, but it will prove useful when analyzing the effects of various policies. ■

QuickQuiz

3. In the model just developed, two markets determine two prices, which are
   a. the nominal exchange rate and the nominal interest rate.
   b. the nominal exchange rate and the real interest rate.
   c. the real exchange rate and the nominal interest rate.
   d. the real exchange rate and the real interest rate.

4. Other things being equal, an increase in the U.S. net capital outflow _____ the demand for loanable funds and _____ the supply of dollars in the market for foreign currency exchange.
   a. increases; increases
   b. increases; decreases
   c. decreases; increases
   d. decreases; decreases

Answers are at the end of the chapter.

33-3 How Policies and Events Affect an Open Economy

Let's now use this model of the open economy to analyze how changes in policy and other events alter the economy's equilibrium. Keep in mind that this model is just supply and demand in two markets: the market for loanable funds and the market for foreign-currency exchange. When using the model to analyze any event, apply the three steps introduced in Chapter 4. First, determine which of the supply and demand curves the event affects. Second, determine the direction in which the curves shift. Third, use the supply-and-demand diagrams to compare the old equilibrium with the new one.

33-3a Government Budget Deficits

Earlier in this book, when we applied the loanable-funds model to a closed economy, we examined the effects of government budget deficits, which occur when government spending exceeds government revenue. Because a government budget

deficit represents **negative** public saving, it reduces national saving (the sum of public and private saving). We concluded that a government budget deficit reduces the supply of loanable funds, drives up the interest rate, and crowds out investment.

Now let's consider the effects of a budget deficit in an open economy. First, which curve in our model shifts? As in a closed economy, the initial impact of the budget deficit is on national saving and, therefore, on the supply curve for loanable funds. Second, in which direction does this supply curve shift? Again, as in a closed economy, a budget deficit represents **negative** public saving, so it reduces national saving and shifts the supply curve for loanable funds to the left. This result is shown as the shift from S_1 to S_2 in panel (a) of Figure 5.

Finally, compare the old and new equilibria. Panel (a) shows the effect of a budget deficit on the market for loanable funds. Because fewer funds are available for borrowers in financial markets, the interest rate rises from r_1 to r_2 to balance supply and demand. Faced with a higher interest rate, borrowers choose to borrow less. In other words, the quantity of loanable funds demanded falls. This change is represented in the figure as the movement along the demand curve from point A to point B. With reduced borrowing comes a decline in the purchase of capital goods by households and firms. As in a closed economy, the budget deficit crowds out domestic investment.

In an open economy, however, the reduced supply of loanable funds has additional effects. Panel (b) shows that the increase in the interest rate from r_1 to r_2 reduces net capital outflow. [This fall in net capital outflow is also part of the decrease in the quantity of loanable funds demanded in the movement from point A to point B in panel (a).] Because savings kept at home now earn higher rates of return, investing abroad is less attractive, and domestic residents buy fewer foreign assets. Higher interest rates also attract foreign investors who want to earn the higher returns on U.S. assets. Thus, when the budget deficit increases the interest rate, both domestic and foreign behavior cause net capital outflow to decline.

Panel (c) shows how the budget deficit affects the market for foreign-currency exchange. Because net capital outflow is reduced, domestic residents need less foreign currency to buy foreign assets and, therefore, supply fewer dollars in the market for foreign-currency exchange. The supply curve for dollars shifts leftward from S_1 to S_2. The reduced supply of dollars causes the real exchange rate to appreciate from E_1 to E_2. That is, the dollar becomes more valuable relative to foreign currencies. This appreciation, in turn, makes domestic goods more expensive compared with foreign goods. People both at home and abroad respond to this change in relative prices, so exports fall and imports rise. For both reasons, net exports fall.

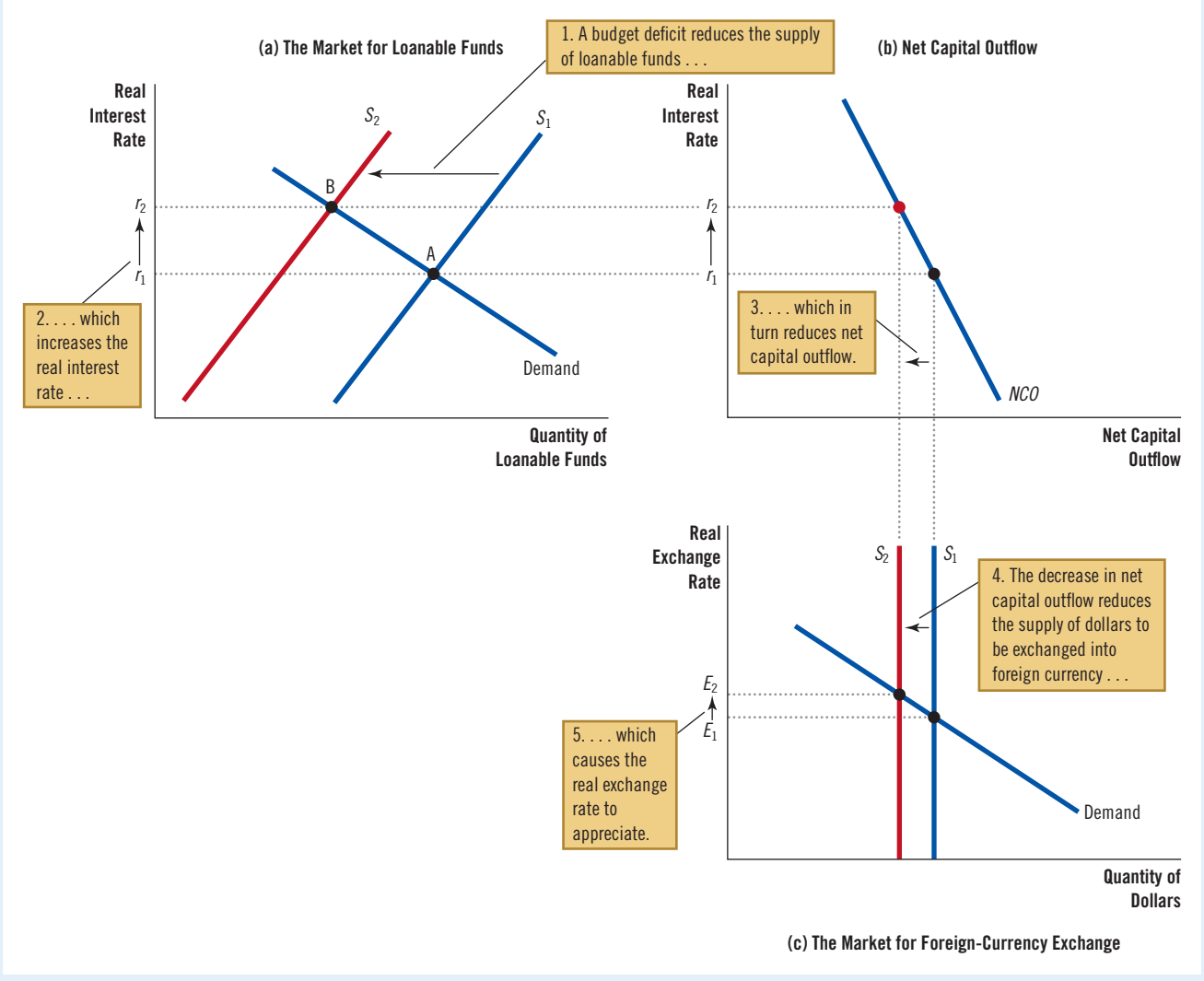
To sum up: **In an open economy, government budget deficits raise real interest rates, crowd out domestic investment, cause the currency to appreciate, and push the trade balance toward deficit.**

An example of this lesson occurred in the United States in the 1980s. Shortly after Ronald Reagan was elected president in 1980, the fiscal policy of the U.S. federal government radically changed. The president and Congress enacted large tax cuts but did not reduce government spending by nearly as much. The result was a large budget deficit. Our model of the open economy predicts that such a policy should have led to a trade deficit, and, in fact, it did, as we saw in a case study in the preceding chapter. Because the budget deficit and trade deficit during this period were so closely related in both theory and practice, they were nicknamed the **twin deficits**. It would be a mistake to view these twins as identical, however, because many factors beyond fiscal policy can influence the trade deficit.

Figure 5

The Effects of a Government Budget Deficit

When the government runs a budget deficit, it reduces the supply of loanable funds from S_1 to S_2 in panel (a). The interest rate rises from r_1 to r_2 to balance the supply and demand for loanable funds. In panel (b), the higher interest rate reduces net capital outflow. That reduced net capital outflow, in turn, reduces the supply of dollars in the market for foreign-currency exchange from S_1 to S_2 in panel (c). The fall in the supply of dollars causes the real exchange rate to appreciate from E_1 to E_2. The appreciation of the exchange rate pushes the trade balance toward deficit.



trade policy
government policy that directly influences the quantity of goods and services that a country imports or exports

33-3b Trade Policy

Trade policy is government policy that directly influences the quantity of goods and services that a country imports or exports. Trade policy takes various forms, usually with the purpose of supporting a particular domestic industry. One common trade policy is a **tariff**, a tax on imported goods. Another is an **import quota**, a limit on the quantity of a good produced abroad that can be sold domestically.

Consider the macroeconomic impact of trade policy. Suppose that the U.S. steel industry, concerned about competition from European producers, convinces the U.S. government to impose a quota on the amount of steel that can be imported from Europe. In making their case, lobbyists for the steel industry assert that the trade restriction would shrink the size of the U.S. trade deficit. Are they right? Our model, illustrated in Figure 6, offers an answer.

The first step in analyzing the trade policy is to determine which curve shifts. The initial impact of the import restriction is, not surprisingly, on imports. Because net exports equal exports minus imports, the policy also affects net exports. And because net exports are the source of demand for dollars in the market for foreign-currency exchange, the policy affects the demand curve in this market.

The second step is to determine the direction in which this demand curve shifts. Because the quota restricts the amount of European steel sold in the United States, it reduces imports at any given real exchange rate. Net exports, which equal exports minus imports, will therefore **rise** for any given real exchange rate. Because foreigners need dollars to buy U.S. net exports, the rise in net exports increases the demand for dollars in the market for foreign-currency exchange. This increase in the demand for dollars is shown in panel (c) of Figure 6 as the shift from D_1 to D_2.

The third step is to compare the old and new equilibria. As we can see in panel (c), the increase in the demand for dollars causes the real exchange rate to appreciate from E_1 to E_2. Because nothing has happened in the market for loanable funds in panel (a), the real interest rate remains the same. Because there is no change in the real interest rate, there is also no change in net capital outflow, shown in panel (b). And because there is no change in net capital outflow, there can be no change in net exports, even though the import quota has reduced imports.

It might seem puzzling that net exports stay the same while imports fall. This puzzle is resolved by noting the change in the real exchange rate: When the dollar appreciates in the market for foreign-currency exchange, domestic goods become more expensive compared with foreign goods. This appreciation encourages imports and discourages exports, and both changes offset the direct increase in net exports due to the import quota. In the end, an import quota reduces both imports and exports, but net exports (exports minus imports) are unchanged.

We have thus arrived at a surprising result: **Trade policies do not affect the trade balance.** That is, policies that directly influence exports or imports do not alter net exports. This conclusion seems less surprising if one recalls the accounting identity:

$$NX = NCO = S - I.$$

Net exports equal net capital outflow, which equals national saving minus domestic investment. Trade policies do not alter the trade balance because they do not alter national saving or domestic investment. For given levels of national saving and domestic investment, the real exchange rate adjusts to keep the trade balance the same, regardless of the trade policies the government puts in place.

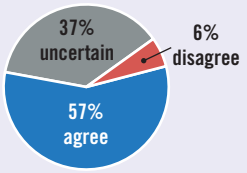
Although trade policies do not affect a country's overall trade balance, they affect specific firms, industries, and countries. When the U.S. government imposes an

Ask the Experts

Deficits

"If the United States reduced its fiscal deficit, then its trade deficit would also shrink."

What do economists say?

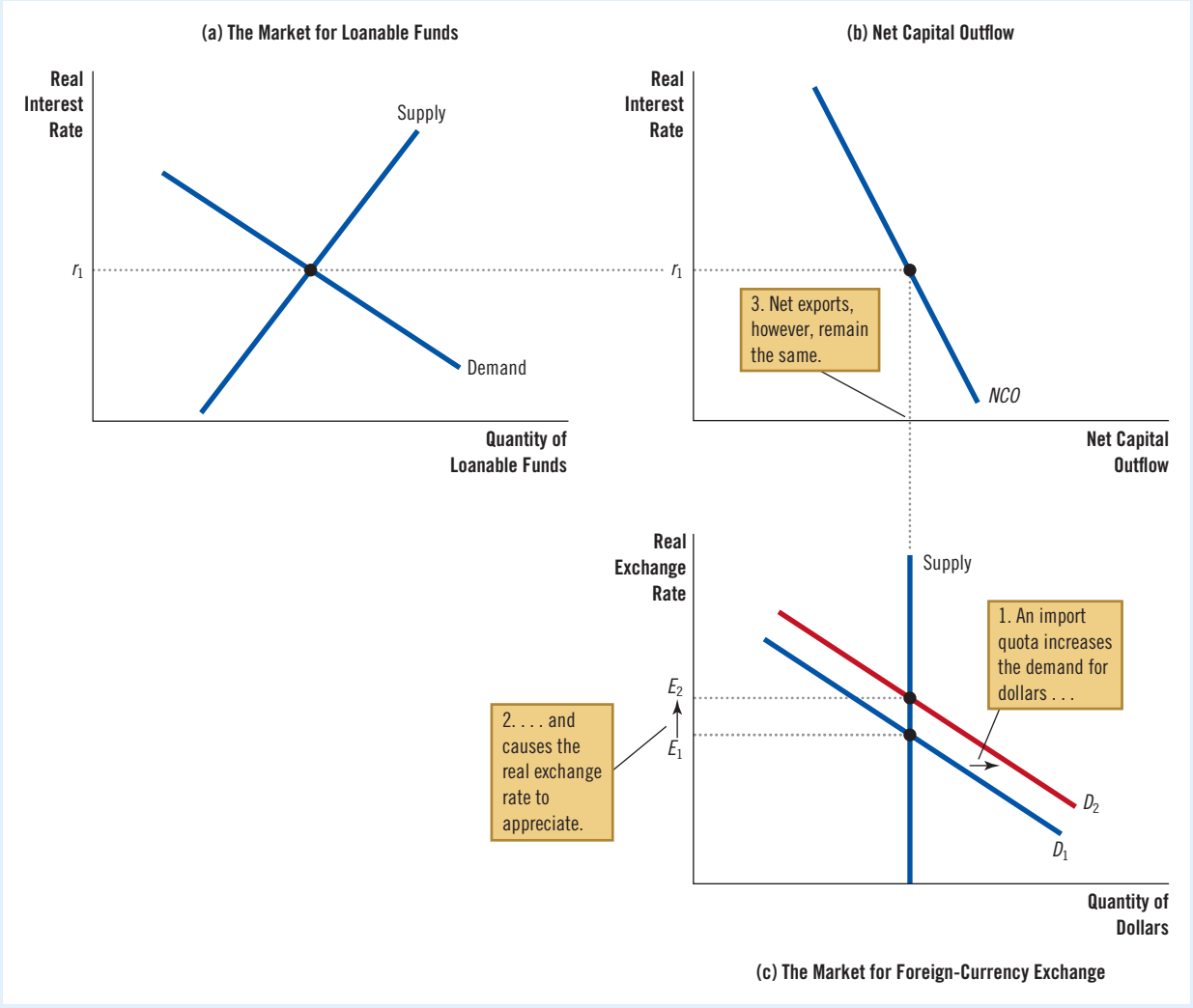


Source: IGM Economic Experts Panel, June 21, 2017.

Figure 6

The Effects of an Import Quota

When the U.S. government imposes a quota on the import of European steel, nothing happens in the market for loanable funds in panel (a) or to net capital outflow in panel (b). The only effect is a rise in net exports (exports minus imports) for any given real exchange rate. As a result, the demand for dollars in the market for foreign-currency exchange rises, as shown by the shift from D_1 to D_2 in panel (c). This increase in the demand for dollars causes the value of the dollar to appreciate from E_1 to E_2. This appreciation of the dollar tends to reduce net exports, offsetting the direct effect of the import quota on the trade balance.



import quota on European steel, the U.S. Steel Corporation faces less competition from abroad and will likely sell more steel and make greater profits. At the same time, because the dollar has appreciated, U.S. farmers find their products more expensive on world markets and will sell less of them to, say, China. In this case, the import quota on European steel increases net exports of steel and decreases net exports of agricultural products. In addition, it pushes the trade balance between

the United States and Europe toward surplus, and it pushes the trade balance between the United States and China toward deficit. The overall trade balance of the U.S. economy, however, stays the same.

The effects of trade policies are, therefore, more microeconomic than macroeconomic. Although advocates of trade policies sometimes suggest (contrary to what our model predicts) that these policies can alter a country's trade balance, they are usually more motivated by concerns about specific firms or industries. One should not be surprised, for instance, to hear an executive from U.S. Steel advocating import quotas on steel from Europe. But economists usually oppose such trade policies. Free trade may hurt some workers and firms, but it allows economies to specialize in doing what they do best, increasing prosperity in all countries. Trade restrictions interfere with these gains from trade.

33-3c Political Instability and Capital Flight

In 1994, political instability in Mexico, including the assassination of a prominent political leader, made world financial markets nervous. The nation suddenly seemed like a riskier place to invest. People decided to pull some of their assets out of Mexico and move these funds to the United States and other "safe havens." Such a large and sudden movement of funds out of a country is called **capital flight**. To see the implications of capital flight for the Mexican economy, let's follow our three steps for analyzing a change in equilibrium. But this time, we apply our model from the Mexican perspective.

capital flight
a large and sudden reduction in the demand for assets located in a country

Consider first which curves are affected by capital flight. When investors around the world observe political problems in Mexico, they decide to sell some of their Mexican assets and use the proceeds to buy U.S. assets. This increases Mexican net capital outflow and, therefore, affects both markets in our model. Most obviously, it affects the net-capital-outflow curve, and this change in net capital outflow, in turn, influences the supply of pesos in the market for foreign-currency exchange. In addition, because the demand for loanable funds comes from both domestic investment and net capital outflow, capital flight affects the demand curve in the Mexican market for loanable funds.

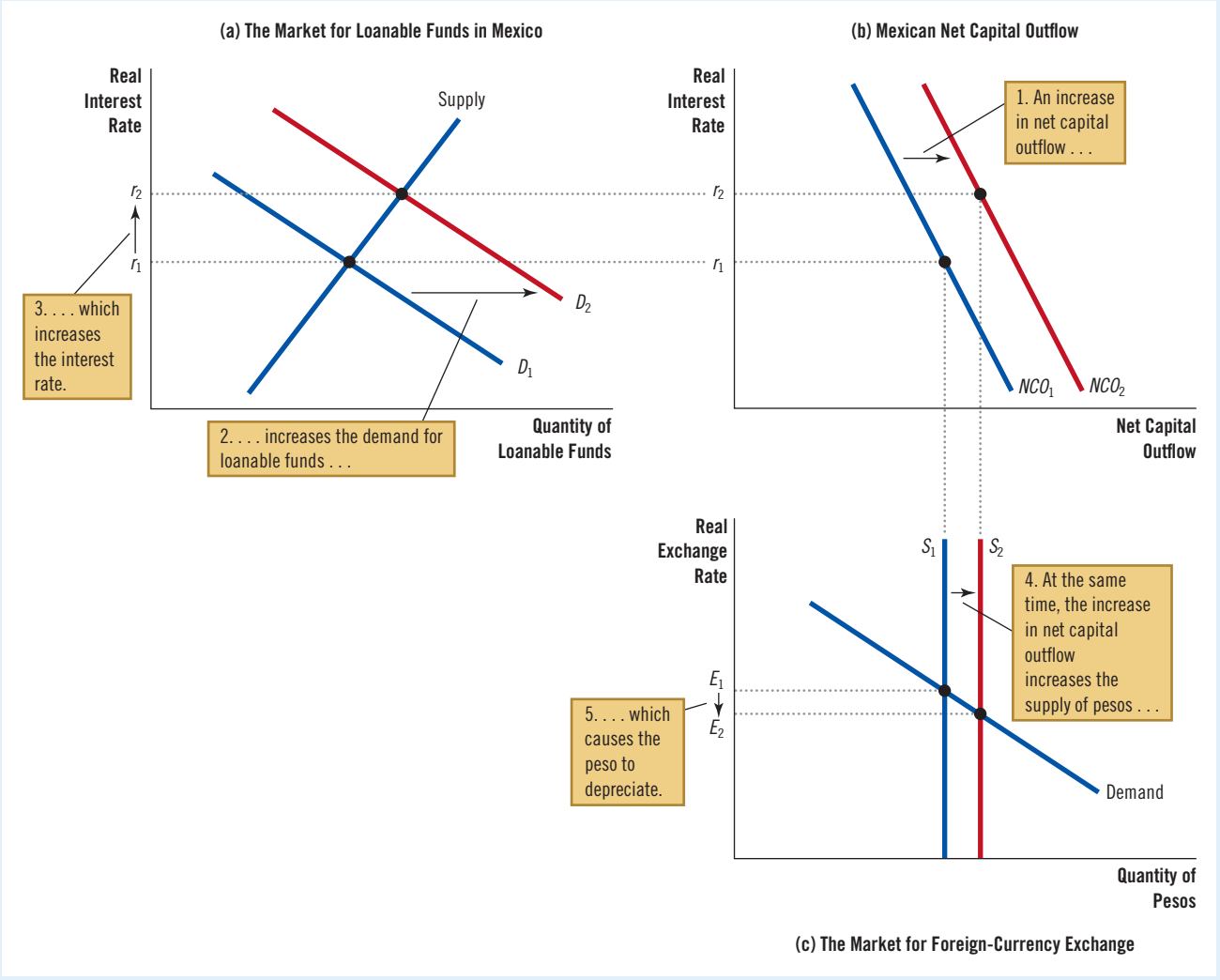
Now consider the direction in which these curves shift. When net capital outflow increases, there is greater demand for loanable funds to finance these purchases of capital assets abroad. As panel (a) of Figure 7 shows, the demand curve for loanable funds shifts to the right from D_1 to D_2. In addition, because net capital outflow is higher for any interest rate, the net-capital-outflow curve also shifts to the right from NCO_1 to NCO_2, as in panel (b).

To see the effects of capital flight on the Mexican economy, compare the old and new equilibria. Panel (a) of Figure 7 shows that the increased demand for loanable funds causes the interest rate in Mexico to rise from r_1 to r_2. Panel (b) shows that Mexican net capital outflow increases. (The rise in the interest rate makes Mexican assets more attractive, but this change only partly offsets the impact of capital flight on net capital outflow.) Panel (c) shows that the increase in net capital outflow raises the supply of pesos in the market for foreign-currency exchange from S_1 to S_2. That is, as people try to get out of Mexican assets, there is a large supply of pesos to be converted into dollars. This increase in supply causes the peso to depreciate from E_1 to E_2. **Thus, capital flight from Mexico increases Mexican interest rates and decreases the value of the Mexican peso in the market for foreign-currency exchange.** This is exactly what happened in 1994. From November 1994 to March 1995, the interest rate on short-term Mexican government bonds rose from 14 percent to 70 percent, and the peso depreciated from 29 to 15 U.S. cents per peso.

Figure 7

The Effects of Capital Flight

If people decide that Mexico is a risky place to keep their savings, they will move their funds to safe havens such as the United States, resulting in an increase in Mexican net capital outflow. Consequently, the demand for loanable funds in Mexico rises from D_1 to D_2, as shown in panel (a), driving up the Mexican real interest rate from r_1 to r_2. Because net capital outflow is higher for any interest rate, that curve also shifts to the right from NCO_1 to NCO_2 in panel (b). At the same time, in the market for foreign-currency exchange, the supply of pesos rises from S_1 to S_2, as shown in panel (c). This increase in the supply of pesos causes the peso to depreciate from E_1 to E_2, so the peso becomes less valuable relative to other currencies.



The price changes that result from capital flight influence key macroeconomic quantities. The depreciation of the currency makes exports cheaper and imports more expensive, pushing the trade balance toward surplus. At the same time, the increase in the interest rate reduces domestic investment, slowing capital accumulation and economic growth.

Capital flight has its largest impact on the country from which capital is fleeing, but it also affects other countries. When capital flows out of Mexico into the United States, for instance, it has the opposite effect on the U.S. economy as it has on the Mexican economy. In particular, the rise in Mexican net capital outflow coincides with a fall in U.S. net capital outflow. As the peso depreciates and Mexican interest rates rise, the dollar appreciates, and U.S. interest rates fall. The impact on the U.S. economy is small, however, because the economy of the United States is much larger than that of Mexico.

The events that we have been describing in Mexico could happen to any economy, and, in fact, they do from time to time. In 1997, the world learned that the banking systems of several Asian economies, including Thailand, South Korea, and Indonesia, were at or near the point of bankruptcy, and this news induced capital to flee from these nations. In 1998, the Russian government defaulted on its debt, prompting international investors to take whatever money they could and run. A similar (but more complicated) set of events unfolded in Argentina in 2002 and again in 2019 and 2020. In each of these cases of capital flight, the results were much as our model predicts: rising interest rates and a depreciating currency.

Capital Flows from China

According to our analysis of capital flight, a nation that experiences an outflow of capital sees its currency weaken in foreign exchange markets, and this depreciation, in turn, increases its net exports. The country into which the capital is flowing sees its currency strengthen, and this appreciation pushes its trade balance toward deficit.

With these lessons in mind, consider this question: Suppose a nation's government, as a matter of policy, encourages capital to flow to another country, perhaps by making foreign investments itself. What effects would this policy have? The answer is much the same: Other things being equal, it leads to a weaker currency and a trade surplus for the nation encouraging the capital outflows and a stronger currency and a trade deficit for the recipient of those capital flows.

This analysis sheds light on a long-running policy dispute between the United States and China. The Chinese government has, at times, tried to depress the value of its currency, the renminbi, in foreign exchange markets to promote its export industries. It did this by accumulating foreign assets, including substantial amounts of U.S. government bonds. From 2000 to 2014, China's total reserves of foreign assets rose from $160 billion to about $4 trillion.

The U.S. government often objected to China's interventions in foreign exchange markets. By holding down the value of the renminbi, the policy made Chinese goods less expensive compared with American goods, pushing the U.S. trade balance toward deficit and hurting American producers who competed with imports from China. Because of these effects, the U.S. government tried to get China to stop influencing the exchange value of its currency using government-sponsored capital flows. Some members of Congress advocated tariffs on Chinese imports unless China ceased its "currency manipulation." And during the Trump administration, the government enacted tariffs, sometimes citing this argument.

Yet the impact of the Chinese policy on the U.S. economy was not all bad. American consumers of Chinese imports benefited from lower prices.

In the News

Separating Fact from Fiction

Politicians on both the right and left often hold mistaken views about the role of international trade in a nation's economic well-being.

Five Big Truths about Trade

By Alan S. Blinder

International trade is, once again, a hot-button political issue, making this an unpropitious time for rational discourse about the subject. Nonetheless, here are five issues on which the overwhelming majority of economists, liberal and conservative, agree.

1. Most job losses are not due to international trade. Every month roughly five million new jobs are created in the U.S. and almost that many are destroyed, leaving a small net increment. International trade accounts for only a minor share of that staggering job churn. Vastly more derives from the hurly-burly of competition and from technological change, which literally creates and destroys entire industries. Competition and technology are widely and correctly applauded—international trade is not so fortunate.

2. Trade is more about efficiency—and hence wages—than about the number of jobs. You probably don't sew your own clothes or grow your own food. Instead, you buy these things from others, using the wages you earn doing something you do better. Imagine how much lower your standard of living would be if you had to sew your own clothes, grow your own food . . . and a thousand other things.

The case for international trade is no different. It's not mainly about creating or destroying jobs. It's about using labor more efficiently, which is one key to higher wages.

But there is a catch: Whenever trade patterns change, some people will gain (either jobs or wages) but others will lose. The federal government could and should help them more, but it doesn't. So Americans who do lose their jobs due to international trade have a legitimate gripe.

3. Bilateral trade imbalances are inevitable and mostly uninteresting. Each month I run a trade deficit with Public Service Electric & Gas. They sell me gas and electricity; I sell them nothing. But I run a bilateral trade surplus with Princeton University, to which I sell teaching services but from which I buy little. Should I seek balanced trade with PSE&G or Princeton? Of course not. Neither should countries.

4. Running an overall trade deficit does not make us "losers." The U.S. multilateral trade balance—its balance with all of its trading partners—has been in deficit for decades. Does that mean that our country is in some sort of trouble? Probably not. For example, people who claim that our trade deficit kills jobs need to explain how the U.S. managed to achieve 4% unemployment in 2000, when our trade deficit was larger, as a share of GDP, than it is today.

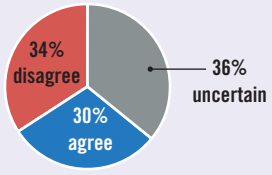
A trade deficit means that foreigners send us more goods and services than we send them. To balance the books, they get our IOUs, which

Ask the Experts

Currency Manipulation

"Economic analysis can identify whether countries are using their exchange rates to benefit their own people at the expense of their trading partners' welfare."

What do economists say?



Source: IGM Economic Experts Panel, June 16, 2015.

In addition, the inflow of capital from China reduced U.S. interest rates, increasing investment in the U.S. economy. To some extent, the Chinese government was financing U.S. economic growth. The Chinese policy of investing in the U.S. economy created winners and losers among Americans. All things considered, the net impact on the U.S. economy was probably small.

The harder question concerns the motives behind the policy: Why were Chinese leaders during these years interested in producing for export and investing abroad rather than producing for domestic consumption and investment? One possibility is that China wanted to accumulate a reserve of foreign assets on which it could draw in emergencies—a kind of national "rainy-day fund." In any case, after 2014, as growth in the Chinese economy slowed, the Chinese government started to spend some of the fund. From 2014 to 2021, its reserves of foreign assets fell by almost $1 trillion. ●

means they wind up holding paper—U.S. Treasury bills, corporate bonds or other private debt instruments. That doesn't sound so terrible for us, does it?

One exceptional country—the U.S.—is the source of the world's major international reserve currency, the U.S. dollar. Since ever-expanding world commerce requires ever more dollars, the U.S. must run trade deficits regularly. That's sometimes called our "exorbitant privilege," since we get to import more than we export.

5. Trade agreements barely affect a nation's trade balance. Much of the political angst is directed not at trade in general, but at specific international trade agreements. The North American Free Trade Agreement allegedly shipped U.S. jobs to Mexico. . . .

There is a grain of truth here. Some U.S. jobs were indeed destroyed when NAFTA liberalized trade with Mexico—and those people deserved better treatment from the government than they got. But NAFTA also created a number of new jobs in the U.S. (See No. 2.)

But there's more. "Trade" and "trade agreements" are not synonyms. We traded with Mexico long before NAFTA, and that trade was growing. Our trade with China has burgeoned in recent decades without a succession of trade agreements.

Most fundamentally, but least understood, a nation's overall trade balance is determined by its domestic decisions, not by trade deals. Think about the accounting involved here.

As noted above, borrowing from abroad is the bookkeeping counterpart of running a trade deficit. One implies the other. The amount we borrow from abroad must equal the gap between our total spending as a nation (including government spending) and our total income (including the government's income from taxation). Spendthrift nations like the U.S. have trade deficits because we don't save much. But these saving decisions are domestic; they do not derive from trade agreements.

America's chronic trade deficits stem from the dollar's international role and from Americans' decisions not to save much, not from trade deals. Trade deficits are not a major cause of either job losses or job gains. But some people do lose their jobs from shifting trade patterns; and the government should do more to help them. Importantly, trade makes American workers more productive and, presumably, better paid.

Now, would someone please tell this to Bernie Sanders and Donald Trump? ■

Source: *The Wall Street Journal,* April 22, 2016.

Questions to Discuss

1. Do you think running a trade deficit necessarily puts a nation at a disadvantage? Why or why not? According to the dictionary, the word **deficit** means "an excess of expenditure over revenue," but another definition is "deficiency or disadvantage." Might this dual meaning mislead pundits and policymakers into being more worried about trade deficits than Professor Blinder is?
2. How do you think the government should help workers who lose their jobs because of changing patterns of trade? Should these workers receive government assistance different from that given to workers who lose jobs for other reasons, such as changing technology?

Mr. Blinder is a professor of economics at Princeton University.

QuickQuiz

5. The government in an open economy cuts spending to reduce the budget deficit. As a result, the interest rate _____, leading to a capital _____ and a currency _____.
   a. falls; outflow; appreciation
   b. falls; outflow; depreciation
   c. falls; inflow; appreciation
   d. rises; inflow; appreciation
6. The nation of Elbonia has long banned the export of its highly prized puka shells. A newly elected president, however, removes the export ban. This change in policy causes the nation's currency to _____, making the goods Elbonia imports _____ expensive.
   a. depreciate; less
   b. depreciate; more
   c. appreciate; less
   d. appreciate; more
7. A civil war abroad causes foreign investors to move their funds to the safe haven of the United States, leading to _____U.S. interest rates and a _____ U.S. dollar.
   a. higher; weaker
   b. higher; stronger
   c. lower; weaker
   d. lower; stronger
8. If business leaders in Great Britain become more confident in their economy, they will increase investment, causing the British pound to _____ and pushing the British trade balance toward _____.
   a. appreciate; deficit
   b. appreciate; surplus
   c. depreciate; deficit
   d. depreciate; surplus

Answers are at the end of the chapter.

33-4 Conclusion

International economics is a topic of increasing importance. In many nations, people buy large quantities of goods from abroad, and much of what they produce is sold overseas. Through mutual funds and other financial institutions, people borrow and lend in world financial markets. As a result, a full analysis of a modern economy requires an understanding of how it interacts with others around the world. This chapter has provided a basic model for thinking about the macroeconomics of open economies.

The study of international economics is valuable, but we should be careful not to exaggerate its importance. In the United States and many other countries, policymakers and commentators are often quick to blame foreigners for economic problems. By contrast, economists typically view most problems as homegrown. For example, politicians are prone to see foreign competition as a threat to national living standards, while economists are more likely to lament excessive budget deficits and insufficient national saving. Low saving impedes growth in capital, productivity, and living standards, regardless of whether the economy is open or closed. Foreigners are a convenient target for politicians because blaming foreigners is a way to avoid responsibility without insulting any domestic constituency. When you hear popular discussions of international trade and finance, it is important to distinguish myth from reality. The tools you have learned in this chapter and previous ones should help in that endeavor.

Chapter in a Nutshell

- Two markets are central to the macroeconomics of open economies: the market for loanable funds and the market for foreign-currency exchange. In the market for loanable funds, the real interest rate adjusts to balance the supply of loanable funds (from national saving) and the demand for loanable funds (for domestic investment and net capital outflow). In the market for foreign-currency exchange, the real exchange rate adjusts to balance the supply of dollars (from net capital outflow) and the demand for dollars (for net exports). Because net capital outflow contributes to the demand for loanable funds and also provides the supply of dollars for foreign-currency exchange, it is the variable that connects these two markets.
- A policy that reduces national saving, such as a government budget deficit, reduces the supply of loanable funds and drives up the interest rate. The higher interest rate causes the net capital outflow to decline, reducing the supply of dollars in the market for foreign-currency exchange. The dollar appreciates, and net exports fall.
- Although restrictive trade policies, such as tariffs or quotas on imports, are sometimes advocated as a way to alter the trade balance, they do not necessarily have that effect. A trade restriction increases net exports for any given exchange rate and, therefore, increases the demand for dollars in the market for foreign-currency exchange. As a result, the dollar appreciates, making domestic goods more expensive relative to foreign goods. This appreciation offsets the initial impact of the trade restriction on net exports.
- When investors change their attitudes about holding assets of a country, the ramifications for the country's economy can be profound. In particular, political instability can lead to capital flight, which tends to increase interest rates and cause the currency to depreciate.

Key Concepts

trade policy, p. 698

capital flight, p. 701

Questions for Review

1. Describe supply and demand in the market for loanable funds and in the market for foreign-currency exchange. How are these markets linked?
2. Why are budget deficits and trade deficits sometimes called the twin deficits?
3. Suppose that a textile workers' union encourages people to buy only American-made clothes. What would this policy do to the trade balance and the real exchange rate? What is the impact on the textile industry? What is the impact on the auto industry?
4. What is capital flight? When a country experiences capital flight, what is the effect on its interest rate and exchange rate?

Problems and Applications

1. Japan generally runs a significant trade surplus. Do you think this surplus is most related to high foreign demand for Japanese goods, low Japanese demand for foreign goods, a high Japanese saving rate relative to Japanese investment, or structural barriers against imports into Japan? Explain your answer.
2. Suppose that Congress is considering an investment tax credit, which subsidizes domestic investment.
   a. How does this policy affect national saving, domestic investment, net capital outflow, the interest rate, the exchange rate, and the trade balance?
   b. Representatives of several large exporters oppose the policy. Why might that be the case?
3. The chapter notes that the rise in the U.S. trade deficit during the 1980s was largely due to the rise in the U.S. budget deficit. On the other hand, the popular press sometimes claims that the increased trade deficit resulted from a decline in the quality of U.S. products relative to foreign products.
   a. Assume that U.S. products did decline in relative quality during the 1980s. How did this decline affect net exports **at any given exchange rate**?
   b. Draw a three-panel diagram to show the effect of this shift in net exports on the U.S. real exchange rate and trade balance.
   c. Is the claim in the popular press consistent with the model in this chapter? Does a decline in the quality of U.S. products have any effect on the U.S. standard of living? (Hint: When a country sells its goods to foreigners, what does it receive in return?)
4. An economist discussing trade policy in *The New Republic* wrote, "One of the benefits of the United States removing its trade restrictions [is] the gain to U.S. industries that produce goods for export. Export industries would find it easier to sell their goods abroad—even if other countries didn't follow our example and reduce their trade barriers." Explain in words why U.S. **export** industries would benefit from a reduction in restrictions on **imports** to the United States.
5. Suppose the French suddenly develop a strong taste for California wines. Answer the following questions in words and with a diagram.
   a. What happens to the demand for dollars in the market for foreign-currency exchange?
   b. What happens to the value of the dollar in the market for foreign-currency exchange?
   c. What happens to U.S. net exports?
6. Senator Blowhard renounces his past support for protectionism: "The U.S. trade deficit must be reduced, but import quotas only annoy our trading partners. If we subsidize U.S. exports instead, we can reduce the deficit by increasing our competitiveness." Using a three-panel diagram, show the effect of an export subsidy on net exports and the real exchange rate. Do you agree with the senator?
7. Suppose the United States decides to subsidize the export of U.S. agricultural products, but it does not increase taxes or decrease any other government spending to offset this expenditure. Using a three-panel diagram, show what happens to national saving, domestic investment, net capital outflow, the interest rate, the exchange rate, and the trade balance. Also explain in words how this U.S. policy affects the amount of imports, exports, and net exports.
8. Suppose that real interest rates increase across Europe. Explain how this development affects U.S. net capital outflow. Then explain how it affects U.S. net exports by using a formula from the chapter and

by drawing a diagram. What happens to the U.S. real interest rate and real exchange rate?

9. Suppose that Americans decide to increase their saving.
   a. If the elasticity of U.S. net capital outflow with respect to the real interest rate is very high, will this increase in private saving have a large or small effect on U.S. domestic investment?
   b. If the elasticity of U.S. exports with respect to the real exchange rate is very low, will this increase in private saving have a large or small effect on the U.S. real exchange rate?

Quick**Quiz Answers**

1. **c** 2. **b** 3. **d** 4. **a** 5. **b** 6. **c** 7. **d** 8. **a**

Chapter

34

Aggregate Demand and Aggregate Supply

Economic activity fluctuates from year to year. In most years, the production of goods and services rises. Because of increases in the labor force, increases in the capital stock, and advances in technological knowledge, the economy can produce more and more over time. This growth allows people to enjoy a higher standard of living. On average, over the past half century, the production of the U.S. economy as measured by real GDP has grown by about 3 percent per year.

In some years, however, instead of expanding, the economy contracts. Firms find themselves unable to sell all the goods and services they have to offer, so they reduce production. Workers are laid off, unemployment becomes widespread, and factories are left idle. With the economy producing fewer goods and services, real GDP and other measures of income decline. Such a period of

SHIFT DRIVE/SHUTTERSTOCK.COM; ERDARKVE+/GETTY IMAGES

recession
a period of declining real incomes and rising unemployment

depression
a severe recession

falling incomes and rising unemployment is called a **recession**. When it is severe, it is sometimes called a **depression**.

In 2008 and 2009, the U.S. economy experienced a downturn that is now called "The Great Recession." From the fourth quarter of 2007 to the second quarter of 2009, real GDP fell by 4.0 percent. The unemployment rate rose from 4.4 percent in May 2007 to 10.0 percent in October 2009, the highest level in more than a quarter century, and it remained above 8 percent for the next three years. Not surprisingly, for students graduating during this time, good jobs were hard to find.

The next recession occurred in 2020 during the coronavirus pandemic, and it was quicker and steeper. Real GDP fell by 10 percent from the fourth quarter of 2019 to the second quarter of 2020. Unemployment rose from 3.5 percent in February 2020 to 14.8 percent just two months later. This time, the economy rebounded quickly. In December 2021, the unemployment rate was back to 3.9 percent, and the number of job openings reached a record high.

What causes short-run fluctuations in economic activity? What, if anything, can public policy do to prevent periods of falling incomes and rising unemployment? When downturns occur, how can policymakers reduce their length and severity? We now take up these questions.

The variables at the center of our analysis are familiar from previous chapters. They include GDP, unemployment, interest rates, and the price level. Also familiar are the policy instruments of government spending, taxes, and the money supply. What differs from our earlier discussions is the time horizon. So far, our goal has been to explain the behavior of these variables in the long run. Our goal now is to explain their short-run deviations from long-run trends. In other words, instead of focusing on the forces that explain economic growth from generation to generation, we are now interested in the forces that explain economic fluctuations from year to year.

Economists debate how best to explain short-run fluctuations, but most use the **model of aggregate demand and aggregate supply**. This chapter introduces the model's two pieces: the aggregate-demand curve and the aggregate-supply curve. But first, let's look at some facts that describe the ups and downs of the economy.

34-1 Three Key Facts about Economic Fluctuations

Short-run fluctuations in economic activity have occurred in all countries throughout history. Three facts are most important in describing these fluctuations.

34-1a Fact 1: Economic Fluctuations Are Irregular and Unpredictable

Fluctuations in the economy are often called **the business cycle**. As this term suggests, economic fluctuations correspond to changes in business conditions. When real GDP grows rapidly, business is good. During such expansionary periods, most firms find that customers are plentiful and profits are growing. But when real GDP falls during recessions, businesses have trouble. During periods of economic contraction, most firms experience declining sales and dwindling profits.

The term **business cycle** is somewhat misleading, however. It suggests that economic fluctuations follow a regular, predictable pattern, like a sine wave or a heartbeat. In fact, fluctuations in economic activity are not at all regular, and they are almost impossible to predict with much accuracy.

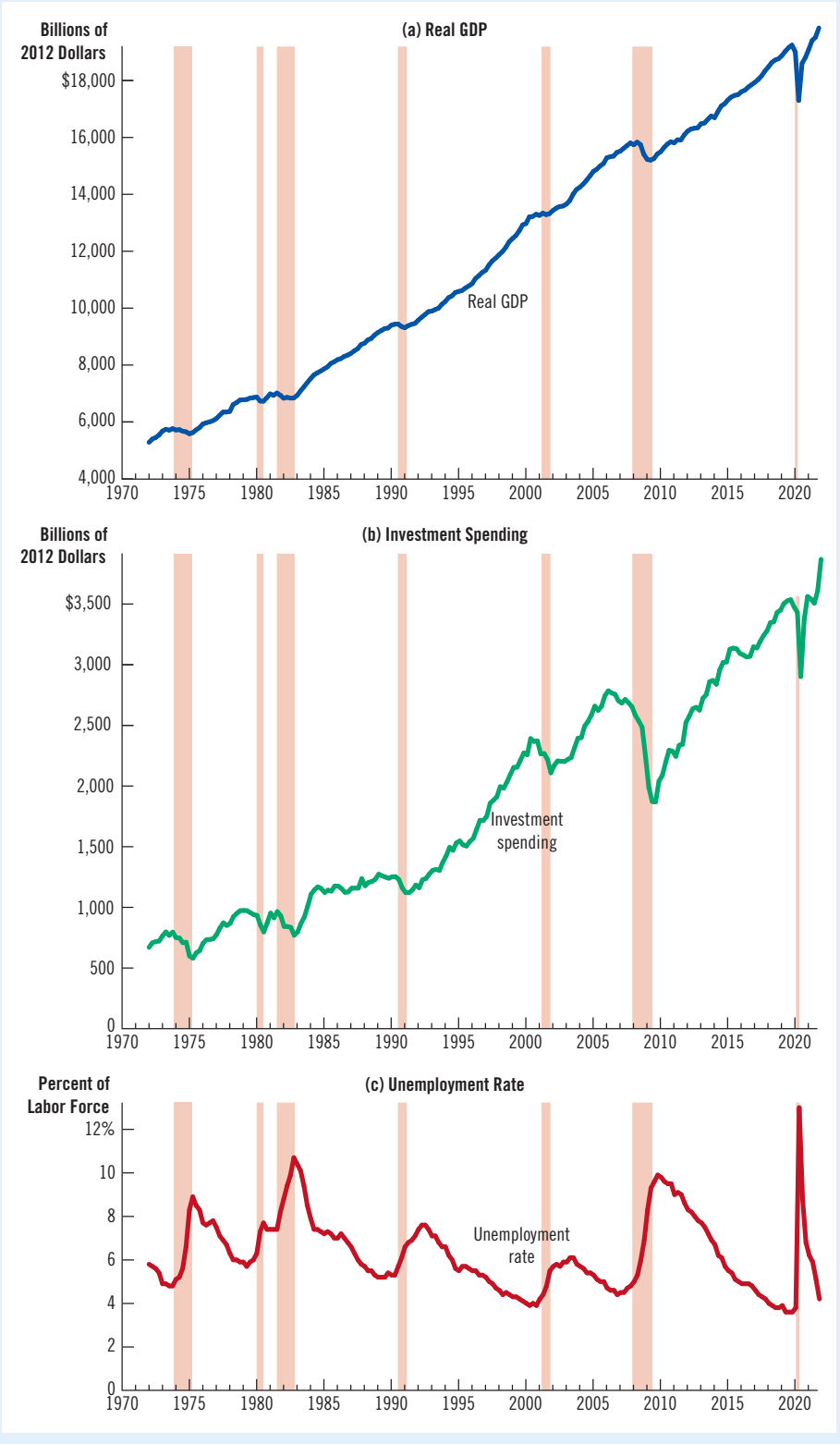
Panel (a) of Figure 1 shows real GDP for the U.S. economy since 1972. The shaded areas represent recessions, and you can see that they don't come at regular intervals.

Figure 1

A Look at Short-Run Economic Fluctuations

This figure shows real GDP in panel (a), investment spending in panel (b), and unemployment in panel (c) for the U.S. economy. Recessions are shown as the shaded areas. Notice that real GDP and investment spending decline during recessions, while unemployment rises.

Source: U.S. Department of Commerce; U.S. Department of Labor.



Sometimes recessions are close together, such as those of 1980 and 1982. Sometimes the economy goes many years without a recession. The longest period in U.S. history without a recession lasted 128 months. It began in June 2009 and ended in February 2020, when the coronavirus recession began.

34-1b Fact 2: Most Macroeconomic Quantities Fluctuate Together

Real GDP is the variable most commonly used to monitor short-run changes in the economy because it is the most comprehensive measure of economic activity. It measures the value of all final goods and services produced within a given period of time as well as the total income (adjusted for inflation) of everyone in the economy.

ROBERT MANKOFF/CARTOONSTOCK LTD

"You're fired. Pass it on."

It turns out, however, that for monitoring short-run fluctuations, it does not really matter which measure of economic activity you use. Most macroeconomic variables that measure some type of income, spending, or production fluctuate closely together. When real GDP falls in a recession, so do personal income, corporate profits, consumer spending, investment spending, industrial production, retail sales, home sales, auto sales, and so on. Because recessions are economy-wide phenomena, they appear in many sources of macroeconomic data.

Although many macroeconomic variables fluctuate together, they move by different amounts. In particular, as panel (b) of Figure 1 shows, investment spending varies greatly over the business cycle. Investment averages only about one-sixth of GDP, but it accounts for about two-thirds of the fall in GDP during recessions. In other words, when the economy contracts, much of the decline is due to reduced spending on new factories, housing, and inventories.

34-1c Fact 3: As Output Falls, Unemployment Rises

Changes in the economy's output of goods and services are strongly correlated with changes in the utilization of the economy's labor force. In other words, when real GDP declines, the unemployment rate rises. This fact is hardly surprising: When firms choose to reduce production, they lay off workers, expanding the pool of unemployed.

Panel (c) of Figure 1 shows the unemployment rate in the U.S. economy since 1972. Once again, the shaded areas indicate recessions. The figure shows that, in each recession, the unemployment rate rises substantially. When the recession ends and real GDP starts to expand, the unemployment rate gradually declines. Because there are always some workers between jobs, the unemployment rate is never zero. Instead, it fluctuates around its natural rate of about 5 percent.

QuickQuiz

1. When the economy goes into a recession, real GDP _____ and unemployment _____.
   a. rises; rises
   b. rises; falls
   c. falls; rises
   d. falls; falls

2. Recessions occur
   a. regularly, about every 3 years.
   b. regularly, about every 7 years.
   c. regularly, about every 12 years.
   d. irregularly.

Answers are at the end of the chapter.

34-2 Explaining Short-Run Economic Fluctuations

Describing what happens to economies as they fluctuate is easy. Explaining why these fluctuations occur is more difficult. Compared with the topics in previous chapters, the theory of economic fluctuations remains controversial. This chapter begins to develop the model that most economists use to explain short-run fluctuations in economic activity.

34-2a The Assumptions of Classical Economics

Previous chapters developed theories to explain what determines most important macroeconomic variables in the long run. Chapter 26 explained the level and growth of productivity and real GDP. Chapters 27 and 28 explained how the financial system works and how the real interest rate adjusts to balance saving and investment. Chapter 29 explained why there is always some unemployment in the economy. Chapters 30 and 31 explained the monetary system and how changes in the money supply affect the price level, the inflation rate, and the nominal interest rate. Chapters 32 and 33 extended this analysis to open economies to explain the trade balance and the exchange rate.

This previous analysis was based on two related ideas: the classical dichotomy and monetary neutrality. Recall that the classical dichotomy is the separation of variables into real variables (those that measure quantities or relative prices) and nominal variables (those measured in terms of money). According to classical macroeconomic theory, changes in the money supply affect nominal variables but not real variables. As a result of this monetary neutrality, we were able to examine the determinants of real variables (real GDP, the real interest rate, and unemployment) without introducing nominal variables (the money supply and the price level).

In a sense, money does not matter in a classical world. If the quantity of money in the economy were to double, everything would cost twice as much, and everyone's income would be twice as high. But so what? The change would be **nominal** (by the standard meaning of "nearly insignificant"). The things that people **really** care about—whether they have a job, how many goods and services they can afford, and so on—would be exactly the same.

This classical view is sometimes described by the saying, "Money is a veil." That is, nominal variables may be the first things we see when we observe an economy because economic variables are often expressed in units of money. But more important are the real variables and the forces that determine them. According to classical theory, to understand these real variables, we need to look behind the veil.

34-2b The Reality of Short-Run Fluctuations

Do these assumptions of classical macroeconomic theory apply to the world in which we live? The answer to this question is central to understanding how the economy works. **Most economists believe that classical theory describes the world in the long run but not in the short run.**

Consider again the impact of money on the economy. Most economists believe that, beyond a period of several years, changes in the money supply affect prices and other nominal variables but do not affect real GDP, unemployment, and other real variables—just as classical theory says. When studying year-to-year changes in the economy, however, the assumption of monetary neutrality is no longer appropriate. In the short run, real and nominal variables are highly intertwined, and changes in the money supply can temporarily push real GDP away from its long-run trend.

Even classical economists, such as David Hume, realized that monetary neutrality did not hold in the short run. From his vantage point in 18th-century England, Hume observed that when the money supply expanded after gold discoveries, it took some time for prices to rise and that, in the meantime, the economy enjoyed higher employment and production.

To understand how the economy works in the short run, we need a new model. It can be built using many of the tools developed in previous chapters, but it must abandon the classical dichotomy and the neutrality of money. We can no longer separate the analysis of real variables such as output and employment from the analysis of nominal variables such as money and the price level. Our new model focuses on how real and nominal variables interact.

model of aggregate demand and aggregate supply
the model that most economists use to explain short-run fluctuations in economic activity around its long-run trend

aggregate-demand curve
a curve that shows the quantity of goods and services that households, firms, the government, and customers abroad want to buy at each price level

aggregate-supply curve
a curve that shows the quantity of goods and services that firms choose to produce and sell at each price level

34-2c The Model of Aggregate Demand and Aggregate Supply

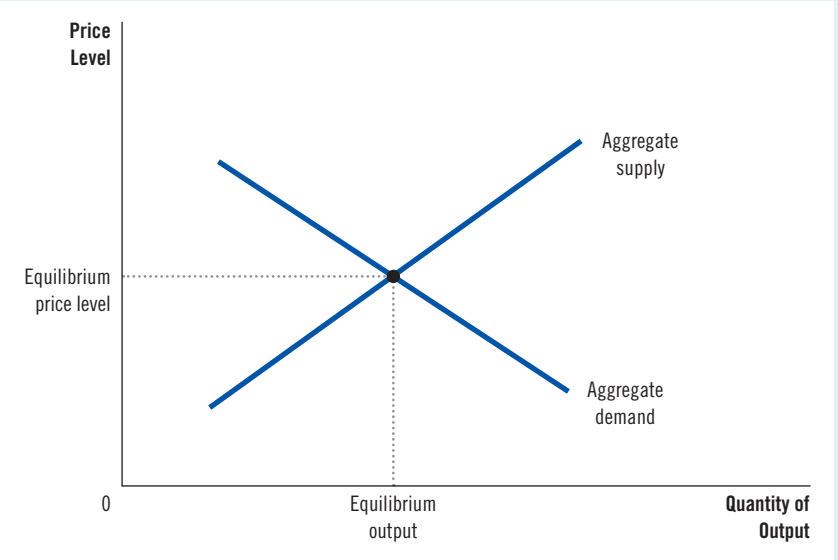
Our model of short-run economic fluctuations focuses on the behavior of two variables. The first is the economy's output of goods and services, as measured by real GDP. The second is the average level of prices, as measured by the CPI or the GDP deflator. Notice that output is a real variable, and the price level is a nominal variable. By focusing on the relationship between these two variables, we are departing from the classical assumption that real and nominal variables can be studied separately.

We analyze fluctuations in the economy as a whole using the **model of aggregate demand and aggregate supply**, which is illustrated in Figure 2. On the vertical axis is the overall price level in the economy. On the horizontal axis is the overall quantity of goods and services produced in the economy. The **aggregate-demand curve** shows the quantity of goods and services that households, firms, the government, and customers abroad want to buy at each price level. The **aggregate-supply curve** shows the quantity of goods and services that firms produce and sell at each price level. According to this model, the price level and the quantity of output adjust to bring aggregate demand and aggregate supply into balance.

Figure 2

Aggregate Demand and Aggregate Supply

Economists use the model of aggregate demand and aggregate supply to analyze economic fluctuations. On the vertical axis is the overall level of prices. On the horizontal axis is the economy's total output of goods and services. Output and the price level adjust to the point at which the aggregate-supply and aggregate-demand curves intersect.



It is tempting to view the model of aggregate demand and aggregate supply as nothing more than a large version of the model of market demand and market supply introduced in Chapter 4. But, in fact, this model is quite different. When we consider demand and supply in the market for a specific good—ice cream, for instance—the behavior of buyers and sellers depends on the ability of resources to move from one market to another. When the price of ice cream rises, the quantity demanded falls because buyers can use their incomes to buy products other than ice cream. Similarly, a higher price of ice cream raises the quantity supplied because firms that produce ice cream can increase production by hiring workers from other parts of the economy. This **microeconomic** substitution from one market to another is impossible for the economy as a whole. After all, the quantity that our model is trying to explain—real GDP—measures the **total** quantity of goods and services produced by **all** firms in **all** markets. To understand why the aggregate-demand curve slopes downward and why the aggregate-supply curve slopes upward, we need a **macroeconomic** theory that explains the total quantity of goods and services demanded and the total quantity of goods and services supplied. Developing such a theory is our next task.

QuickQuiz

3. According to classical macroeconomic theory and monetary neutrality, changes in the money supply affect
   a. the unemployment rate.
   b. real GDP.
   c. the GDP deflator.
   d. none of the above.
4. Most economists believe that classical macroeconomic theory
   a. is valid only in the long run.
   b. is valid only in the short run.
   c. is always valid.
   d. is never valid.
5. In the model of aggregate demand and aggregate supply, the quantity of _____ is on the horizontal axis, and the _____ is on the vertical axis.
   a. output; interest rate
   b. output; price level
   c. money; interest rate
   d. money; price level

Answers are at the end of the chapter.

34-3 The Aggregate-Demand Curve

The aggregate-demand curve depicts the quantity of all goods and services demanded in the economy at any price level. As Figure 3 illustrates, the aggregate-demand curve slopes downward. Other things being equal, a decrease in the economy's overall level of prices (from, say, P_1 to P_2) raises the quantity of goods and services demanded (from Y_1 to Y_2). Conversely, an increase in the price level reduces the quantity of goods and services demanded.

34-3a Why the Aggregate-Demand Curve Slopes Downward

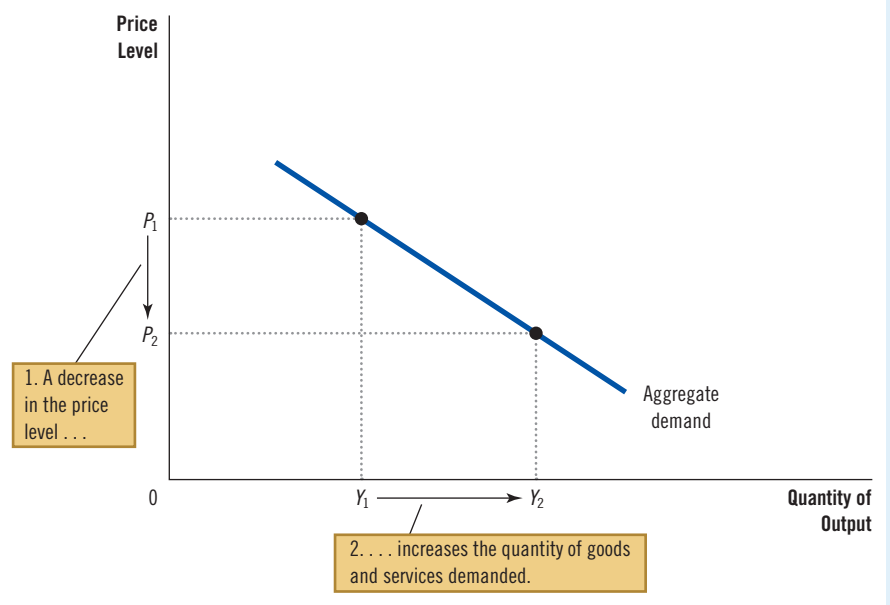
Why does a change in the price level move the quantity of goods and services demanded in the opposite direction? To answer this question, it is useful to recall that an economy's GDP (which we denote as Y) is the sum of its consumption (C), investment (I), government purchases (G), and net exports (NX):

$$Y = C + I + G + NX.$$

Figure 3

The Aggregate-Demand Curve

A fall in the price level from P_1 to P_2 increases the quantity of goods and services demanded from Y_1 to Y_2. There are three reasons for this negative relationship. As the price level falls, real wealth rises, interest rates fall, and the exchange rate depreciates. These effects stimulate spending on consumption, investment, and net exports. Increased spending on any or all of these components of output means a larger quantity of goods and services demanded.



Each of these four components contributes to the aggregate demand for goods and services. For now, we assume that government spending is fixed by policy. The other three components of spending—consumption, investment, and net exports—depend on economic conditions and, in particular, on the price level. To understand the downward slope of the aggregate-demand curve, we must examine how the price level affects the quantity of goods and services demanded for consumption, investment, and net exports.

The Price Level and Consumption: The Wealth Effect Consider the money you hold in your wallet and bank account. The nominal value of this money is fixed: One dollar is always worth one dollar. Yet the **real** value of a dollar can change. If a candy bar costs one dollar, then a dollar is worth one candy bar. If the price of a candy bar falls to 50 cents, one dollar is worth two candy bars. Thus, when the price level falls, the dollars you hold rise in value, increasing your real wealth and your ability to buy goods and services.

This logic provides the first reason the aggregate-demand curve slopes downward. **A decrease in the price level raises the real value of money and makes consumers wealthier, encouraging them to spend more. The increase in consumer spending means a larger quantity of goods and services demanded. Conversely, an increase in the price level reduces the real value of money and makes consumers poorer, reducing consumer spending and the quantity of goods and services demanded.**

The Price Level and Investment: The Interest-Rate Effect The price level is one determinant of the quantity of money demanded. When the price level is lower, people do not need to hold as much money to buy the goods and services they want. Therefore, when the price level falls, people try to reduce their money holdings by lending some of it out. For instance, they might use their excess money to buy interest-bearing bonds. Or they might deposit the excess money in an interest-bearing savings account, and the bank would lend out these funds. In either case,

as people try to convert some of their money into interest-bearing assets, they drive down interest rates. (The next chapter analyzes this process in more detail.)

Interest rates, in turn, affect spending on goods and services. Because a lower interest rate makes borrowing less expensive, it encourages firms to borrow more to invest in new plants and equipment, and it encourages households to borrow more to invest in new housing. (A lower interest rate might also stimulate consumer spending, especially spending on large, durable purchases such as cars, which are often bought on credit.) In short, a lower interest rate increases the quantity of goods and services demanded.

This logic provides the second reason the aggregate-demand curve slopes downward. **A lower price level reduces the interest rate, encourages spending on investment goods, and increases the quantity of goods and services demanded. Conversely, a higher price level raises the interest rate, discourages investment spending, and decreases the quantity of goods and services demanded.**

The Price Level and Net Exports: The Exchange-Rate Effect For the reasons just discussed, a lower price level in the United States lowers the U.S. interest rate. In response to the lower interest rate, some investors will seek higher returns by moving their assets out of the United States. For instance, as the interest rate on American bonds falls, a mutual fund might sell American bonds to buy German bonds. As the mutual fund tries to convert its dollars into euros to buy the German bonds, it increases the supply of dollars in the market for foreign-currency exchange.

The increased supply of dollars to be exchanged for euros causes the dollar to depreciate relative to the euro. This alters the real exchange rate—the relative price of domestic and foreign goods. Because each dollar buys fewer units of foreign currencies, foreign goods become more expensive compared with domestic goods.

The change in relative prices affects spending at home and abroad. Because foreign goods are more expensive, Americans buy less from other countries, reducing U.S. imports of goods and services. At the same time, because U.S. goods have become relatively cheap, foreigners buy more from the United States, increasing U.S. exports. Net exports equal exports minus imports, so both of these changes cause U.S. net exports to increase. In this way, the depreciation of the dollar leads to an increase in the quantity of goods and services demanded.

This logic yields the third reason the aggregate-demand curve slopes downward. **When a fall in the U.S. price level causes U.S. interest rates to fall, the real value of the dollar declines in foreign exchange markets. This depreciation stimulates U.S. net exports and increases the quantity of goods and services demanded. Conversely, when the U.S. price level rises and causes U.S. interest rates to rise, the real value of the dollar increases, and this appreciation reduces U.S. net exports and the quantity of goods and services demanded.**

Summing Up There are three distinct but related reasons a fall in the price level increases the quantity of goods and services demanded:

1. Consumers become wealthier, stimulating the demand for consumption goods.
2. Interest rates fall, stimulating the demand for investment goods.
3. The currency depreciates, stimulating the demand for net exports.

The same three effects work in reverse: When the price level rises, decreased wealth depresses consumer spending, higher interest rates depress investment spending, and currency appreciation depresses net exports.

Here is a thought experiment to hone your intuition about these effects. Imagine that one day you wake up and notice that, for some mysterious reason, the prices of all goods and services have fallen by half, so the dollars you are holding are worth twice as much. In real terms, you now have twice as much money as you had when you went to bed the night before. What would you do with the extra money? You could spend it at your favorite restaurant, increasing consumer spending. You could lend it out (by buying a bond or depositing it in a bank), reducing interest rates and increasing investment spending. Or you could invest it overseas (by buying shares in an international mutual fund), reducing the real exchange value of the dollar and increasing net exports. No matter which of these options you choose, the fall in the price level leads to an increase in the quantity of goods and services demanded. This relationship is what the downward slope of the aggregate-demand curve represents.

Keep in mind that the aggregate-demand curve (like all demand curves) is drawn holding "other things equal." In particular, these three explanations of the downward-sloping aggregate-demand curve assume that the money supply is fixed. That is, we have been considering how a change in the price level affects the demand for goods and services, holding the amount of money in the economy constant. As we will see, a change in the quantity of money shifts the aggregate-demand curve. At this point, just remember that the aggregate-demand curve is drawn for a given quantity of the money supply.

34-3b Why the Aggregate-Demand Curve Might Shift

The downward slope of the aggregate-demand curve shows that a fall in the price level increases the overall quantity of goods and services demanded. Many other factors also affect the quantity of goods and services demanded. When one of these other factors changes, the quantity of goods and services demanded at every price level changes, and the aggregate-demand curve shifts.

Here are examples of events that shift aggregate demand. We can group them according to the component of spending that is most directly affected.

Shifts Arising from Changes in Consumption Suppose Americans suddenly become more concerned about saving for retirement and, as a result, reduce current consumption. Because the quantity of goods and services demanded at any price level is now lower, the aggregate-demand curve shifts to the left. Conversely, imagine that a stock market boom makes people wealthier and less concerned about saving. The resulting increase in consumer spending means a greater quantity of goods and services demanded at any price level, so the aggregate-demand curve shifts to the right.

Any event that changes how much people want to consume at a given price level shifts the aggregate-demand curve. One policy variable that has this effect is taxation. When the government cuts taxes, it encourages people to spend more, so the aggregate-demand curve shifts to the right. When the government raises taxes, people cut back on their spending, and the aggregate-demand curve shifts to the left.

Shifts Arising from Changes in Investment Any event that changes how much firms want to invest at a given price level also shifts the aggregate-demand curve. For instance, imagine that the computer industry introduces a faster line of computers, and many firms decide to invest in new computer systems. Because the quantity of

goods and services demanded at any price level is higher, the aggregate-demand curve shifts to the right. Conversely, if firms become pessimistic about future business conditions, they may cut back on investment spending, shifting the aggregate-demand curve to the left.

Tax policy can also influence aggregate demand through investment. For example, other things being equal, an investment tax credit (a tax rebate tied to a firm's investment spending) increases the quantity of investment goods that firms want to buy and therefore shifts the aggregate-demand curve to the right. The repeal of an investment tax credit reduces investment and shifts the aggregate-demand curve to the left.

Another policy variable that can influence investment and aggregate demand is the money supply. As the next chapter explains, an increase in the money supply lowers the interest rate in the short run. This decrease in the interest rate makes borrowing less costly, stimulating investment spending and shifting the aggregate-demand curve to the right. Conversely, a decrease in the money supply raises the interest rate, discourages investment spending, and shifts the aggregate-demand curve to the left. Many economists believe that throughout U.S. history, changes in monetary policy have been an important source of shifts in aggregate demand.

Shifts Arising from Changes in Government Purchases The most direct way that policymakers shift the aggregate-demand curve is through government purchases. For example, suppose Congress decides to reduce purchases of new weapons systems. Because the quantity of goods and services demanded at any price level is lower, the aggregate-demand curve shifts to the left. Conversely, if state governments start building more highways, the result is a greater quantity of goods and services demanded at any price level, so the aggregate-demand curve shifts to the right.

Shifts Arising from Changes in Net Exports Any event that changes net exports for a given price level also shifts aggregate demand. For instance, when Europe has a recession, it buys fewer goods from the United States. U.S. net exports decline at every price level, shifting the aggregate-demand curve for the U.S. economy to the left. When Europe recovers from its recession, it buys more U.S. goods, and the aggregate-demand curve shifts to the right.

Net exports can also change because international speculators affect the exchange rate. Suppose, for instance, that these speculators lose confidence in foreign economies and want to move wealth into the U.S. economy. In doing so, they bid up the value of the U.S. dollar in the foreign exchange market. This appreciation of the dollar makes U.S. goods more expensive relative to foreign goods, depressing net exports and shifting the aggregate-demand curve to the left. Conversely, speculation that causes a depreciation of the dollar stimulates net exports and shifts the aggregate-demand curve to the right.

Summing Up The next chapter analyzes the aggregate-demand curve in more detail. There, we examine more precisely how the tools of monetary and fiscal policy can shift aggregate demand and when policymakers should use these tools for that purpose. For now, you should have some understanding of why the aggregate-demand curve slopes downward and what kinds of events and policies can shift this curve. Table 1 summarizes what we have learned so far.

Table 1

The Aggregate-Demand Curve: Summary

Why Does the Aggregate-Demand Curve Slope Downward?

1. **The Wealth Effect:** A lower price level increases real wealth, stimulating spending on consumption.
2. **The Interest-Rate Effect:** A lower price level reduces the interest rate, stimulating spending on investment.
3. **The Exchange-Rate Effect:** A lower price level causes the real exchange rate to depreciate, stimulating spending on net exports.

Why Might the Aggregate-Demand Curve Shift?

1. **Shifts Arising from Changes in Consumption:** An event that causes consumers to spend more at a given price level (a tax cut, a stock market boom) shifts the aggregate-demand curve to the right. An event that causes consumers to spend less at a given price level (a tax hike, a stock market decline) shifts the curve to the left.
2. **Shifts Arising from Changes in Investment:** An event that causes firms to invest more at a given price level (optimism about the future, a fall in interest rates due to an increase in the money supply) shifts the aggregate-demand curve to the right. An event that causes firms to invest less at a given price level (pessimism about the future, a rise in interest rates due to a decrease in the money supply) shifts the curve to the left.
3. **Shifts Arising from Changes in Government Purchases:** An increase in government purchases of goods and services (greater spending on defense or highway construction) shifts the aggregate-demand curve to the right. A decrease in government purchases of goods and services (a cutback in defense or highway spending) shifts the curve to the left.
4. **Shifts Arising from Changes in Net Exports:** An event that raises spending on net exports at a given price level (a boom overseas, speculation that causes a currency depreciation) shifts the aggregate-demand curve to the right. An event that reduces spending on net exports at a given price level (a recession overseas, speculation that causes a currency appreciation) shifts the curve to the left.

QuickQuiz

6. The aggregate-demand curve slopes downward because a fall in the price level causes
   a. real wealth to decrease.
   b. the interest rate to decline.
   c. the currency to appreciate.
   d. all of the above.

7. Which of the following would shift the aggregate-demand curve to the left?
   a. a decline in the stock market
   b. an increase in taxes
   c. a decrease in government spending
   d. all of the above

Answers are at the end of the chapter.

34-4 The Aggregate-Supply Curve

The aggregate-supply curve shows the total quantity of goods and services that firms produce and sell at any price level. Unlike the aggregate-demand curve, which always slopes downward, the slope of the aggregate-supply curve depends on the time horizon being examined. **In the long run, the aggregate-supply curve**

is vertical, while in the short run, it slopes upward. This section explains both the long-run aggregate-supply curve and the short-run aggregate-supply curve. In doing so, it shows why the economy can, in the short run, deviate from the long-run equilibrium described by classical theory.

34-4a Why the Aggregate-Supply Curve Is Vertical in the Long Run

What determines the quantity of goods and services supplied in the long run? We implicitly answered this question earlier in the book when discussing economic growth. **In the long run, an economy's production of goods and services (its real GDP) depends on its supplies of labor, capital, and natural resources and on the available technology used to turn these factors of production into goods and services.**

When we analyzed the forces that govern long-run growth, we did not mention the overall level of prices. The price level was explained in a separate chapter, which developed the quantity theory of money. We learned that if two economies were identical in every way except that one had twice as much money in circulation as the other, the price level would be twice as high in the economy with more money. But because the amount of money does not affect technology or the supplies of labor, capital, and natural resources, the output of goods and services in the two economies would be the same.

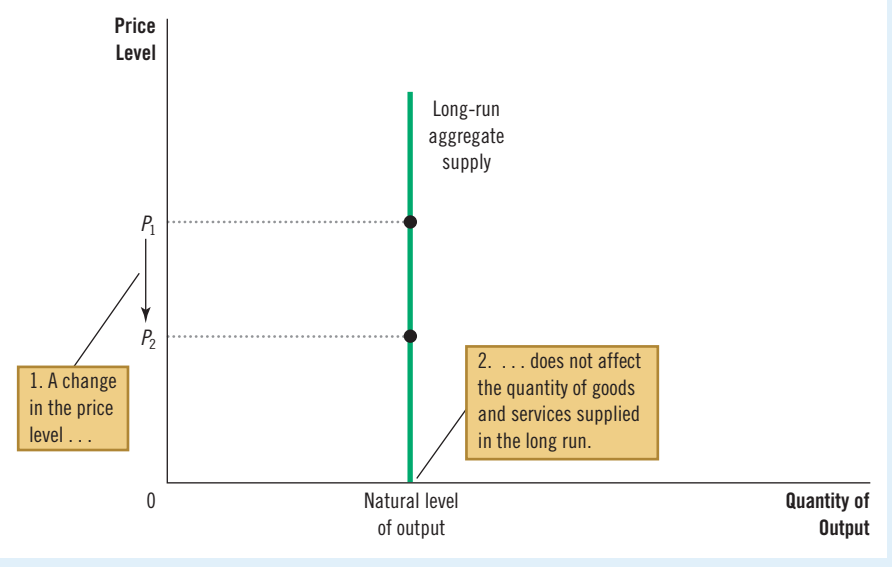
Because the price level does not affect the long-run determinants of real GDP, the long-run aggregate-supply curve is vertical, as in Figure 4. In other words, in the long run, the economy's labor, capital, natural resources, and technology determine the total quantity of goods and services supplied, and this quantity supplied is the same regardless of the price level.

The vertical long-run aggregate-supply curve is a graphical representation of the classical dichotomy and monetary neutrality. As we have discussed, classical

Figure 4

The Long-Run Aggregate-Supply Curve

In the long run, the quantity of output supplied depends on the economy's quantities of labor, capital, and natural resources and on the technology for turning these inputs into output. Because the quantity supplied does not depend on the overall price level, the long-run aggregate-supply curve is vertical at the natural level of output.



macroeconomic theory is based on the assumption that real variables do not depend on nominal variables. The long-run aggregate-supply curve is consistent with this idea because it implies that the quantity of output (a real variable) does not depend on the level of prices (a nominal variable). As noted earlier, most economists believe this principle works well when studying the economy over a period of many years but not for year-to-year changes. **For this reason, the aggregate-supply curve is vertical only in the long run.**

34-4b Why the Long-Run Aggregate-Supply Curve Might Shift

Because classical macroeconomic theory describes the quantity of goods and services produced by an economy in the long run, it also explains the position of the long-run aggregate-supply curve. The long-run level of production is sometimes called **potential output** or **full-employment output**. To be more precise, we call it the **natural level of output** because it shows what the economy produces when unemployment is at its natural, or normal, rate. The natural level of output is the rate of production toward which the economy gravitates in the long run.

natural level of output the production of goods and services that an economy achieves in the long run when unemployment is at its normal rate

Any change in the economy that alters the natural level of output shifts the long-run aggregate-supply curve. Because output in the classical model depends on labor, capital, natural resources, and technological knowledge, we can categorize shifts in the long-run aggregate-supply curve as arising from these four sources.

Shifts Arising from Changes in Labor Imagine that an economy experiences an increase in immigration. Because this increase results in a greater number of workers, the quantity of goods and services supplied would also increase. As a result, the long-run aggregate-supply curve would shift to the right. Conversely, if many workers left the economy to go abroad, the long-run aggregate-supply curve would shift to the left.

The position of the long-run aggregate-supply curve also depends on the natural rate of unemployment, so any change in this rate shifts the long-run aggregate-supply curve. For example, if Congress made unemployment insurance substantially more generous, unemployed workers might not search as hard for new jobs, increasing the natural rate of unemployment and reducing the economy's production of goods and services. As a result, the long-run aggregate-supply curve would shift to the left. Conversely, if Congress enacted a successful job training program for unemployed workers, the natural rate of unemployment would fall, and the long-run aggregate-supply curve would shift to the right.

Shifts Arising from Changes in Capital An increase in the economy's capital stock raises productivity, thereby increasing the quantity of goods and services supplied. As a result, the long-run aggregate-supply curve shifts to the right. Conversely, a decrease in the economy's capital stock reduces productivity and the quantity of goods and services supplied, shifting the long-run aggregate-supply curve to the left.

Notice that the same logic applies regardless of whether we are discussing physical capital such as machines and factories or human capital such as college degrees. An increase in either type of capital will raise the economy's ability to produce goods and services and shift the long-run aggregate-supply curve to the right.

Shifts Arising from Changes in Natural Resources An economy's production depends on its natural resources, including its land, minerals, and weather. The

discovery of a new mineral deposit shifts the long-run aggregate-supply curve to the right. A change in weather patterns that makes farming more difficult shifts the long-run aggregate-supply curve to the left.

In many countries, crucial natural resources are imported. A change in the availability of these resources can also shift the aggregate-supply curve. For example, as we discuss later in this chapter, developments in the world oil market have historically been an important source of shifts in aggregate supply for oil-importing nations.

Shifts Arising from Changes in Technological Knowledge Perhaps the most important reason that the economy today produces more than it did a generation ago is that technological knowledge has advanced. The development of industrial robots, for instance, has allowed firms to produce more goods and services from any given amount of labor, capital, and natural resources. As robot use has spread, it has shifted the long-run aggregate-supply curve to the right.

Although not precisely technological, many other events act like changes in technology. For instance, opening up international trade has effects similar to inventing new production processes because it allows a country to specialize in higher-productivity industries; therefore, it also shifts the long-run aggregate-supply curve to the right. Conversely, if the government passes new regulations preventing firms from using some production methods, perhaps to address worker safety or environmental concerns, the result is a leftward shift in the long-run aggregate-supply curve.

Summing Up Because the long-run aggregate-supply curve reflects the classical model, it provides a new way to express the analysis in earlier chapters. Any policy or event that raised real GDP in previous chapters can be described as increasing the quantity of goods and services supplied and shifting the long-run aggregate-supply curve to the right. Any policy or event that lowered real GDP in previous chapters can be said to reduce the quantity of goods and services supplied and to shift the long-run aggregate-supply curve to the left.

34-4c Using Aggregate Demand and Aggregate Supply to Depict Long-Run Growth and Inflation

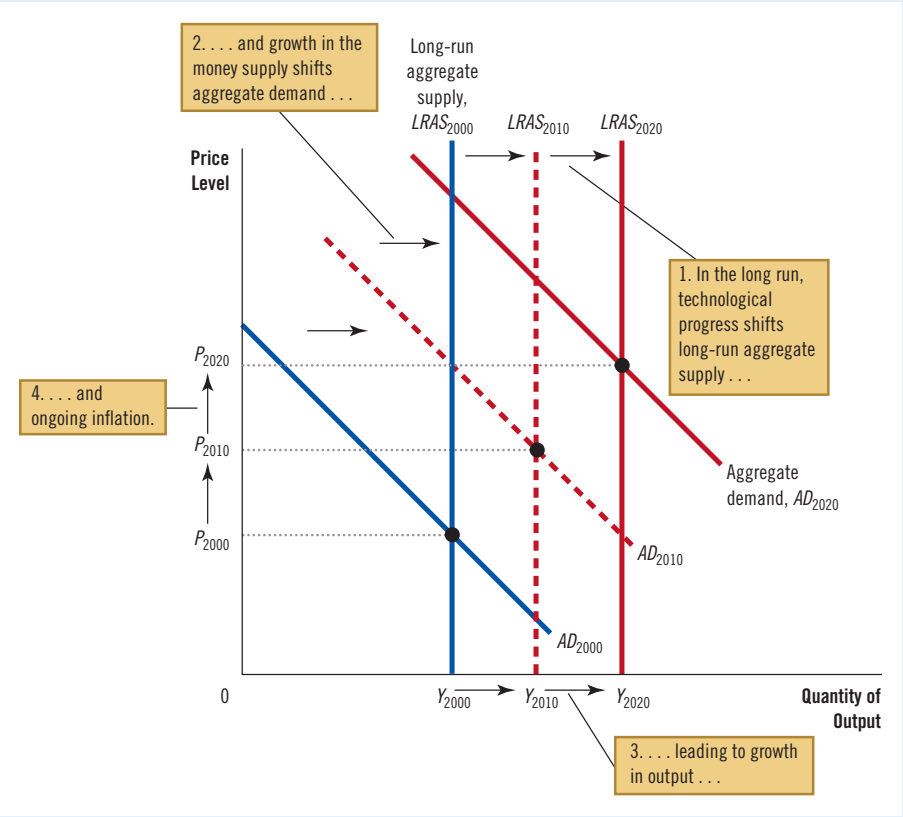
Having introduced the economy's aggregate-demand curve and the long-run aggregate-supply curve, we have a new way to describe the economy's long-run trends. Figure 5 illustrates the changes that occur in an economy from decade to decade. Notice that both curves are shifting. Although many forces influence the economy in the long run and can, in theory, cause such shifts, the two most important forces in practice are technology and monetary policy. Technological progress enhances an economy's ability to produce goods and services, and the resulting increases in output are reflected in continual shifts of the long-run aggregate-supply curve to the right. At the same time, because the Fed increases the money supply over time, the aggregate-demand curve also shifts to the right. As the figure shows, the result is continuing growth in output (as shown by increasing Y) and continuing inflation (as shown by increasing P). This is just another way of representing the classical analysis of growth and inflation conducted in earlier chapters.

The purpose of developing the model of aggregate demand and aggregate supply, however, is not to dress our previous long-run conclusions in new clothing. Instead, it is to provide a framework for short-run analysis, as we will see in a moment. As we develop the short-run model, we simplify the analysis by omitting the ongoing growth

Figure 5

Long-Run Growth and Inflation in the Model of Aggregate Demand and Aggregate Supply

As the economy becomes better able to produce goods and services, primarily because of technological progress, the long-run aggregate-supply curve shifts to the right. At the same time, as the Fed increases the money supply, the aggregate-demand curve also shifts to the right. In this figure, output grows from Y_{2000} to Y_{2010} and then to Y_{2020}, and the price level rises from P_{2000} to P_{2010} and then to P_{2020}. Thus, the model of aggregate demand and aggregate supply offers a new way to describe the classical analysis of growth and inflation.



and inflation shown by the shifts in Figure 5. But always remember that long-run trends are the background on which short-run fluctuations are superimposed. **The short-run fluctuations in output and the price level that we will be studying should be viewed as deviations from the long-run trends of output growth and inflation.**

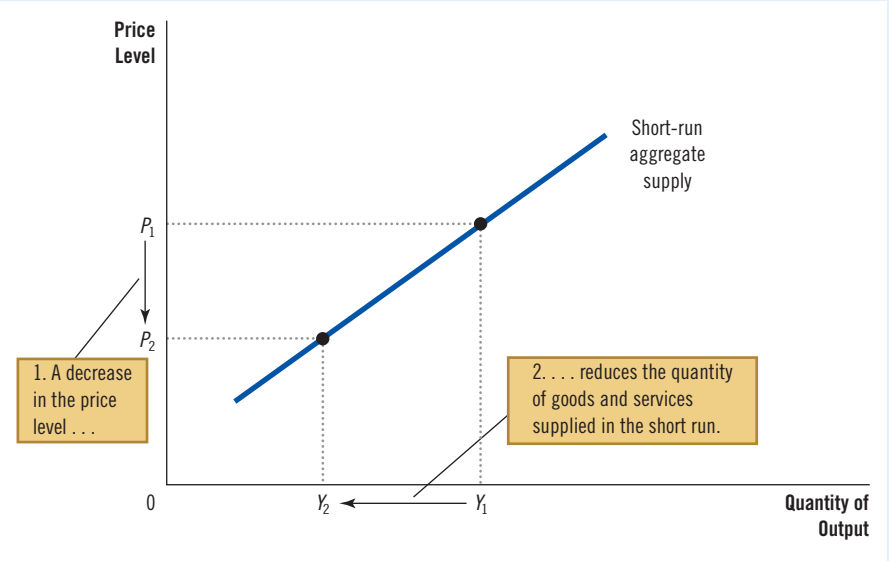
34-4d Why the Aggregate-Supply Curve Slopes Upward in the Short Run

The key difference between the economy in the short run and in the long run is the behavior of aggregate supply. The long-run aggregate-supply curve is vertical because, in the long run, the overall level of prices does not affect the economy's ability to produce goods and services. By contrast, in the short run, the price level **does** affect the economy's output. That is, over a period of a year or two, an increase in the price level tends to increase the quantity of goods and services supplied, and a decrease in the price level tends to reduce the quantity of goods and services supplied. As a result, the short-run aggregate-supply curve slopes upward, as in Figure 6.

Figure 6

The Short-Run Aggregate-Supply Curve

In the short run, a fall in the price level from P_1 to P_2 reduces the quantity of output supplied from Y_1 to Y_2. This positive relationship could be due to sticky wages, sticky prices, or misperceptions. Over time, wages, prices, and perceptions adjust, so this positive relationship is only temporary.



Why do changes in the price level affect output in the short run? Macroeconomists have proposed three theories for the upward slope of the short-run aggregate-supply curve. In each one, some market imperfection causes the supply side of the economy to behave differently in the short run than it does in the long run. The theories differ in their details but share a common theme: in a word, surprise. **The quantity of output supplied deviates from its long-run, or natural, level when the actual price level in the economy deviates from the price level that people expected.** When the price level rises above the expected level, output rises above its natural level, and when the price level falls below the expected level, output falls below its natural level.

The Sticky-Wage Theory The first explanation for the upward slope of the short-run aggregate-supply curve is the sticky-wage theory. This theory is the simplest of the three approaches to aggregate supply, and some economists believe it highlights the most important reason why the short run and long run differ. This theory of short-run aggregate supply is one emphasized in this book.

According to this theory, the short-run aggregate-supply curve slopes upward because nominal wages are slow to adjust to changing economic conditions. In other words, wages are "sticky" in the short run. To some extent, the slow adjustment of nominal wages is attributable to long-term contracts between workers and firms that fix nominal wages, sometimes for as long as three years. This gradual adjustment may also be attributable to slowly changing social norms and notions of fairness, which can influence wage setting.

An example can help explain how sticky nominal wages can result in a short-run aggregate-supply curve that slopes upward. Imagine that a year ago, a firm expected the price level today to be at a certain level: We'll call it 100. Based on this expectation, it signed a contract with its workers agreeing to pay them, say, $30 an hour. In fact, the price level turns out to be only 95. Because prices have fallen below expectations, the firm gets 5 percent less than expected for each unit of the product that it sells. The cost of labor used to make the output, however, is stuck at

$30 per hour. Production is now less profitable, so the firm hires fewer workers and reduces the quantity of output supplied. Over time, the labor contract will expire, and the firm can renegotiate with its workers for a lower nominal wage (which they may accept because prices are lower), but in the meantime, employment and production will remain below their long-run levels.

The same logic works in reverse. Suppose the price level turns out to be 105 and the nominal wage remains stuck at $30. The firm sees that the amount it is paid for each unit sold is up by 5 percent, while its labor costs are not. In response, it hires more workers and increases the quantity of output supplied. Eventually, the workers will demand higher nominal wages to compensate for the higher price level. But for a while, the firm can take advantage of the profit opportunity by increasing employment and production above their long-run levels.

In short, according to the sticky-wage theory, the short-run aggregate-supply curve slopes upward because nominal wages are based on expected prices and do not respond immediately when the actual price level turns out to be different from what was expected. This wage stickiness gives firms an incentive to produce less output when the price level turns out lower than expected and to produce more when the price level turns out higher than expected.

The Sticky-Price Theory Some economists have advocated another approach to explaining the upward slope of the short-run aggregate-supply curve, called the sticky-price theory. While the sticky-wage theory emphasizes that nominal wages adjust slowly over time, the sticky-price theory focuses on the prices of some goods and services, which also adjust sluggishly in response to changing economic conditions. This slow adjustment occurs in part because there are costs to changing prices, called **menu costs**. These menu costs include the expense of printing and distributing catalogs, the time required to change price tags, and even the managerial effort needed to decide on new prices. Because of these costs, prices as well as wages may be sticky in the short run.

To see how sticky prices explain the aggregate-supply curve's upward slope, suppose that each firm in the economy announces its prices in advance based on the economic conditions it expects for the coming year. Suppose further that after prices are announced, the economy experiences an unexpected contraction in the money supply, which (as we have learned) reduces the overall price level in the long run. What happens in the short run? Some firms reduce their prices quickly in response to the change in economic conditions, but many others want to avoid additional menu costs and temporarily lag behind in cutting prices. Because these lagging firms' prices are too high, their sales decline. Declining sales, in turn, cause them to cut back production and employment. In other words, an unexpected fall in the price level leaves some firms with higher-than-desired prices, which depress their sales and induce them to reduce the quantity of goods and services they produce.

Similar reasoning applies when the money supply and price level turn out to be higher than what firms expected when they set their prices. Some firms raise prices quickly in response to the new environment, but others lag behind. Their low prices attract customers, inducing these firms to increase employment and production. Again, for the overall economy, there is a positive association between an unexpected movement in the price level and the quantity of output produced. The upward slope of the short-run aggregate-supply curve reflects this positive association.

The Misperceptions Theory A third approach to explaining the upward slope of the short-run aggregate-supply curve is the misperceptions theory. It holds that

changes in the overall price level can temporarily mislead suppliers about what is happening in the markets in which they sell their output. Because of these short-run misperceptions, suppliers respond to changes in the level of prices, leading to an upward-sloping aggregate-supply curve.

To see how this might work, suppose the overall price level falls below the level that suppliers expected. When suppliers see the prices of their products fall, they may mistakenly believe that their **relative** prices have fallen; that is, they may believe that their prices have fallen compared with other prices in the economy. For example, wheat farmers may notice a fall in the price of wheat before they notice a fall in the prices of the many items they buy as consumers. They may infer that the reward for producing wheat is temporarily low, and they may respond by reducing the quantity of wheat they supply. Similarly, workers may notice a fall in their nominal wages before they notice that the prices of the goods they buy are also falling. They may infer that the reward for working is temporarily low and respond by working less. In both cases, a lower price level causes misperceptions about relative prices, and these misperceptions induce suppliers to respond to the lower price level by reducing the quantity of goods and services supplied.

Similar misperceptions arise when the price level is above what was expected. Suppliers of goods and services may notice the price of their output rising and infer, mistakenly, that their relative prices are rising. They would conclude that it is a good time to produce more. Until their misperceptions are corrected, they respond to the higher price level by increasing the quantity of goods and services supplied. This behavior results in a short-run aggregate-supply curve that slopes upward.

Summing Up There are three explanations for the upward slope of the short-run aggregate-supply curve: (1) sticky wages, (2) sticky prices, and (3) misperceptions about relative prices. Economists debate which of these theories is correct, and it is possible that each contains an element of truth. For our purposes, the similarities of the theories are more important than the differences. All three suggest that output deviates in the short run from its natural level when the actual price level deviates from the price level that people had expected. We can express this mathematically as follows:

$$\begin{pmatrix}\text{Quantity}\\ \text{of output}\\ \text{supplied}\end{pmatrix} = \begin{pmatrix}\text{Natural}\\ \text{level of}\\ \text{output}\end{pmatrix} + a\left(\begin{pmatrix}\text{Actual}\\ \text{price}\\ \text{level}\end{pmatrix} - \begin{pmatrix}\text{Expected}\\ \text{price}\\ \text{level}\end{pmatrix}\right),$$

where a is a number that determines how much output responds to unexpected changes in the price level.

Notice that each of the three theories of short-run aggregate supply emphasizes a problem that is likely to be temporary. Whether the upward slope of the aggregate-supply curve is attributable to sticky wages, sticky prices, or misperceptions, these conditions will not last forever. Over time, nominal wages and prices will become unstuck, and misperceptions about relative prices will be corrected. It is reasonable to assume that, in the long run, wages and prices are flexible rather than sticky, and people accurately perceive relative prices. Each of the three theories explains not only a short-run aggregate-supply curve that slopes upward but also a long-run aggregate-supply curve that is vertical.

34-4e Why the Short-Run Aggregate-Supply Curve Might Shift

The short-run aggregate-supply curve depicts the quantity of goods and services supplied in the short run for any level of prices. It is similar to the long-run aggregate-supply curve, but it is upward-sloping rather than vertical because of sticky wages, sticky prices, and misperceptions. When thinking about what shifts the short-run aggregate-supply curve, we have to consider all the variables that shift the long-run aggregate-supply curve, plus an additional one: the expected price level. It influences the wages that are stuck, the prices that are stuck, and the perceptions about relative prices that may be flawed.

Let's start with what we know about the long-run aggregate-supply curve. As we have discussed, shifts in the long-run aggregate-supply curve normally arise from changes in labor, capital, natural resources, or technological knowledge. These variables also shift the short-run aggregate-supply curve. For example, when an increase in the capital stock raises productivity, the economy can produce more output, so both the long-run and short-run aggregate-supply curves shift to the right. When large sectoral shifts raise the natural rate of unemployment, the economy has fewer employed workers and produces less output, so both the long-run and short-run aggregate-supply curves shift to the left.

The new variable that affects the position of the short-run aggregate-supply curve is the price level that people expected to prevail. As we have discussed, the quantity of goods and services supplied depends, in the short run, on sticky wages, sticky prices, and misperceptions. Yet wages, prices, and perceptions are set based on the expected price level. So when the expected price level changes, the short-run aggregate-supply curve shifts.

For example, the sticky-wage theory says that workers and firms agree on a level of nominal wages based on their expectations about the price level. The expected price level, therefore, affects firms' costs and, for any actual price level, the quantity of goods and services supplied. When the expected price level rises, wages are set higher, costs increase, and firms produce a smaller quantity of goods and services at any actual price level. Thus, the short-run aggregate-supply curve shifts to the left. Conversely, when the expected price level falls, wages are set lower, costs decline, firms increase output at any actual price level, and the short-run aggregate-supply curve shifts to the right.

A similar logic applies in each of the theories of aggregate supply. The general lesson is the following: **An increase in the expected price level reduces the quantity of goods and services supplied and shifts the short-run aggregate-supply curve to the left. A decrease in the expected price level raises the quantity of goods and services supplied and shifts the short-run aggregate-supply curve to the right.**

As the next section shows, the influence of expectations on the position of the short-run aggregate-supply curve plays a key role in explaining how the economy makes the transition from the short run to the long run. In the short run, expectations are fixed, and the economy finds itself at the intersection of the aggregate-demand curve and the short-run aggregate-supply curve. Over time, if the price level differs from what people expected, expectations change, and the short-run aggregate-supply curve shifts. This shift ensures that, in the long run, the economy moves to the intersection of the aggregate-demand curve and the long-run aggregate-supply curve.

Table 2 summarizes what we've learned about the short-run aggregate-supply curve.

Table 2

The Short-Run Aggregate-Supply Curve: Summary

Why Does the Short-Run Aggregate-Supply Curve Slope Upward?

1. **The Sticky-Wage Theory:** An unexpectedly low price level, for given nominal wages, causes firms to hire fewer workers and produce a smaller quantity of goods and services.
2. **The Sticky-Price Theory:** An unexpectedly low price level leaves some firms with higher-than-desired prices, depressing their sales and leading them to cut back production.
3. **The Misperceptions Theory:** An unexpectedly low price level leads some suppliers to think their relative prices have fallen, inducing a fall in production.

Why Might the Short-Run Aggregate-Supply Curve Shift?

1. **Shifts Arising from Changes in Labor:** An increase in the quantity of labor available (perhaps due to a fall in the natural rate of unemployment) shifts the aggregate-supply curve to the right. A decrease in the quantity of labor available (perhaps due to a rise in the natural rate of unemployment) shifts the aggregate-supply curve to the left.
2. **Shifts Arising from Changes in Capital:** An increase in physical or human capital shifts the aggregate-supply curve to the right. A decrease in physical or human capital shifts the aggregate-supply curve to the left.
3. **Shifts Arising from Changes in Natural Resources:** An increase in the availability of natural resources shifts the aggregate-supply curve to the right. A decrease in the availability of natural resources shifts the aggregate-supply curve to the left.
4. **Shifts Arising from Changes in Technology:** An advance in technological knowledge shifts the aggregate-supply curve to the right. A decrease in the available technology (perhaps due to government regulation) shifts the aggregate-supply curve to the left.
5. **Shifts Arising from Changes in the Expected Price Level:** A decrease in the expected price level shifts the short-run aggregate-supply curve to the right. An increase in the expected price level shifts the short-run aggregate-supply curve to the left.

QuickQuiz

8. One reason the short-run aggregate-supply curve slopes upward is that a higher price level
   a. raises nominal wages if real wages are sticky.
   b. reduces nominal wages if real wages are sticky.
   c. raises real wages if nominal wages are sticky.
   d. reduces real wages if nominal wages are sticky.

9. A change in which of the following would shift the short-run aggregate-supply curve but not the long-run aggregate-supply curve?
   a. the labor force
   b. the capital stock
   c. the state of technology
   d. the expected price level

Answers are at the end of the chapter.

34-5 Two Causes of Economic Fluctuations

Let's now use the model of aggregate demand and aggregate supply to examine the two basic causes of short-run fluctuations: shifts in aggregate demand and shifts in aggregate supply.

To keep things simple, assume the economy begins in long-run equilibrium, as shown in Figure 7. Output and the price level are determined in the long run by the intersection of the aggregate-demand curve and the long-run aggregate-supply curve, shown as point A in the figure. At this point, output is at its natural level. Because the economy is always in a short-run equilibrium, the short-run aggregate-supply curve passes through this point as well, indicating that the expected price level has adjusted to this long-run equilibrium. That is, when an economy is in its long-run equilibrium, the expected price level equals the actual price level so the intersection of aggregate demand with short-run aggregate supply is the same as the intersection of aggregate demand with long-run aggregate supply.

34-5a The Effects of a Shift in Aggregate Demand

Suppose that a wave of pessimism overtakes the economy. The cause might be a scandal in the White House, a crash in the stock market, or the outbreak of war overseas. Whatever the specific reason, many people lose confidence in the future and alter their plans. Households cut back on their spending and delay major purchases, and firms put off buying new equipment.

What is the macroeconomic impact of such a wave of pessimism? To answer this question, let's follow the three steps introduced in Chapter 4 for analyzing supply and demand. First, determine whether the event affects aggregate demand or aggregate supply. Second, determine the direction in which the curve shifts. Third, use the diagram of aggregate demand and aggregate supply to compare the initial and new equilibria. The new wrinkle is that we need to add a fourth step: We have to keep track of a new short-run equilibrium, a new long-run equilibrium, and the transition between them. Table 3 summarizes the four steps for analyzing economic fluctuations.

Figure 7

The Long-Run Equilibrium

The long-run equilibrium of the economy is found where the aggregate-demand curve crosses the long-run aggregate-supply curve (point A). When the economy reaches this long-run equilibrium, the expected price level will have adjusted to equal the actual price level. As a result, the short-run aggregate-supply curve crosses this point as well.

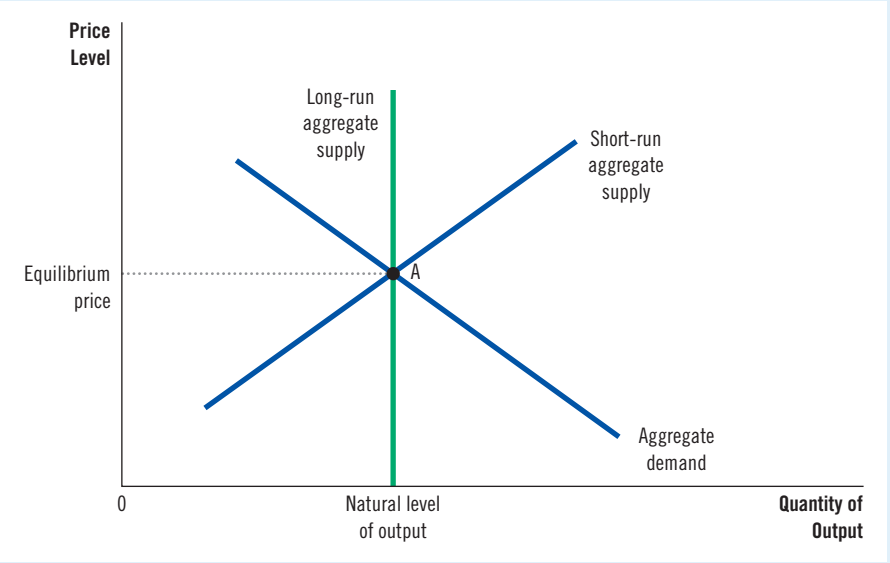


Table 3

Four Steps for Analyzing Macroeconomic Fluctuations

1. Decide whether the event shifts the aggregate-demand curve or the aggregate-supply curve (or perhaps both).
2. Decide the direction in which the curve shifts.
3. Use the diagram of aggregate demand and aggregate supply to determine the impact on output and the price level in the short run.
4. Use the diagram of aggregate demand and aggregate supply to analyze how the economy moves from its new short-run equilibrium to its new long-run equilibrium.

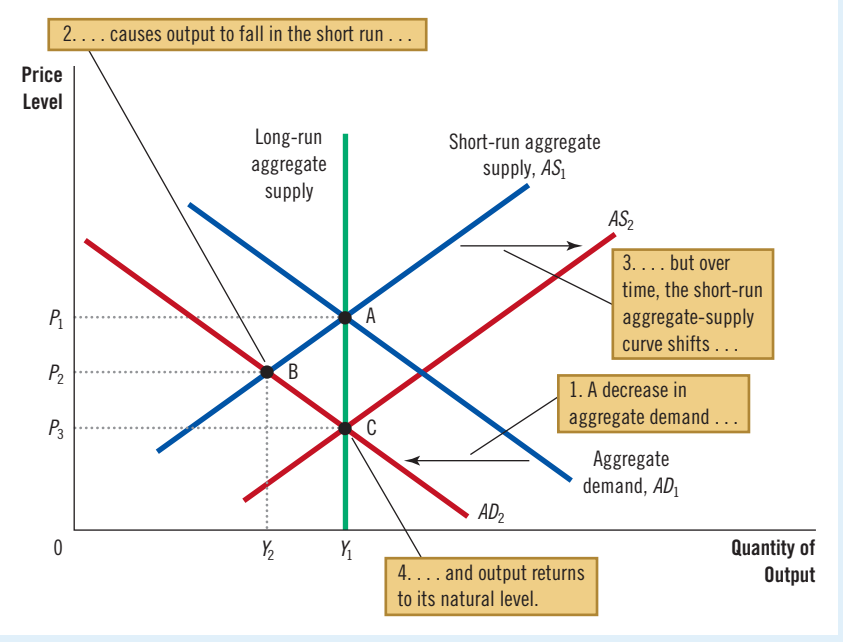
The first two steps are straightforward. First, because the wave of pessimism affects spending plans, it affects the aggregate-demand curve. Second, because households and firms now want to buy a smaller quantity of goods and services for any price level, the event reduces aggregate demand. As Figure 8 shows, the aggregate-demand curve shifts to the left from AD_1 to AD_2.

Using this figure, we can perform step three: By comparing the initial and new equilibria, we can see the effects of the fall in aggregate demand. In the short run, the economy moves along the initial short-run aggregate-supply curve, AS_1, going from point A to point B. As the economy moves between these two points, output falls from Y_1 to Y_2, and the price level falls from P_1 to P_2. The decline in output indicates that the economy is in a recession. Although not shown in the

Figure 8

A Contraction in Aggregate Demand

A fall in aggregate demand is represented by a leftward shift in the aggregate-demand curve from AD_1 to AD_2. In the short run, the economy moves from point A to point B. Output falls from Y_1 to Y_2, and the price level falls from P_1 to P_2. But as the expected price level adjusts, the short-run aggregate-supply curve shifts to the right from AS_1 to AS_2, and the economy reaches point C, where the new aggregate-demand curve crosses the long-run aggregate-supply curve. In the long run, the price level falls to P_3, and output returns to its natural level, Y_1.



figure, firms respond to lower sales and production by reducing employment. The pessimism that caused the shift in aggregate demand is, to some extent, self-fulfilling: Pessimism about the future leads to falling incomes and rising unemployment.

Now comes step four—the transition from the short-run equilibrium to the new long-run equilibrium. Because of the reduction in aggregate demand, the price level initially falls from P_1 to P_2. The price level is below the level that people were expecting (P_1) before the sudden fall in aggregate demand. People can be surprised in the short run, but they will not remain surprised forever. Eventually, their expectations catch up with the new reality. The expected price level falls, altering wages, prices, and perceptions, and these changes, in turn, affect the position of the short-run aggregate-supply curve. For example, according to the sticky-wage theory, once workers and firms come to expect a lower level of prices, they agree on lower nominal wages; the reduction in labor costs encourages firms to hire more workers and expand production at any level of prices. In this way, the fall in the expected price level shifts the short-run aggregate-supply curve to the right from AS_1 to AS_2 in Figure 8. This shift allows the economy to approach point C, where the new aggregate-demand curve (AD_2) crosses the long-run aggregate-supply curve.

In the new long-run equilibrium, point C, output is back to its natural level. The economy has corrected itself: The decline in output is reversed in the long run, even without action by policymakers. Although the wave of pessimism has reduced aggregate demand, the price level has fallen sufficiently (to P_3) to offset the shift in the aggregate-demand curve, and people have come to expect this new lower price level as well. In the long run, the shift in aggregate demand is reflected fully in the price level and not at all in the level of output. In other words, the long-run effect of a shift in aggregate demand is a nominal change (the price level is lower) but not a real change (output is the same).

What should policymakers do when faced with a sudden fall in aggregate demand? This analysis assumed they did nothing. But another possibility is that, as soon as the economy heads into recession (moving from point A to point B), policymakers take action to increase aggregate demand. As we noted earlier, an increase in government spending or an increase in the money supply would increase the quantity of goods and services demanded at any price and thereby shift the aggregate-demand curve to the right. If policymakers act with sufficient speed and precision, they can offset the initial shift in aggregate demand, return the aggregate-demand curve to AD_1, and bring the economy back to point A. If the policy is successful, the painful period of depressed output and employment can be reduced in length and severity. The next chapter discusses in more detail the ways in which monetary and fiscal policy influence aggregate demand, as well as some of the practical difficulties in using these policy instruments.

To sum up, this story about shifts in aggregate demand has three important lessons:

- In the short run, shifts in aggregate demand cause fluctuations in the economy's output of goods and services.
- In the long run, shifts in aggregate demand affect the overall price level but do not affect output.
- Because policymakers influence aggregate demand, they can potentially mitigate the severity of economic fluctuations.

FYI

Monetary Neutrality Revisited

According to classical economic theory, money is neutral. That is, changes in the quantity of money affect nominal variables such as the price level but not real variables such as output. Earlier in this chapter, we noted that most economists accept this conclusion as a description of how the economy works in the long run but not in the short run. Using the model of aggregate demand and aggregate supply, we can explain this conclusion more fully.

Suppose that the Fed reduces the quantity of money in the economy. What effect does this change have? As discussed, the money supply is one determinant of aggregate demand. The reduction in the money supply shifts the aggregate-demand curve to the left.

The analysis looks just like Figure 8. Even though the cause of the shift in aggregate demand is different, we would observe the same effects on output and the price level. In the short run, both output and the price level fall. The economy has a recession. But over time, the expected price level falls as well. Firms and workers respond to their new expectations by, for instance, agreeing to lower nominal wages. As they do so, the short-run aggregate-supply curve shifts to the right. Eventually, the economy finds itself back on the long-run aggregate-supply curve.

Figure 8 shows when money matters for real variables and when it does not. In the long run, money is neutral, as represented by the movement of the economy from point A to point C. But in the short run, a change in the money supply has real effects, as represented by the movement of the economy from point A to point B. An old saying summarizes the analysis: "Money is a veil, but when the veil flutters, real output sputters." ■

Case Study

Two Big Shifts in Aggregate Demand: The Great Depression and World War II

This chapter began by establishing three facts about economic fluctuations using data since 1972. Let's now take a longer look at U.S. economic history. Figure 9 shows data since 1900 on the percentage change in real GDP over the previous three years. In an average 3-year period, real GDP grows about 10 percent—a bit more than 3 percent per year. The business cycle, however, causes fluctuations around this average. Two episodes jump out from this figure as being particularly significant: the large drop in real GDP in the early 1930s and the large increase in the early 1940s. Shifts in aggregate demand caused both of these events.

The economic calamity of the early 1930s is called the **Great Depression**, and it is by far the largest economic downturn in U.S. history. Real GDP fell by 26 percent from 1929 to 1933, and unemployment rose from 3 percent to 25 percent. At the same time, the price level fell by 22 percent over these four years. Many other countries experienced similar declines in output and prices during this period.

Economic historians continue to debate the causes of the Great Depression, but most explanations center on a large decline in aggregate demand. What caused aggregate demand to contract? Here is where the disagreement arises.

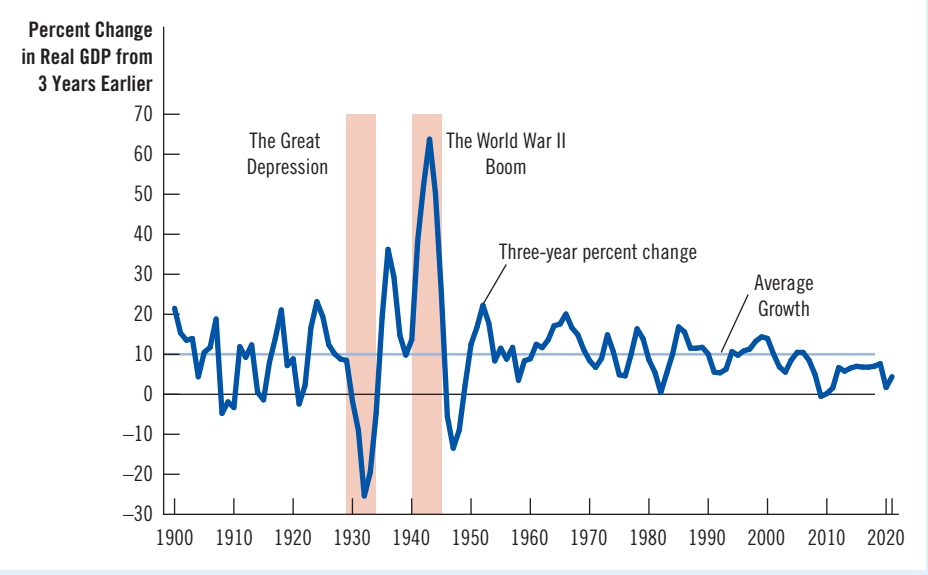
Many economists place primary blame on the decline in the money supply: From 1929 to 1933, it fell by 28 percent. As you may recall from our discussion of the monetary system, this decline in the money supply stemmed from problems in the banking system. As households withdrew their money from financially shaky banks and bankers became more cautious and started holding greater reserves, the process of money creation under fractional-reserve banking went into reverse. The Fed, meanwhile, failed to offset this fall in the money multiplier with expansionary open-market operations, so the money supply declined. Many economists blame the Fed's inaction for the Great Depression's severity.

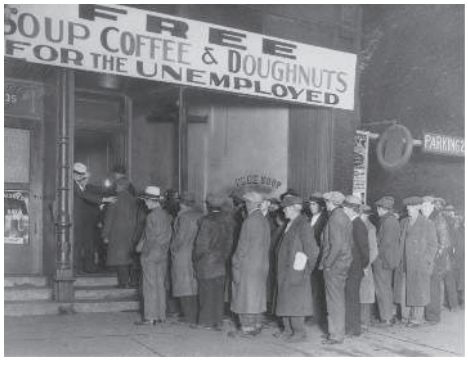
Figure 9

U.S. Real GDP Growth since 1900

Over the course of U.S. economic history, two fluctuations stand out as especially large. During the early 1930s, the economy endured the Great Depression, when the production of goods and services plummeted. During the early 1940s, the United States entered World War II, and production rose rapidly. Both events are usually explained by large shifts in aggregate demand.

Source: Louis D. Johnston and Samuel H. Williamson, "What Was GDP Then?" http://www.measuringworth.com/usgdp/; Department of Commerce.





BETTMANN/GETTY IMAGES

The outcome of a massive decrease in aggregate demand

Some economists have suggested other reasons for the collapse in aggregate demand. For example, stock prices fell about 90 percent during this period, depressing household wealth and consumer spending. In addition, the banking problems may have prevented some firms from obtaining the financing they wanted for new projects and business expansions, reducing investment spending. It is possible that all these forces acted together to contract aggregate demand.

The second significant episode in Figure 9—the rapid economic expansion of the early 1940s—is easier to explain. World War II caused the boom. As the United States entered the war overseas, the federal government devoted more resources to the military. Government purchases of goods and services increased almost fivefold from 1939 to 1944. This huge expansion in aggregate demand almost doubled the economy's production of goods and services and led to a 20 percent increase in the price level (although government price controls limited the rise in prices). Unemployment fell from 17 percent in 1939 to about 1 percent in 1944—the lowest level in U.S. history. ●

Case Study

The Great Recession of 2008–2009

In 2008 and 2009, the U.S. economy experienced a financial crisis and severe downturn in economic activity. In many ways, it was the worst macroeconomic event in more than half a century.

The story of this downturn begins a few years earlier with a boom in the housing market that was partly fueled by low interest rates. In the aftermath of the recession of 2001, the Fed dropped rates to historically low levels. These low interest rates helped the economy recover, but by making it less expensive to get a mortgage and buy a home, they contributed to a rise in house prices.

In addition to low interest rates, changes in the mortgage market made it easier for **subprime borrowers**—those with a higher risk of default based on their income and credit history—to get loans to buy homes. One development was **securitization**, the process by which a financial institution (specifically, a mortgage originator) makes loans and then (with the help of an investment bank) bundles them together into financial instruments called mortgage-backed securities. These **mortgage-backed securities** were then sold to other institutions (such as banks and insurance companies), which may not have fully appreciated the risks in these securities. Some economists blame inadequate regulation for these high-risk loans. Others blame government policies that encouraged this lending to make homeownership more attainable for low-income families. Together, these forces drove up housing demand and house prices. From 1995 to 2006, average house prices in the United States more than doubled.

Those prices proved unsustainable. From 2006 to 2009, house prices nationwide fell about 30 percent. Such price fluctuations are not necessarily a problem in a market economy. After all, price movements are how markets equilibrate supply and demand. In this case, however, the price decline had two repercussions that led to a large fall in aggregate demand.

The first was a rise in mortgage defaults and home foreclosures. During the housing boom, many people bought their homes with mostly borrowed money and minimal down payments. When house prices declined, these homeowners were **underwater** (they owed more on their mortgages than their homes were worth). Many of these homeowners stopped repaying their loans. The banks servicing the mortgages responded to these defaults by taking the houses away in foreclosure procedures and then selling them off. The banks' goal was to recoup whatever they could from the bad loans. But the increase in the supply of houses for sale exacerbated the downward spiral of house prices. As house prices fell, spending on residential construction collapsed.

A second repercussion was that financial institutions that owned mortgage-backed securities suffered large losses. In essence, by borrowing large sums to buy high-risk mortgages, these companies had bet that house prices would keep rising; when this bet turned bad, they found themselves at or near bankruptcy. Because of these losses, many financial institutions did not have funds to loan out, and the ability of the financial system to channel resources to those who could best use them was impaired. Even creditworthy customers found themselves unable to borrow to finance investment spending. Such an event is called a **credit crunch**.

As a result of the residential investment collapse and credit crunch, the economy experienced a contractionary shift in aggregate demand. Real GDP and employment both fell sharply. The figures cited in this chapter's introduction are worth repeating: Real GDP declined by 4.0 percent between the fourth quarter of 2007 and the second quarter of 2009, and the rate of unemployment rose from 4.4 percent in May 2007 to 10.0 percent in October 2009. This experience served as a vivid reminder that deep economic downturns and the personal hardship they cause are not relics of history but a constant risk in the modern economy.

As the crisis unfolded, the U.S. government responded in various ways. Three policy actions—all aimed in part at returning aggregate demand to its previous level—are most noteworthy.

First, the Fed cut its target for the federal funds rate from 5.25 percent in September 2007 to about zero in December 2008. In addition, in a policy called **quantitative easing**, the Fed started buying mortgage-backed securities and other long-term debt in open-market operations. The goals of quantitative easing were to lower long-term interest rates and to provide the financial system with additional funds so banks would make loans more readily available.

Second, in an even more unusual move in October 2008, Congress appropriated $700 billion for the Treasury to use to rescue the financial system. Much of this money was injected into banks as capital. That is, the Treasury put funds into the banking system, which the banks could use to make loans and otherwise continue normal operations; in exchange for these funds, the U.S. government became a part owner of these banks, at least temporarily. The policy's goal was to stem the crisis on Wall Street and make it easier for businesses and individuals to borrow.

Finally, when Barack Obama became president in January 2009, his first major initiative was a large increase in government spending. After a brief congressional debate, he signed a $787 billion stimulus bill on February 17, 2009.

The recovery from this recession began in June 2009, but it was meager by historical standards. Over the next seven years, real GDP growth averaged only 2.2 percent per year, well below the average rate of growth over the past half century of about 3 percent. The unemployment rate did not fall below 5.0 percent until 2016.

Which of the policy moves were most important for ending the recession? What other policies might have promoted a more robust recovery? Macroeconomic historians continue to debate these questions. ●

34-5b The Effects of a Shift in Aggregate Supply

Imagine once again an economy in long-run equilibrium. Now suppose that the costs of production suddenly increase for some firms. For example, bad weather in farm states might destroy crops, driving up the cost of food products. Or a war in the Middle East might interrupt the shipping of crude oil, driving up the cost of oil-intensive goods.

To analyze the macroeconomic impact of such an increase in production costs, follow the same four steps as always. First, which curve is affected? Because production costs affect the firms that supply goods and services, changes in production costs alter the position of the aggregate-supply curve. Second, in which direction does the curve shift? Because higher production costs make selling goods and services less profitable, firms now supply a smaller quantity of output for any price level. As Figure 10 shows, the short-run aggregate-supply curve shifts to the left, from AS_1 to AS_2. (Depending on the event, the long-run aggregate-supply curve might also shift. To keep things simple, however, we will assume that it does not.)

The figure allows us to perform step three of comparing the initial and new equilibria. In the short run, the economy goes from point A to point B, moving along the existing aggregate-demand curve. The output of the economy falls from Y_1 to Y_2, and the price level rises from P_1 to P_2. Because the economy is experiencing both **stagnation** (falling output) and **inflation** (rising prices), such an event is sometimes called **stagflation**.

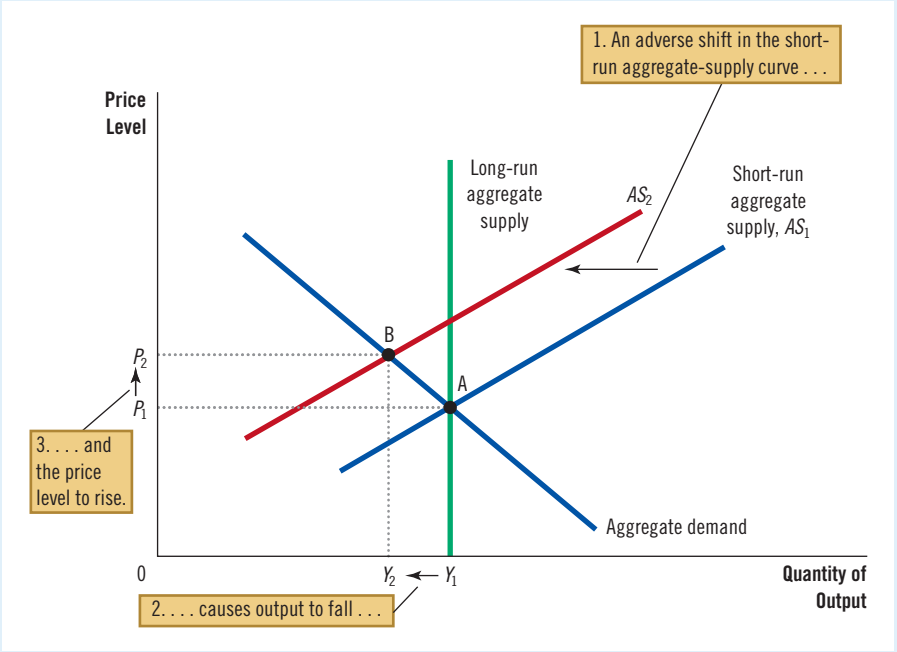
stagflation
a period of falling output and rising prices

Now consider step four—the transition from the short-run equilibrium to the long-run equilibrium. According to the sticky-wage theory, the key issue is how stagflation affects nominal wages. Firms and workers may, at first, respond to the higher level of prices by raising their expectations of the price level and

Figure 10

An Adverse Shift in Aggregate Supply

When some event increases firms' costs, the short-run aggregate-supply curve shifts to the left from AS_1 to AS_2. The economy moves from point A to point B. The result is stagflation: Output falls from Y_1 to Y_2, and the price level rises from P_1 to P_2.



setting higher nominal wages. In this case, firms' costs will rise yet again, and the short-run aggregate-supply curve will shift farther to the left, making the problem of stagflation even worse. This phenomenon of higher prices leading to higher wages, which in turn leads to even higher prices, is sometimes called a **wage-price spiral**.

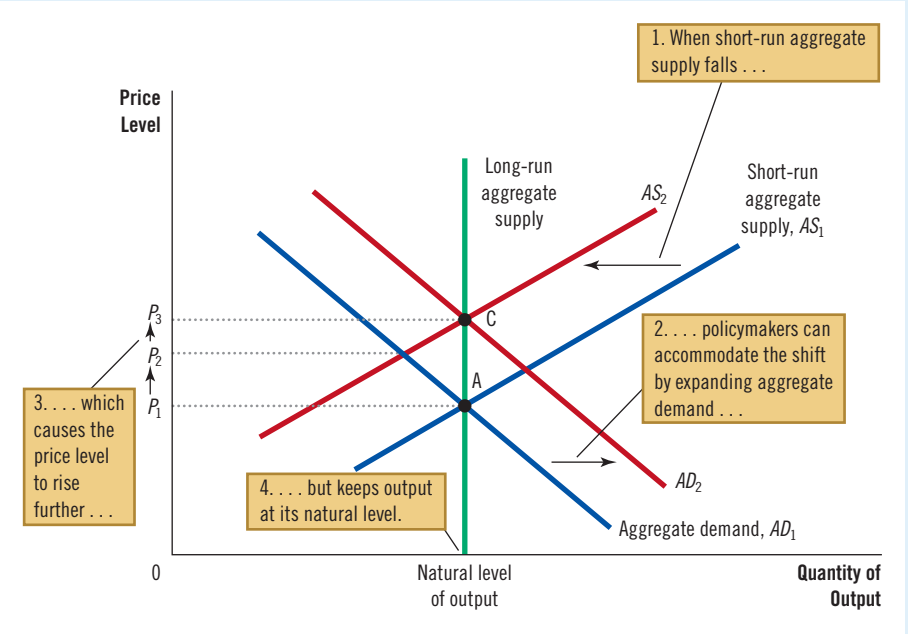
At some point, the spiral of ever-rising wages and prices will slow. The low level of output and employment will put downward pressure on wages because workers have less bargaining power when unemployment is high. As nominal wages fall, producing goods and services becomes more profitable, and the short-run aggregate-supply curve shifts to the right. As it shifts back toward AS_1, the price level falls, and the quantity of output approaches its natural level. In the long run, the economy returns to point A, where the aggregate-demand curve crosses the long-run aggregate-supply curve.

This transition back to the initial equilibrium assumes, however, that aggregate demand is held constant throughout the process. In the real world, that may not be the case. Monetary and fiscal policymakers might attempt to offset some of the effects of the shift in the short-run aggregate-supply curve by shifting the aggregate-demand curve. This possibility is shown in Figure 11. In this case, changes in policy shift the aggregate-demand curve to the right, from AD_1 to AD_2—exactly enough to prevent the shift in aggregate supply from affecting output. The economy moves directly from point A to point C. Output remains at its natural level, and the price level rises from P_1 to P_3. In this case, policymakers are said to **accommodate** the shift in aggregate supply. An accommodative policy accepts a permanently higher level of prices to maintain a higher level of output and employment.

Figure 11

Accommodating an Adverse Shift in Aggregate Supply

Faced with an adverse shift in aggregate supply from AS_1 to AS_2, policymakers who can influence aggregate demand might try to shift the aggregate-demand curve to the right from AD_1 to AD_2. The economy would move from point A to point C. This policy would prevent the supply shift from reducing output in the short run, but the price level would permanently rise from P_1 to P_3.



To sum up, this story about shifts in aggregate supply has two important lessons:

- Shifts in aggregate supply can cause stagflation—a combination of recession (falling output) and inflation (rising prices).
- Policymakers who can influence aggregate demand can mitigate the adverse impact on output but only at the cost of exacerbating the problem of inflation.

Case Study

Oil and the Economy

Some of the largest economic fluctuations in the U.S. economy since 1970 originated in the oil fields of the Middle East. Crude oil is an input into the production of many goods and services, and much of the world's oil comes from Saudi Arabia, Kuwait, and other Middle Eastern countries. When some event (often political in origin) reduces the supply of crude oil flowing from this region, the price of oil rises around the world. Firms in the United States that produce gasoline, tires, and many other products experience rising costs, so they find it less profitable to supply their output of goods and services at any price level. The result is a leftward shift in the aggregate-supply curve, which, in turn, leads to stagflation.

The first episode of this sort occurred in the mid-1970s. The countries with large oil reserves started to exert their influence on the world economy as members of OPEC, the Organization of the Petroleum Exporting Countries. OPEC is a **cartel**—a group of sellers that attempts to thwart competition and reduce production to raise prices. And indeed, oil prices rose substantially. From 1973 to 1975, oil approximately doubled in price. Oil-importing countries around the world experienced simultaneous inflation and recession. The U.S. inflation rate as measured by the CPI exceeded 10 percent for the first time in decades. Unemployment rose from 4.9 percent in 1973 to 8.5 percent in 1975.

Almost the same thing happened a few years later. In the late 1970s, the OPEC countries again restricted the supply of oil. From 1978 to 1981, the price of oil more than doubled. Once again, the result was stagflation. Inflation, which had subsided somewhat after the first OPEC event, again rose above 10 percent per year. But because the Fed was not willing to accommodate such a large rise in inflation, a recession soon followed. Unemployment rose from about 6 percent in 1978 and 1979 to about 10 percent a few years later.

Developments in the world market for oil can also be a source of favorable shifts in aggregate supply. In 1986, squabbling broke out among members of OPEC. Member countries reneged on their agreements to restrict oil production. In the world market for crude oil, prices fell by about half. This fall in oil prices reduced costs to U.S. firms, which now found it more profitable to supply goods and services at any price level. As a result, the aggregate-supply curve shifted to the right. The U.S. economy experienced the opposite of stagflation: Output grew rapidly, unemployment fell, and the inflation rate reached its lowest level in many years.

In recent years, developments in the world oil market have not been as important a source of fluctuations for the U.S. economy. One reason is that conservation efforts, advancing technology, and the increased availability of alternative energy sources have reduced the economy's dependence on oil. The amount of oil used to produce a unit of real GDP has declined by more than 60 percent since the OPEC shocks of the 1970s. As a result, the impact of any change in oil prices on the U.S. economy is smaller today than it was in the past. ●

YASSER AL-ZAYYAT/AFP/GETTY IMAGES

Changes in Middle East oil production are one source of U.S. economic fluctuations.

FYI

The Origins of the Model of Aggregate Demand and Aggregate Supply

Now that we have seen the model of aggregate demand and aggregate supply, let's step back and consider its history. How did this model of short-run fluctuations develop? The answer is that this model is largely a by-product of the Great Depression of the 1930s. Economists and policymakers at the time were puzzled about what had caused this calamity and were uncertain about how to deal with it.

KEYSTONE/HULTON ARCHIVE/GETTY IMAGES

John Maynard Keynes

In 1936, the economist John Maynard Keynes published *The General Theory of Employment, Interest, and Money*, a landmark book that attempted to explain short-run economic fluctuations in general and the Great Depression in particular. Keynes's main message was that recessions and depressions can occur because of inadequate aggregate demand for goods and services.

Keynes had long been a critic of classical economic theory because it could explain only the long-run effects of policies. A few years before offering *The General Theory*, Keynes had written the following about classical economics:

> The long run is a misleading guide to current affairs. In the long run we are all dead. Economists set themselves too easy, too useless a task if in tempestuous seasons they can only tell us when the storm is long past, the ocean will be flat.

Keynes's message was aimed at policymakers as well as economists. As the world's economies suffered from high unemployment, Keynes advocated policies to increase aggregate demand, including government spending on public works.

The next chapter examines in detail how policymakers can use the tools of monetary and fiscal policy to influence aggregate demand. The analysis in that chapter, as well as in this one, owes much to the legacy of John Maynard Keynes. ■

Case Study

The Covid Recession of 2020

In 2020, the U.S. economy and most others around the world experienced a downturn with three unusual features.

The first was its cause: the Covid-19 pandemic. This infectious and dangerous virus initially appeared in China in late 2019 and then in the United States in early 2020. To slow its spread, health experts advised people to avoid close interactions with others. Elected leaders ordered large segments of the economy to be closed, including movie theaters, sporting events, concerts, restaurants (except for take-out), and non-essential retail stores. Commercial air travel stopped almost completely.

The second unusual feature of the 2020 downturn was its speed and depth. From February 2020 to April 2020, employment in the United States fell from 61.1 percent of the adult population to 51.3 percent—by far the largest two-month drop ever recorded. The unemployment rate in April 2020 was 14.8 percent, the highest level since the Great Depression.

The third unusual feature of this downturn was that, in a sense, it was intentional. Most recessions are accidents: Some unexpected event shifts aggregate

In the News

The Strange Downturn of 2020

The economic events that resulted from the Covid-19 pandemic were not at all typical.

The Unusual Covid Recession

By Austan Goolsbee

Americans everywhere remain concerned about getting past what feels like a never-ending litany of Covid-19-induced economic problems, from out-of-stock products to inflation and continued fears of exposure to illness when going out. When will it finally end?

Professional economic forecasters are struggling to answer this question.

Many are turning to previous recessions as a guide to how things will go. But one of the most important things to understand is that while the pandemic created a collapse—and for some even an economic disaster—it really wasn't a recession in the normal sense.

That sounds strange. Indeed, the arbiter of these things, the National Bureau of Economic Research, declared that the United States had a two-month recession in March and April 2020.

But past business cycles look nothing like what the United States has gone through in the pandemic, so they are the wrong place to find lessons for where things are going now.

The causes of recessions vary, but they follow a basic pattern: The hardest-hit industries are the cyclically sensitive sectors where demand dries up. Those sectors include sales of big-ticket items like furniture, construction materials, appliances and cars, as the Bureau of Labor Statistics and others have documented. These are purchases that can be delayed when times are bad. Recoveries begin when demand returns to these cyclical industries, when prices fall enough or interest rates get cut enough or pent-up needs build sufficiently for demand to return.

Recessions have much smaller impacts on noncyclical industries like hospitals, nursing care, gas and electric utilities, and the like. Demand there is steady regardless of the cycle. Some service sector industries, like education, see demand rise in recessions.

None of these familiar patterns held during the pandemic economic collapse. Spending on consumer durables went up. Indeed, sales of TVs with screens larger than 65 inches rose 77 percent from April to June 2020, compared to the year before, as the bottom dropped out of the economy. Watching TV was one of the few things people could still do during lockdown. Demand for other cyclical industry goods like housing and construction materials boomed too.

During the pandemic downturn, Americans also reversed a decades-long trend toward spending on services rather than goods. For 75 years, consumers in the United States have been spending less and less of their money on physical goods (from 60 percent of spending in the 1940s to 31 percent in 2019). Counter to this trend (and contrasting with previous

supply or aggregate demand, reducing production and employment. When this occurs, policymakers usually want to return the economy to normal levels of production and employment as quickly as possible. By contrast, the 2020 downturn was a recession by design. To curb the Covid-19 pandemic, policymakers compelled changes in behavior that reduced production and employment. Of course, the pandemic itself was neither intended nor desired. But given the circumstances, a large, temporary decline in economic activity was arguably the best possible outcome.

The economic downturn of 2020 can be interpreted using the model of aggregate supply and aggregate demand. Consider first aggregate demand. Starting in March 2020, many places where people normally buy things, such as restaurants and retail stores, were closed by government decree. And people avoided many businesses that remained open to reduce the risk of infection. As a result, the quantity of goods and services demanded was lower at every price level. The aggregate demand curve shifted to the left.

Now consider aggregate supply. When the health crisis caused many businesses to temporarily shut down, it caused a sudden, massive reduction in the quantity of goods and services supplied at every price level. The aggregate supply curve

recessions), the share of consumer spending on physical goods actually jumped during the pandemic to the highest level in 17 years and among the biggest jumps ever recorded.

In other words, this was a recession like no other in recent memory. The pandemic downturn was driven by all those industries that are supposed to be recession-proof—trips to the dentist, electricity usage in offices and malls, and so on. And the normally *counter*cyclical education sector had big enrollment drops despite the bad economy.

Of course, this was because of the coronavirus. But it means that the recovery from past recessions doesn't really say much about how the recovery will go now. Everyone is trying to predict when there will be a rebound in service sector industries that normally don't decline, like health care, child care and education. That's really more of a question about how quickly we can control the spread of the virus than it is about recession fundamentals.

At the same time, the unusually large demand for physical goods in the United States and other rich countries has exceeded supply, driving up inflation and leading to shortages.

So the most important thing to watch if you want to understand the economy is, as has been the case for a year and half now, the progress made against the virus. Related, and also worth watching, is how much Americans spend on goods relative to services. (It was 31 percent in 2019 and has risen to 35 percent now.)

While economic growth in the United States was disappointing in the third quarter of 2021, it could easily turn around if coronavirus case numbers improve. The U.S. employment numbers released on Friday were encouraging. New Covid-19 cases are down significantly and millions of children are now eligible for vaccination, which could reduce infection rates even further.

Looking beyond the coming months, though, the most interesting questions aren't really about recession and recovery. They center on whether any of the pandemic changes will last. Some companies, for example, are now trying to hold more inventory and keep their supply chains local to avoid disruption. Many people are working partly from home and some have moved to the exurbs. But how long before they rediscover why we ended up with lean manufacturing and a global supply chain in the first place? And Americans are already moving back to cities.

My view is that reversals of longstanding economic trends are not likely to become permanent. Once the economic memory of the pandemic has faded, the old lessons from the regular business cycle will probably become relevant once more. Until that happens, though, best to get in line for a vaccine booster and keep your eye on the case numbers. ■

Questions to Discuss

1. How did your family's spending change during the coronavirus pandemic?
2. In your opinion, what economic policies should the government have pursued during the pandemic?

Mr. Goolsbee is a professor of economics at the University of Chicago.

Source: *New York Times*, November 10, 2021.

shifted to the left. The simultaneous shifts in aggregate demand and aggregate supply led to a sharp reduction in production and employment.

Once the enormity of the downturn became clear, policymakers responded swiftly. On March 27, 2020, the Coronavirus Aid, Relief, and Economic Security (CARES) Act was signed into law. Together with other legislation enacted around the same time, it authorized a combination of spending increases and tax reductions of about $2 trillion, roughly 10 percent of GDP, making it the largest fiscal response to a recession in history. The CARES Act is sometimes called a stimulus bill, but the goal was not actually to end the recession by stimulating the economy. The recession was inevitable, given the pandemic. The policy's goal was to alleviate the hardship people faced and to prevent the downturn from leaving permanent scars on the economy.

A large part of the policy response might be called social insurance or disaster relief. All households, except those with high incomes, were given tax rebates of $1,200 per adult and $500 per child. Eligibility for unemployment insurance was expanded, and benefits were temporarily increased by $600 per week. Small businesses were offered loans that would be forgiven and turned into grants if they did not lay off any workers for the next two months.

To prevent permanent damage from the recession, the CARES Act had various provisions to promote business continuity. This was part of the motivation for the forgivable loans to small businesses. Not only did workers continue getting paychecks, but they stayed connected to their employers, so normal business could quickly resume when the crisis passed. The CARES Act also provided funds that enabled the Federal Reserve, working with the Treasury, to lend to larger businesses, states, and municipalities, expanding the Fed's role as lender of last resort. At the same time, the Fed cut its target for the federal funds rate to near zero.

In the United States, the number of daily deaths from Covid-19 peaked in late April 2020 and then started to slowly decline. By June 2020, many restrictions on economic activity were relaxed, leading to a quick economic rebound. The unemployment rate, after increasing from 3.5 percent in February 2020 to 14.8 percent in April 2020, declined to 6.9 percent in October 2020.

The pandemic, however, was not over. Cases surged in January 2021 and then again in October 2021 and January 2022.

As the pandemic persisted, subsequent legislation continued and expanded the relief offered by the CARES Act. President Trump signed a $900 billion relief package in December 2020, and President Biden signed a $1.9 trillion package in March 2021.

Some economists—most notably the former Treasury Secretary Lawrence Summers—suggested that the fiscal response was excessive. They worried that, together with supply-chain disruptions from the pandemic, these monetary and fiscal policies might lead to excessive inflation. As measured by the CPI, the 12-month inflation rate rose to 7.5 percent in January 2022—the highest level in 40 years. Policymakers initially believed that this inflation surge would prove temporary. But in March 2022, the Fed started raising interest rates to contain inflationary pressures.

The eventual solution to this economic downturn came more from microbiology than from macroeconomics. Economic activity started to return to normal after several vaccines were developed and distributed in 2021. But the hesitancy of large segments of the U.S. population to get vaccinated, along with the appearance of the new Omicron variant of the virus, delayed the pandemic's end and slowed the pace of economic recovery. ●

QuickQuiz

10. A sudden increase in business pessimism shifts the aggregate-_____ curve, leading to _____ output.
    a. supply; lower
    b. supply; higher
    c. demand; lower
    d. demand; higher
11. An increase in the aggregate demand for goods and services has a larger impact on output _____ and a larger impact on the price level _____.
    a. in the short run; in the long run
    b. in the long run; in the short run
    c. in the short run; also in the short run
    d. in the long run; also in the long run
12. Stagflation is caused by a
    a. leftward shift in the aggregate-demand curve.
    b. rightward shift in the aggregate-demand curve.
    c. leftward shift in the aggregate-supply curve.
    d. rightward shift in the aggregate-supply curve.

Answers are at the end of the chapter.

34-6 Conclusion

This chapter had two goals. First, it discussed some of the important facts about short-run fluctuations in economic activity. Second, it introduced a basic model to explain those fluctuations, called the model of aggregate demand and aggregate supply. We continue our study of this model in the next chapter to understand more fully what causes economic fluctuations and how policymakers might respond to them.

Chapter in a Nutshell

- All societies experience short-run economic fluctuations around long-run trends. These fluctuations are irregular and largely unpredictable. When recessions occur, real GDP and other measures of income, spending, and production fall, while unemployment rises.
- Classical economic theory is based on the assumption that nominal variables such as the money supply and the price level don't influence real variables such as output and employment. Most economists believe that this assumption is accurate in the long run but not in the short run. Economists analyze short-run economic fluctuations using the model of aggregate demand and aggregate supply. According to this model, the output of goods and services and the overall level of prices adjust to balance aggregate demand and aggregate supply.
- The aggregate-demand curve slopes downward for three reasons. The first is the wealth effect: A lower price level raises the real value of households' money holdings, stimulating consumer spending. The second is the interest-rate effect: A lower price level reduces the quantity of money households demand. As households try to convert money into interest-bearing assets, interest rates fall, stimulating investment spending. The third is the exchange-rate effect: As a lower price level reduces interest rates, the dollar depreciates in the market for foreign-currency exchange, stimulating net exports.
- Any event or policy that raises consumption, investment, government purchases, or net exports at any price level increases aggregate demand. Any event or policy that reduces consumption, investment, government purchases, or net exports at any price level decreases aggregate demand.
- The long-run aggregate-supply curve is vertical. In the long run, the quantity of goods and services supplied depends on the economy's labor, capital, natural resources, and technology but not on the overall level of prices.
- Three theories have been proposed to explain the upward slope of the short-run aggregate-supply curve. According to the sticky-wage theory, an unexpected fall in the price level, for given nominal wages, induces firms to reduce employment and production.

According to the sticky-price theory, an unexpected fall in the price level leaves some firms with prices that are temporarily too high, reducing their sales and causing them to cut back production. According to the misperceptions theory, an unexpected fall in the price level leads suppliers to mistakenly believe that their relative prices have fallen, inducing them to reduce production. All three theories imply that output deviates from its natural level when the actual price level deviates from the price level that people expected.

- Events that alter the economy's ability to produce output, such as changes in labor, capital, natural resources, or technology, shift the short-run aggregate-supply curve (and may shift the long-run aggregate-supply curve as well). In addition, the position of the short-run aggregate-supply curve depends on the expected price level.
- One possible cause of economic fluctuations is a shift in aggregate demand. When the aggregate-demand curve shifts to the left, for instance, output and prices fall in the short run. Over time, as a change in the expected price level causes wages, prices, and perceptions to adjust, the short-run aggregate-supply curve shifts to the right. This shift returns the economy to its natural level of output at a new, lower price level.
- A second possible cause of economic fluctuations is a shift in aggregate supply. When the short-run aggregate-supply curve shifts to the left, the effect is falling output and rising prices—a combination called stagflation. Over time, as wages, prices, and perceptions adjust, the short-run aggregate-supply curve shifts back to the right, returning the price level and output to their original levels.

Key Concepts

recession, p. 710
depression, p. 710
model of aggregate demand and aggregate supply, p. 714
aggregate-demand curve, p. 714
aggregate-supply curve, p. 714
natural level of output, p. 722
stagflation, p. 736

Questions for Review

1. Name two macroeconomic variables that decline when the economy goes into a recession. Name one macroeconomic variable that rises during a recession.
2. Draw a diagram showing aggregate demand, short-run aggregate supply, and long-run aggregate supply. Be careful to label the axes correctly.
3. List and explain the three reasons the aggregate-demand curve slopes downward.
4. Explain why the long-run aggregate-supply curve is vertical.
5. List and explain the three theories for why the short-run aggregate-supply curve slopes upward.
6. What might shift the aggregate-demand curve to the left? Use the model of aggregate demand and aggregate supply to trace the short-run and long-run effects of such a shift on output and the price level.
7. What might shift the aggregate-supply curve to the left? Use the model of aggregate demand and aggregate supply to trace the short-run and long-run effects of such a shift on output and the price level.

Problems and Applications

1. Suppose the economy is in a long-run equilibrium.
   a. Draw a diagram to illustrate the state of the economy. Be sure to show aggregate demand, short-run aggregate supply, and long-run aggregate supply.
   b. Now suppose that a stock market crash causes aggregate demand to fall. Use your diagram to show what happens to output and the price level in the short run. What happens to the unemployment rate?
   c. Use the sticky-wage theory of aggregate supply to explain what happens to output and the price level in the long run (assuming no change in policy). What role does the expected price level play in this adjustment? Be sure to illustrate your analysis in a graph.

2. Explain whether each of the following events increases, decreases, or has no effect on long-run aggregate supply.
   a. The United States experiences a wave of immigration.
   b. Congress raises the minimum wage to $15 per hour.
   c. Intel invents a new and more powerful computer chip.
   d. A severe hurricane damages factories along the East Coast.
3. Suppose an economy is in long-run equilibrium.
   a. Use the model of aggregate demand and aggregate supply to illustrate the initial equilibrium (call it point A). Be sure to include both short-run aggregate supply and long-run aggregate supply.
   b. The central bank raises the money supply by 5 percent. Use your diagram to show what happens to output and the price level as the economy moves from the initial equilibrium to the new short-run equilibrium (call it point B).
   c. Now show the new long-run equilibrium (call it point C). What causes the economy to move from point B to point C?
   d. According to the sticky-wage theory of aggregate supply, how do nominal wages at point A compare with nominal wages at point B? How do nominal wages at point A compare with nominal wages at point C?
   e. According to the sticky-wage theory of aggregate supply, how do real wages at point A compare with real wages at point B? How do real wages at point A compare with real wages at point C?
   f. Judging by the impact of the money supply on nominal and real wages, is this analysis consistent with the proposition that money has real effects in the short run but is neutral in the long run?
4. In 1939, with the U.S. economy not yet fully recovered from the Great Depression, President Franklin Roosevelt proclaimed that Thanksgiving would fall a week earlier than usual so that the shopping period before Christmas would be longer. (The policy was dubbed "Franksgiving.") Explain what President Roosevelt might have been trying to achieve, using the model of aggregate demand and aggregate supply.
5. Explain why the following statements are false.
   a. "The aggregate-demand curve slopes downward because it is the horizontal sum of the demand curves for individual goods."
   b. "The long-run aggregate-supply curve is vertical because economic forces do not affect long-run aggregate supply."
   c. "If firms adjusted their prices every day, then the short-run aggregate-supply curve would be horizontal."
   d. "Whenever the economy enters a recession, its long-run aggregate-supply curve shifts to the left."
6. For each of the three theories for the upward slope of the short-run aggregate-supply curve, carefully explain the following:
   a. how the economy recovers from a recession and returns to its long-run equilibrium without any policy intervention
   b. what determines the speed of that recovery
7. The economy begins in long-run equilibrium. Then, one day, the president appoints a new Fed chair, who is well known for holding the view that inflation is not a major problem for an economy.
   a. How would this news affect the price level that people expect to prevail?
   b. How would this change in the expected price level affect the nominal wage that workers and firms agree to in their new labor contracts?
   c. How would this change in the nominal wage affect the profitability of producing goods and services at any price level?
   d. How would this change in profitability affect the short-run aggregate-supply curve?
   e. If aggregate demand is held constant, how would this shift in the aggregate-supply curve affect the price level and the quantity of output produced?
   f. Do you think appointing this Fed chair was a good decision?
8. Explain whether each of the following events shifts the short-run aggregate-supply curve, the aggregate-demand curve, both, or neither. For each event that does shift a curve, draw a diagram to illustrate the effect on the economy.
   a. Households decide to save a larger share of their income.
   b. Florida orange groves suffer a prolonged period of below-freezing temperatures.
   c. Increased job opportunities overseas cause many people to leave the country.
9. For each of the following events, explain the short-run and long-run effects on output and the price level, assuming policymakers take no action.
   a. The stock market declines sharply, reducing consumers' wealth.
   b. The federal government increases spending on national defense.
   c. A technological improvement raises productivity.
   d. A recession overseas causes foreigners to buy fewer U.S. goods.

10. Suppose firms become optimistic about future business conditions and invest heavily in new capital equipment.
    a. Draw an aggregate-demand/aggregate-supply diagram to show the short-run effect of this optimism on the economy. Label the new levels of prices and real output. Explain in words why the aggregate quantity of output **supplied** changes.
    b. Now use the diagram from part (a) to show the new long-run equilibrium of the economy. (For now, assume there is no change in the long-run aggregate-supply curve.) Explain in words why the aggregate quantity of output **demanded** changes between the short run and the long run.
    c. How might the investment boom affect the long-run aggregate-supply curve? Explain.

QuickQuiz Answers

1. **c** 2. **d** 3. **c** 4. **a** 5. **b** 6. **b** 7. **d** 8. **d** 9. **d** 10. **c** 11. **a** 12. **c**

Chapter

35

The Influence of Monetary and Fiscal Policy on Aggregate Demand

Imagine that you are a member of the Federal Open Market Committee, the group at the Federal Reserve that sets monetary policy. You observe that the president and Congress are raising taxes to reduce the budget deficit. How should the Fed respond to this change in fiscal policy? Should monetary policy be expansionary, contractionary, or unchanged?

To answer this question, we will need to consider the impact of monetary and fiscal policy on the economy. The preceding chapter used the model of aggregate demand and aggregate supply to explain short-run economic fluctuations. We saw that shifts in the aggregate-demand curve or the aggregate-supply curve cause fluctuations in the overall output of goods and services and the level of prices. As we noted, both monetary and fiscal policy influence aggregate demand, so a change in one of these policies can lead to short-run fluctuations in output and prices. Policymakers may want to adjust the other policy in response.

This chapter examines in greater detail how government policy tools influence the position of the aggregate-demand curve. These tools include monetary policy (the open-market operations conducted by the central bank and the interest rate it pays on reserves) and fiscal policy (the levels of government spending and taxation set by the president and Congress). We have already discussed the long-run effects of these policies. Chapters 26 and 27 showed how

AZMANJAKA/E+/GETTY IMAGES; CARLOS FLORES/EYEEM/GETTY IMAGES

fiscal policy affects saving, investment, and long-run economic growth. Chapters 30 and 31 examined how monetary policy influences the price level in the long run. Now we will see how these policy tools can shift the aggregate-demand curve and affect macroeconomic variables in the short run.

Many factors aside from monetary and fiscal policy influence aggregate demand. Among them is the desired spending by households and firms. When desired spending changes, aggregate demand shifts, and if policymakers don't respond, such shifts cause short-run fluctuations in output and employment. Policymakers sometimes use the policy levers at their disposal to try to offset these effects and stabilize the economy. The theory behind these policy actions, together with the difficulties that arise when using this theory in practice, is the focus of this chapter.

35-1 How Monetary Policy Influences Aggregate Demand

The aggregate-demand curve shows the total quantity of goods and services demanded in the economy for any price level. The preceding chapter discussed three reasons for the curve's downward slope:

- **The wealth effect:** A lower price level raises the real value of households' money holdings, which are part of their wealth. Higher real wealth stimulates consumer spending and thus increases the quantity of goods and services demanded.
- **The interest-rate effect:** A lower price level reduces the amount of money people want to hold. As people try to lend out their excess money, the interest rate falls. The lower interest rate stimulates investment spending and thus increases the quantity of goods and services demanded.
- **The exchange-rate effect:** When a lower price level reduces the interest rate, investors move some of their funds overseas in search of higher returns. This movement causes the real value of the domestic currency to fall in the market for foreign-currency exchange. Domestic goods become less expensive relative to foreign goods. This change in the real exchange rate stimulates spending on net exports and thus increases the quantity of goods and services demanded.

These three effects work simultaneously to increase the quantity of goods and services demanded when the price level falls. And the opposite occurs when the price level rises.

While all three of these effects help explain the downward slope of the aggregate-demand curve, they are not of equal importance. Because money holdings are a small part of household wealth, the wealth effect is the least important. In addition, because exports and imports represent only a small fraction of U.S. GDP, the exchange-rate effect is not large for the U.S. economy. (It is more important for smaller countries, which typically export and import a larger fraction of their GDP.) **For the U.S. economy, the most important reason for the downward slope of the aggregate-demand curve is the interest-rate effect.**

theory of liquidity preference
Keynes's theory that the interest rate adjusts to bring money supply and money demand into balance

For that reason, we now examine the short-run determination of interest rates in more detail, developing what is known as the **theory of liquidity preference**. This theory of interest rates helps explain the downward slope of the aggregate-demand curve as well as how monetary and fiscal policy can shift the curve. It is useful in analyzing the causes of short-run fluctuations and the possible policy responses.

35-1a The Theory of Liquidity Preference

In his book *The General Theory of Employment, Interest, and Money*, John Maynard Keynes proposed the theory of liquidity preference to explain the factors that determine an economy's interest rate. The theory is, in essence, an application of supply and demand. According to Keynes, the interest rate adjusts to balance the supply of and demand for money.

You may recall that economists distinguish between two interest rates: The **nominal interest rate** is the interest rate as usually reported, while the **real interest rate** is the interest rate corrected for the effects of inflation. When there is no inflation, the two rates are the same. But when borrowers and lenders expect prices to rise over the term of the loan, they agree to a nominal interest rate that exceeds the real interest rate by the expected rate of inflation. The higher nominal interest rate compensates for the expectation that the loan will be repaid in less valuable dollars.

Which interest rate are we trying to explain with the theory of liquidity preference? The answer is both. In the analysis that follows, we hold constant the expected rate of inflation. This assumption is reasonable for studying the economy in the short run because expected inflation is typically stable over short periods. In this case, nominal and real interest rates differ by a constant: When the nominal interest rate rises or falls, the real interest rate that people expect to earn rises or falls by the same amount. For the rest of this chapter, changes in the interest rate refer to both the real interest rate and the nominal interest rate.

Now, let's develop the theory of liquidity preference by considering the supply and demand for money.

Money Supply The first piece of the theory of liquidity preference is the supply of money. As Chapter 30 discussed, the Federal Reserve controls the money supply in the United States. Historically, the Fed has altered the money supply primarily by changing the quantity of reserves in the banking system through the purchase and sale of government bonds in open-market operations. When the Fed buys government bonds, the dollars it pays for them are typically deposited in banks and added to bank reserves. When the Fed sells government bonds, the dollars it receives for them are withdrawn from the banking system, and bank reserves fall. These changes in bank reserves, in turn, lead to changes in banks' ability to make loans and create money. In this way, the Fed can alter the money supply.

In addition to open-market operations, the Fed can influence the money supply using various other tools. One option used more recently is for the Fed to change the interest rate it pays on reserves. For example, a decrease in the interest rate paid on reserves discourages banks from holding them, increasing bank lending and, in turn, the money supply. Conversely, an increase in the interest rate paid on reserves encourages banks to hold them, reducing bank lending and the money supply. The Fed can also alter the money supply by changing reserve requirements (the amount of reserves banks must hold against deposits) or the discount rate (the interest rate the Fed charges banks for borrowing reserves from it).

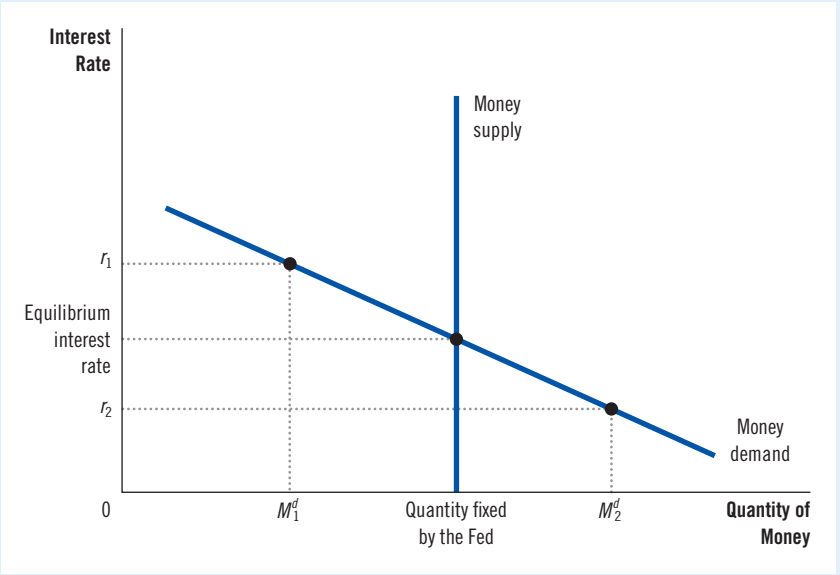
These details about monetary control are important for the implementation of Fed policy, but they are not crucial for the analysis in this chapter. The goal here is to examine how changes in the money supply affect the aggregate demand for goods and services. Toward that end, the theory of liquidity preference makes the simplifying assumption that the Fed controls the money supply directly.

Because the quantity of money supplied in the economy is fixed at whatever level the Fed decides to set it, it does not depend on the prevailing interest rate. We represent a fixed money supply with a vertical supply curve, as in Figure 1.

Figure 1

Equilibrium in the Money Market

According to the theory of liquidity preference, the interest rate adjusts to bring the quantity of money supplied and the quantity of money demanded into balance. If the interest rate is above the equilibrium level (such as at r_1), the quantity of money people want to hold (M_1^d) is less than the quantity the Fed has created, and this surplus puts downward pressure on the interest rate. Conversely, if the interest rate is below the equilibrium level (such as at r_2), the quantity of money people want to hold (M_2^d) exceeds the quantity the Fed has created, and this shortage puts upward pressure on the interest rate. In this manner, the theory says, the forces of supply and demand in the market for money push the interest rate toward the equilibrium interest rate at which people are content holding the quantity of money the Fed has created.



Money Demand The second piece of the theory of liquidity preference is the demand for money. Recall that an asset's **liquidity** refers to the ease with which it can be converted into the economy's medium of exchange. Because money is the economy's medium of exchange, it is the most liquid asset available. This liquidity explains why there is always some demand for money: People choose to hold money instead of other assets that offer higher rates of return because they can use it to buy goods and services.

Although many factors determine the quantity of money demanded, the theory of liquidity preference emphasizes the interest rate because it represents the opportunity cost of holding money. That is, when you hold wealth as cash rather than as an interest-bearing bond or bank deposit, you lose the interest you could have earned. An increase in the interest rate raises the cost of holding money and, as a result, reduces the quantity of money demanded. Conversely, a decrease in the interest rate reduces the cost of holding money and raises the quantity demanded. That's why, as Figure 1 shows, the money demand curve slopes downward.

Equilibrium in the Money Market According to the theory of liquidity preference, the interest rate adjusts to balance the supply and demand for money. There is one interest rate, called the **equilibrium interest rate**, at which the quantity

of money demanded exactly balances the quantity supplied. If the interest rate is at any other level, people will try to adjust their portfolios of money and nonmonetary assets and drive the interest rate toward the equilibrium.

For example, suppose that the interest rate is above the equilibrium level, such as r_1 in Figure 1. In this case, the quantity of money that people want to hold, M_1^d, is less than the quantity that the Fed has supplied. Those people who are holding the surplus of money will try to get rid of it by buying interest-bearing bonds or by depositing it in interest-bearing bank accounts. Because bond issuers and banks prefer to pay lower interest rates, they respond to this surplus of money by lowering the rates they offer. As the interest rate falls, people become more willing to hold money until, at the equilibrium interest rate, people are happy to hold exactly the amount the Fed has supplied.

Conversely, at interest rates below the equilibrium level, such as r_2 in Figure 1, the quantity of money that people want to hold, M_2^d, exceeds the quantity that the Fed has supplied. As a result, people try to increase their holdings of money by reducing their holdings of bonds and other interest-bearing assets. As people cut back on their holdings of bonds, bond issuers find that they have to offer higher interest rates to attract buyers. Ultimately, the interest rate rises until it reaches the equilibrium level.

FYI

Interest Rates in the Long Run and the Short Run

An earlier chapter said that the interest rate adjusts to balance the supply of loanable funds (national saving) and the demand for loanable funds (desired investment). But we just said that the interest rate adjusts to balance the supply of and demand for money. How can these two theories be reconciled?

To answer this question, we need to focus on three macroeconomic variables: the economy's output of goods and services, the interest rate, and the price level. According to the classical macroeconomic theory developed earlier in the book, these variables are determined as follows:

1. **Output** is determined by the supplies of capital and labor and the available production technology for turning capital and labor into output. (We call this the natural level of output.)
2. For any given level of output, the **interest rate** adjusts to balance the supply and demand for loanable funds.
3. Given output and the interest rate, the **price level** adjusts to balance the supply and demand for money. Changes in the money supply lead to proportionate changes in the price level.

These are three of the essential propositions of classical economic theory. Most economists believe that these propositions do a good job of describing how the economy works **in the long run**.

Yet these propositions do not hold in the short run. As we learned in the preceding chapter, many prices are slow to adjust to changes in the money supply. This is reflected in a short-run aggregate-supply curve that is upward-sloping rather than vertical. As a result, **in the short run**, the overall price level cannot, by itself, move to balance the supply of and demand for money. This stickiness of the price level requires the interest rate to move to bring the money market into equilibrium. These changes in the interest rate, in turn, affect the aggregate demand for goods and services. As aggregate demand fluctuates, the economy's output of goods and services moves away from the level determined by factor supplies and technology.

When thinking about the economy in the short run (day to day, week to week, month to month, or quarter to quarter), apply the following logic:

1. The **price level** is stuck at some level (based on previously formed expectations) and, in the short run, is relatively unresponsive to changing economic conditions.
2. For any given (stuck) price level, the **interest rate** adjusts to balance the supply of and demand for money.
3. The interest rate that balances the money market influences the quantity of goods and services demanded and thereby the level of **output**.

Notice that this logic reverses the order of analysis used to study the economy in the long run.

The two different theories of the interest rate are useful for different purposes. When thinking about the long-run determinants of the interest rate, it is best to keep in mind the loanable-funds theory, which highlights the importance of an economy's saving propensities and investment opportunities. But when thinking about the short-run determinants of the interest rate, it is best to keep in mind the liquidity-preference theory, which highlights the importance of monetary policy. ■

35-1b The Downward Slope of the Aggregate-Demand Curve

Having seen how the theory of liquidity preference explains the economy's equilibrium interest rate, we now consider the theory's implications for the aggregate demand for goods and services. As a warm-up exercise, let's first use the theory to reexamine a topic we have already covered—the interest-rate effect and the downward slope of the aggregate-demand curve. In particular, suppose that the overall level of prices in the economy rises. What happens to the interest rate that balances the supply and demand for money, and how does that change affect the quantity of goods and services demanded?

As Chapter 31 discussed, the price level is one determinant of the quantity of money demanded. At higher prices, more money is exchanged every time a good or service is sold, so people will choose to hold a larger quantity of money. That means a higher price level increases the quantity of money demanded for any interest rate. This is shown in panel (a) of Figure 2: An increase in the price level from P_1 to P_2 shifts the money demand curve to the right from MD_1 to MD_2.

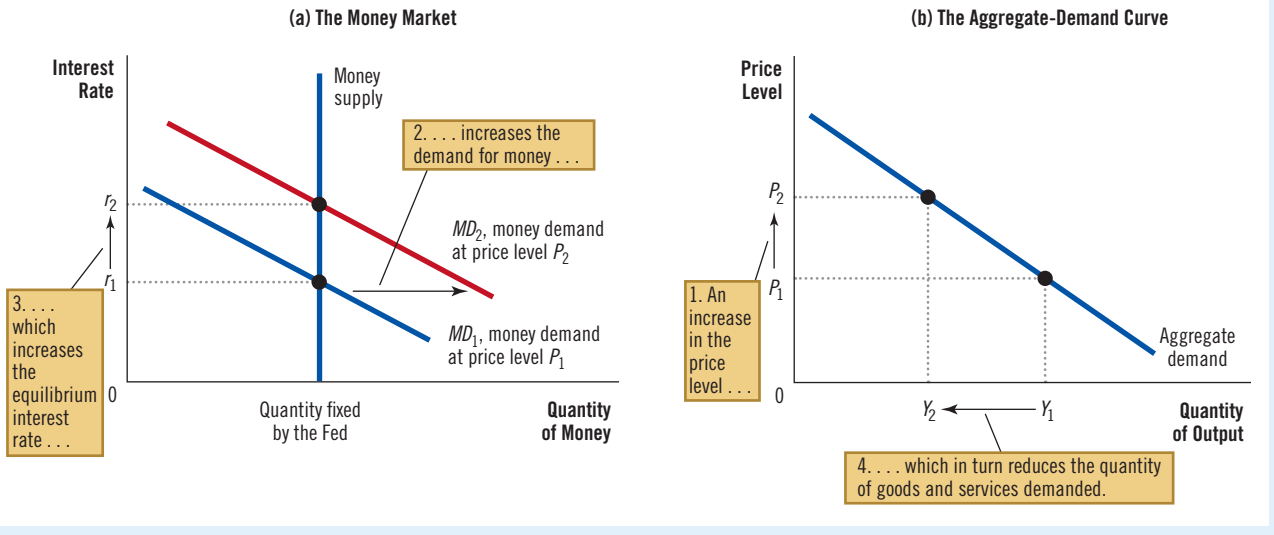
This shift in money demand affects the equilibrium in the money market. As the figure shows, the interest rate must rise to balance money supply and money demand. Because the higher price level has increased the amount of money people want to hold, it has shifted the money demand curve to the right. Yet the quantity of money supplied is fixed, so the interest rate must rise from r_1 to r_2 to forestall the additional demand.

This increase in the interest rate has ramifications not only for the money market but also for the quantity of goods and services demanded, as shown in panel (b).

Figure 2

The Money Market and the Slope of the Aggregate-Demand Curve

An increase in the price level from P_1 to P_2 shifts the money demand curve to the right, as in panel (a). This increase in money demand causes the interest rate to rise from r_1 to r_2. Because the interest rate is the cost of borrowing, the increase in the interest rate reduces the quantity of goods and services demanded from Y_1 to Y_2. This negative relationship between the price level and quantity demanded is represented by a downward-sloping aggregate-demand curve, as in panel (b).



At a higher interest rate, the cost of borrowing and the return to saving are greater. Fewer households choose to borrow to buy a new house, and those who do buy smaller houses, so the demand for residential investment falls. Fewer firms choose to borrow to build new factories and buy new equipment, so business investment falls. Thus, when the price level rises from P_1 to P_2, increasing money demand from MD_1 to MD_2 and raising the interest rate from r_1 to r_2, the quantity of goods and services demanded falls from Y_1 to Y_2.

The interest-rate effect can be summarized in three steps: (1) A higher price level raises money demand. (2) Higher money demand leads to a higher interest rate. (3) A higher interest rate reduces the quantity of goods and services demanded. The same logic works for a decline in the price level: A lower price level reduces money demand, leading to a lower interest rate and a larger quantity of goods and services demanded. The result of this analysis is a negative relationship between the price level and the quantity of goods and services demanded, as illustrated by a downward-sloping aggregate-demand curve.

35-1c Changes in the Money Supply

So far, we have used the theory of liquidity preference to explain more fully how the total quantity of goods and services demanded in the economy changes as the price level changes. That is, we have examined movements along a downward-sloping aggregate-demand curve. The theory also sheds light, however, on some of the other events that alter the quantity of goods and services demanded. Whenever the quantity of goods and services demanded changes **for any price level**, the aggregate-demand curve shifts.

One important variable that shifts the aggregate-demand curve is monetary policy. To see how, suppose that the Fed increases the money supply by buying government bonds in open-market operations. (Why the Fed might do this will become clear later.) Let's consider how this monetary injection influences the equilibrium interest rate for a given price level. This will reveal what the injection does to the position of the aggregate-demand curve.

As panel (a) of Figure 3 shows, an increase in the money supply shifts the money supply curve to the right from MS_1 to MS_2. Because the money demand curve has not changed, the interest rate falls from r_1 to r_2 to balance money supply and money demand. That is, the interest rate must fall to induce people to hold the additional money the Fed has created, restoring equilibrium in the money market.

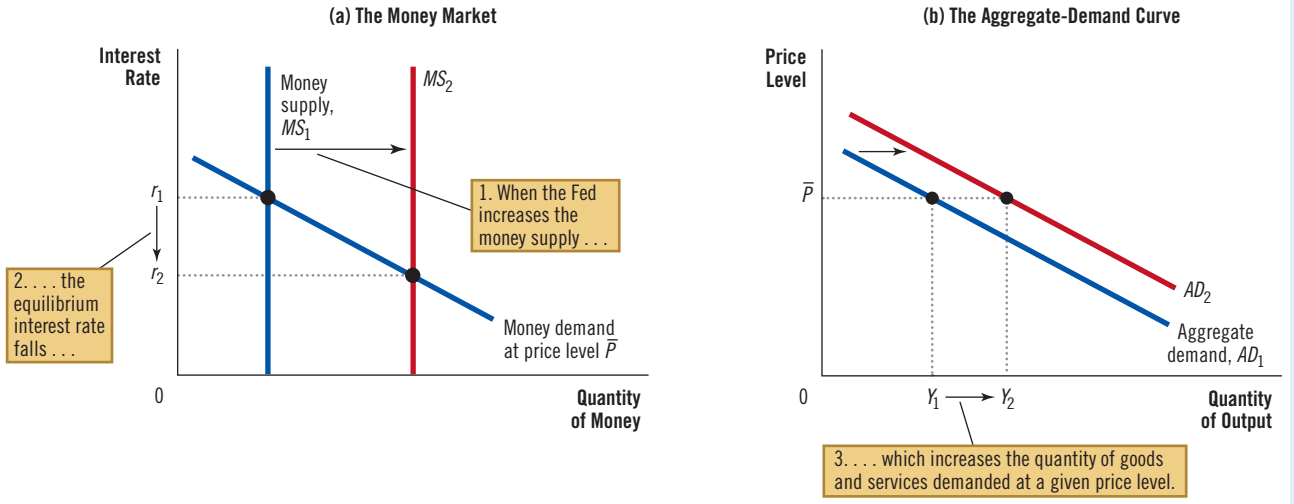
Once again, the interest rate influences the quantity of goods and services demanded, as shown in panel (b) of Figure 3. The lower interest rate reduces the cost of borrowing and the return to saving. Households spend more on new homes, stimulating the demand for residential investment. Firms spend more on new factories and new equipment, stimulating business investment. As a result, the quantity of goods and services demanded at a given price level, $\bar{P}$, rises from Y_1 to Y_2. There is nothing special about $\bar{P}$: The monetary injection raises the quantity of goods and services demanded at every price level. Thus, the entire aggregate-demand curve shifts to the right.

To sum up: **When the Fed increases the money supply, it lowers the interest rate and increases the quantity of goods and services demanded for any price level, shifting the aggregate-demand curve to the right. Conversely, when the Fed contracts the money supply, it raises the interest rate and reduces the quantity of goods and services demanded for any price level, shifting the aggregate-demand curve to the left.**

Figure 3

A Monetary Injection

In panel (a), an increase in the money supply from MS_1 to MS_2 reduces the equilibrium interest rate from r_1 to r_2. When the interest rate falls, the cost of borrowing drops, raising the quantity of goods and services demanded at a given price level from Y_1 to Y_2. In panel (b), therefore, the aggregate-demand curve shifts to the right from AD_1 to AD_2.



35-1d The Role of Interest-Rate Targets in Fed Policy

How does the Federal Reserve affect the economy? So far, this book has treated the money supply as the Fed's policy instrument. For example, when the Fed buys government bonds in open-market operations, it increases the money supply and expands aggregate demand. When the Fed sells government bonds in open-market operations, it decreases the money supply and contracts aggregate demand.

Focusing on the money supply is a good starting point, but another perspective is useful when thinking about recent policy. In the past, the Fed has, at times, set a target for the money supply, but that is no longer the case. The Fed now conducts policy in large part by setting a target for the **federal funds rate**—the interest rate that banks charge one another for short-term loans. This target is reevaluated every six weeks at meetings of the Federal Open Market Committee.

There are several related reasons for the Fed's decision to target the federal funds rate. One is that the money supply is hard to measure precisely. Another is that money demand fluctuates. For any given money supply, these fluctuations can lead to changes in interest rates, aggregate demand, and output. But when the Fed targets the federal funds rate, it essentially accommodates the day-to-day shifts in money demand by adjusting the money supply accordingly.

The Fed's decision to target an interest rate does not fundamentally alter our analysis of monetary policy. The theory of liquidity preference establishes an important principle: **Monetary policy can be described either in terms of the money supply or in terms of the interest rate.** When the Fed sets a target for the federal funds rate of, say, 4 percent, it commits itself to adjusting the money supply to make the equilibrium in the money market hit that target.

As a result, changes in monetary policy can be viewed either as changing the interest rate target or as changing the money supply. When you read in the news that "the Fed has reduced the target for the federal funds rate from 4 to 3 percent," understand that money supply is being adjusted to hit the target. When the Fed reduces the federal funds target, it increases the money supply and reduces the equilibrium interest rate (just as in Figure 3). Conversely, when the Fed raises the target, it decreases the money supply and raises the equilibrium interest rate.

This link between the Fed's interest-rate target and the money supply has become nearly automatic since 2008, when the Fed introduced interest on reserves as a policy tool. The rate at which banks will lend in the federal funds market is closely tied to what they can earn by holding reserves at the Fed. When the Fed reduces its federal funds target, it also reduces the interest rate it pays on reserves, which increases the money supply because banks lend more to the public. And when the Fed raises the federal funds target, it also raises the interest rate on reserves, which reduces the money supply because banks cut back on lending.

The bottom line is simple: **Changes in monetary policy aimed at expanding aggregate demand can be described either as increasing the money supply or as reducing the interest rate. Changes in monetary policy aimed at contracting aggregate demand can be described either as decreasing the money supply or as raising the interest rate.**

Why the Fed Watches the Stock Market (and Vice Versa)

"The stock market has predicted nine out of the past five recessions." So said Paul Samuelson, the distinguished economist (and textbook author). Samuelson was entirely right about the stock market's unimpressive fortune-telling powers. The stock market is highly volatile and often gives the wrong signals about the future of economic activity.

But stock prices sometimes reflect broader economic developments. The economic boom of the 1990s, for example, appeared not only in rapid GDP growth and falling unemployment but also in rising stock prices, which increased about fourfold during this decade. Similarly, the Great Recession of 2008 and 2009 was reflected in falling stock prices: From November 2007 to March 2009, the stock market lost about half its value. And the severe economic downturn during the 2020 pandemic was mirrored by a 34-percent decline in stock prices from February 14 to March 23 of that year.

How should the Fed respond to stock-market fluctuations? It has no reason to care about stock prices in themselves, but it does have the job of monitoring and responding to developments in the overall economy, and the stock market is a piece of that puzzle. When the stock market booms, households become wealthier, and this increased wealth stimulates consumer spending. In addition, a rise in stock prices makes it more attractive for firms to sell new shares, increasing investment spending. For both reasons, a booming stock market expands the aggregate demand for goods and services.

As we discuss more fully later in this chapter, one of the Fed's goals is to stabilize aggregate demand, which, in turn, means greater stability in output and the price level. To promote stability, the Fed might respond to a stock-market boom by keeping the money supply lower and interest rates higher than it otherwise would. The contractionary effects of tighter monetary policy would offset the expansionary effects of higher stock prices. In fact, the Fed kept real interest rates high by historical standards during the stock-market boom of the late 1990s.

The opposite occurs when the stock market falls. At such moments, spending on consumption and investment tends to decline, depressing aggregate demand and pushing the economy toward recession. To stabilize aggregate demand, the Fed might increase the money supply and lower interest rates. On October 19, 1987, for example, the stock market fell by 22.6 percent—one of the biggest one-day drops in history. The Fed responded to the market crash by reducing the federal funds rate from 7.7 percent at the beginning of October to 6.6 percent at the end of the month. In part because of the Fed's quick action, the economy avoided a recession. Similarly, as the preceding chapter discussed, the Fed also reduced interest rates during the economic downturn and stock-market decline of 2008 and 2009, but this time, monetary policy was not sufficient to avert a deep recession. And it cut interest rates substantially during the pandemic-induced stock decline of 2020.

While the Fed watches the stock market, participants in the stock market also watch the Fed. Changes in monetary policy can influence stock valuations. For example, when the Fed raises interest rates, stock prices typically fall. Because interest-bearing bonds are an alternative to stocks, a higher interest rate reduces the present value of any future cash flows associated with stock ownership. In addition, tighter monetary policy depresses the demand for goods and services, reducing economic activity and firms' profits. ●

35-1e The Zero Lower Bound

So monetary policy works through interest rates. This raises a question: What if the Fed's target interest rate has fallen as far as it can? In the Great Recession of 2008 and 2009 and again during the coronavirus recession of 2020, the federal funds rate fell to about zero. In this situation, what, if anything, can monetary policy do to stimulate the economy?

Some economists describe this situation as a **liquidity trap**. According to the *theory of liquidity preference*, expansionary monetary policy works by reducing interest rates and stimulating investment spending. But if interest rates have already fallen to around zero, monetary policy may no longer be effective. Nominal interest rates cannot fall much below zero: Rather than making a loan at a negative nominal interest rate, a person would just hold cash. In this environment, expansionary monetary policy raises the supply of money, making the public's asset portfolio more liquid, but because interest rates can't fall further, the extra liquidity might not have any effect. Aggregate demand, production, and employment may be "trapped" at low levels.

Other economists are skeptical about the relevance of liquidity traps and believe that a central bank continues to have tools to expand the economy, even after its interest rate target hits its lower bound of zero. One option is to have the central bank commit itself to keeping interest rates low for an extended period. Such a policy is sometimes called **forward guidance**. Even if the central bank's current target for the interest rate cannot fall any further, the promise that interest rates will remain low may help stimulate investment spending.

A second option is to have the central bank conduct expansionary open-market operations using a larger variety of financial instruments. Normally, the Fed conducts expansionary open-market operations by buying short-term government bonds. But it could also buy mortgage-backed securities and longer-term government bonds to lower the interest rates on these kinds of loans. This type of unconventional monetary policy is sometimes called **quantitative easing** because it increases the quantity of bank reserves. During both the Great Recession and the coronavirus recession, the Fed engaged in forward guidance and quantitative easing.

Some economists have suggested that the possibility of hitting the zero lower bound for interest rates justifies setting the target rate of inflation well above zero. Under zero inflation, the real interest rate, like the nominal interest rate, can never fall below zero. But if the normal rate of inflation is, say, 4 percent, then the central bank can easily push the real interest rate to negative 4 percent by lowering the nominal interest rate to zero. A higher inflation target gives monetary policymakers more room to stimulate the economy when needed, reducing the risk of hitting the zero lower bound and falling into a liquidity trap.

QuickQuiz

1. According to the theory of liquidity preference, an economy's interest rate adjusts
   a. to balance the supply and demand for loanable funds.
   b. to balance the supply and demand for money.
   c. one-for-one to changes in expected inflation.
   d. to equal the interest rate prevailing in world financial markets.
2. If the central bank wants to contract aggregate demand, it can _____ the money supply and thereby _____ the interest rate.
   a. increase; increase
   b. increase; decrease
   c. decrease; increase
   d. decrease; decrease
3. The Fed's target for the federal funds rate
   a. is an extra policy tool for the central bank in addition to and independent of the money supply.
   b. commits the Fed to set a particular money supply so that it hits the announced target.
   c. is a goal that is rarely achieved because the Fed can determine only the money supply.
   d. matters to banks that borrow and lend federal funds but does not influence aggregate demand.

Answers are at the end of the chapter.

35-2 How Fiscal Policy Influences Aggregate Demand

The government can influence the behavior of the economy not only with monetary policy but also with fiscal policy. **Fiscal policy** refers to the government's choices regarding the overall levels of government purchases and taxes. Earlier in the book, we examined how fiscal policy influences saving, investment, and growth in the long run. In the short run, however, the primary effect of fiscal policy is on the aggregate demand for goods and services.

fiscal policy
the setting of the levels of government spending and taxation by government policymakers

35-2a Changes in Government Purchases

When policymakers change the money supply or the level of taxes, they shift the aggregate-demand curve indirectly by influencing the spending decisions of firms or households. By contrast, when the government alters its own purchases of goods and services, it shifts the aggregate-demand curve directly.

Suppose, for instance, that the U.S. Department of Defense places a $20 billion order for new fighter planes with Boeing, the aircraft manufacturer. This order raises the demand for the output produced by Boeing, inducing the company to hire more workers and increase production. For the overall economy, the increase in the demand for Boeing planes means an increase in the total quantity of goods

and services demanded at each price level. As a result, the aggregate-demand curve shifts to the right.

By how much does this $20 billion order from the government shift the aggregate-demand curve? You might guess that the aggregate-demand curve shifts to the right by exactly $20 billion. It turns out, however, that this is not the case. Two macroeconomic effects cause the size of the shift in aggregate demand to differ from the change in government purchases. The first—the multiplier effect—suggests the shift in aggregate demand could be **larger** than $20 billion. The second—the crowding-out effect—suggests the shift in aggregate demand could be **smaller** than $20 billion. We discuss these two effects in turn.

35-2b The Multiplier Effect

When the government buys $20 billion of goods from Boeing, that purchase has repercussions. The immediate impact is to raise employment and profits at Boeing. Then, as the workers see higher earnings and the firm's owners see higher profits, they respond to this increase in income by raising their own spending on consumer goods. As a result, the government purchase from Boeing raises the demand for the products of many other firms in the economy. Because each dollar spent by the government can raise the aggregate demand for goods and services by more than a dollar, government purchases are said to have a **multiplier effect** on aggregate demand.

multiplier effect
the additional shifts in aggregate demand that result when expansionary fiscal policy increases income and thereby increases consumer spending

This multiplier effect continues even after this first round. When consumer spending rises, the firms that produce these consumer goods hire more people and experience higher profits. Higher earnings and profits stimulate consumer spending once again and so on. A positive feedback loop develops, as higher demand leads to higher income, which in turn leads to even higher demand. Once all these effects are added together, the total impact on the quantity of goods and services demanded can be much larger than the initial boost from higher government spending.

Figure 4 illustrates the multiplier effect. The increase in government purchases of $20 billion initially shifts the aggregate-demand curve to the right from AD_1 to AD_2 by exactly $20 billion. But once consumers respond by increasing their spending, the aggregate-demand curve shifts still farther to AD_3.

The multiplier effect arising from the response of consumer spending can be strengthened by the response of investment to higher levels of demand. For instance, Boeing might react to the higher demand for planes by deciding to buy more equipment or build another plant. In this case, higher government demand spurs higher demand for investment goods. The positive influence of demand on investment is sometimes called the **investment accelerator**.

35-2c A Formula for the Spending Multiplier

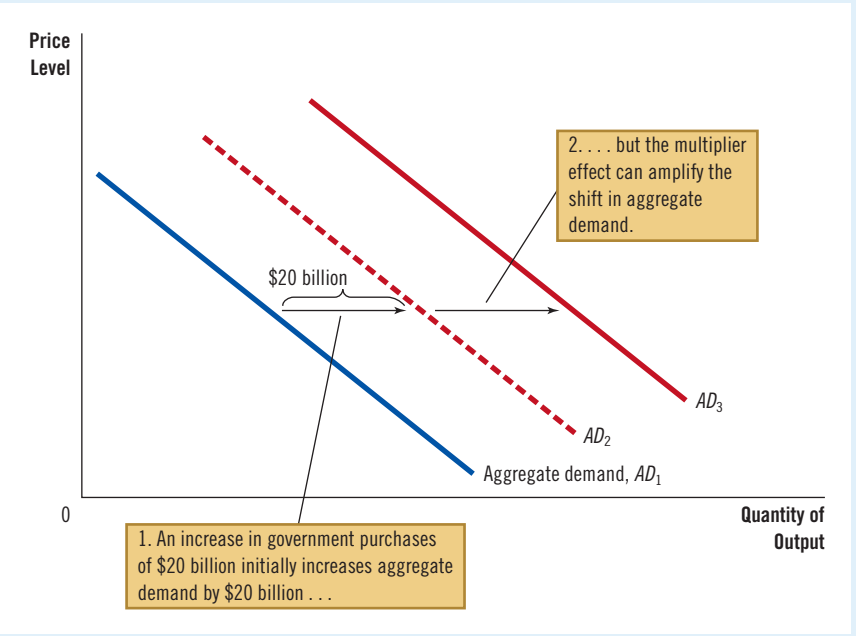
Simple algebra allows us to derive a formula for the size of the multiplier effect that arises when an increase in government purchases induces increases in consumer spending. An important concept, and one that appears in this formula, is the **marginal propensity to consume** (*MPC*)—the fraction of extra income that a household consumes rather than saves. For example, suppose that the marginal propensity to consume is ¾. This means that for every extra dollar a household earns, the household spends $0.75 (¾ of the dollar) and saves $0.25. With an *MPC* of ¾, when the workers and owners of Boeing earn $20 billion from the government contract, they increase their consumer spending by ¾ × $20 billion, or $15 billion.

To gauge the impact of a change in government purchases on aggregate demand, we follow the effects step-by-step. The process begins when the government

Figure 4

The Multiplier Effect

An increase in government purchases of $20 billion can shift the aggregate-demand curve to the right by more than $20 billion. This multiplier effect arises because increases in aggregate income stimulate additional spending by consumers.



spends $20 billion and, as a result, increases national income (earnings and profits) by the same amount. With an extra $20 billion of income, consumers increase spending by $MPC \times \$20$ billion. This additional consumer spending raises the income for the workers and owners of the firms that produce the consumption goods by the same amount. With this second increase in income, consumers increase spending again, this time by $MPC \times (MPC \times \$20 \text{ billion})$. These feedback effects go on and on.

To determine the total impact on the demand for goods and services, we add up all these effects:

$$
\begin{aligned}
&\text{Change in government purchases} = \$20 \text{ billion} \\
&\text{First change in consumption} = MPC \times \$20 \text{ billion} \\
&\text{Second change in consumption} = MPC^2 \times \$20 \text{ billion} \\
&\text{Third change in consumption} = MPC^3 \times \$20 \text{ billion} \\
&\quad \bullet \qquad\qquad \bullet \\
&\quad \bullet \qquad\qquad \bullet \\
&\quad \bullet \qquad\qquad \bullet \\
&\text{Total change in demand} \\
&\quad = (1 + MPC + MPC^2 + MPC^3 + \cdots) \times \$20 \text{ billion}.
\end{aligned}
$$

Here "..." represents an infinite number of similar terms. We can write the multiplier as follows:

$$\text{Multiplier} = 1 + MPC + MPC^2 + MPC^3 + \cdots$$

This multiplier tells us the demand for goods and services that each dollar of government purchases generates.

To simplify this equation for the multiplier, recall from math class that this expression is an infinite geometric series. For x between -1 and $+1$,

$$1 + x + x^2 + x^3 + \cdots = 1/(1 - x).$$

In our case, $x = MPC$. Thus,

$$\text{Multiplier} = 1/(1 - MPC).$$

For example, if *MPC* is $\frac{3}{4}$, the multiplier is $1/(1 - \frac{3}{4})$, which is 4. In this case, the $20 billion of government spending generates $80 billion of demand for goods and services.

This formula shows that the size of the multiplier depends on the marginal propensity to consume. While an *MPC* of $\frac{3}{4}$ leads to a multiplier of 4, an *MPC* of $\frac{1}{2}$ leads to a multiplier of only 2. Thus, a higher *MPC* means a larger multiplier. To see why, remember that the multiplier arises because higher income induces greater consumer spending. The higher the *MPC*, the more consumption responds to a change in income, and the larger the multiplier.

35-2d Other Applications of the Multiplier Effect

Because of the multiplier effect, a dollar of government purchases can generate more than a dollar of aggregate demand. The logic of the multiplier effect, however, is not restricted to changes in government purchases. Instead, it applies to any event that alters spending on any component of GDP—consumption, investment, government purchases, or net exports.

For example, suppose that a recession overseas reduces the demand for U.S. net exports by $10 billion. This reduced spending on U.S. goods and services depresses U.S. national income and, in turn, reduces spending by U.S. consumers. If the marginal propensity to consume is $\frac{3}{4}$ and the multiplier is 4, then the $10 billion fall in net exports leads to a $40 billion contraction in aggregate demand.

As another example, suppose that a stock-market boom increases household wealth and stimulates spending on goods and services by $20 billion. This extra consumer spending increases national income and, in turn, generates even more consumer spending. If the marginal propensity to consume is $\frac{3}{4}$ and the multiplier is 4, then the initial increase of $20 billion in consumer spending translates into an $80 billion increase in aggregate demand.

The multiplier is an important concept in macroeconomics because it shows how the economy can amplify the impact of changes in spending. A small initial change in consumption, investment, government purchases, or net exports can end up having a large effect on aggregate demand and, therefore, the economy's production of goods and services.

35-2e The Crowding-Out Effect

The multiplier effect seems to suggest that when the government buys $20 billion of planes from Boeing, the resulting expansion in aggregate demand is necessarily larger than $20 billion. Yet another force works in the opposite direction. While an increase in government purchases stimulates the aggregate demand for goods and services, it also causes the interest rate to rise, reducing investment spending and putting downward pressure on aggregate demand. The reduction in aggregate

demand that results when a fiscal expansion raises the interest rate is called the **crowding-out effect**.

crowding-out effect the offset in aggregate demand that results when expansionary fiscal policy raises the interest rate and thereby reduces investment spending

To see why crowding out occurs, consider what happens in the money market when the government buys planes from Boeing. The increase in demand raises the incomes of the workers and owners of this firm (and, because of the multiplier effect, of other firms as well). As incomes rise, households plan to buy more goods and services and so choose to hold more of their wealth in liquid form. That is, the increase in income caused by the fiscal expansion raises the demand for money, and this may have consequences.

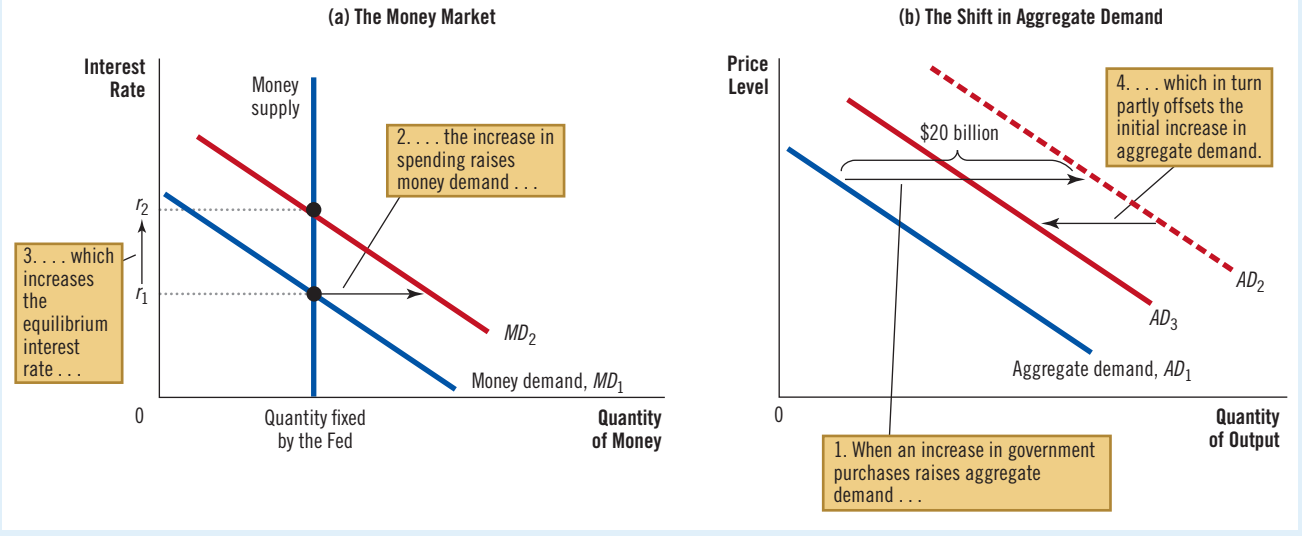
The effect of the increase in money demand is shown in panel (a) of Figure 5. If the Fed does not change the money supply, the vertical supply curve remains the same. When the higher level of income shifts the money demand curve to the right from MD_1 to MD_2, the interest rate must rise from r_1 to r_2 to keep supply and demand in balance.

The increase in the interest rate, in turn, reduces the quantity of goods and services demanded. In particular, because borrowing is more expensive, the demand for residential and business investment goods declines. In other words, as the increase in government purchases increases the demand for goods and services, it may also crowd out investment. This crowding-out effect partially offsets the impact of government purchases on aggregate demand, as illustrated in panel (b) of Figure 5.

Figure 5

The Crowding-Out Effect

Panel (a) shows the money market. When the government increases its purchases of goods and services, income increases, raising the demand for money from MD_1 to MD_2 and increasing the equilibrium interest rate from r_1 to r_2. Panel (b) shows the effects on aggregate demand. The initial impact of the increase in government purchases shifts the aggregate-demand curve from AD_1 to AD_2. Yet because the interest rate is the cost of borrowing, the increase in the interest rate tends to reduce the quantity of goods and services demanded, particularly for investment goods. This crowding out of investment partially offsets the impact of the fiscal expansion on aggregate demand. In the end, the aggregate-demand curve shifts only to AD_3.



The increase in government purchases initially shifts the aggregate-demand curve from AD_1 to AD_2, but once crowding out takes place, the aggregate-demand curve drops back to AD_3.

To sum up: **When the government increases its purchases by $20 billion, the aggregate demand for goods and services could rise by more or less than $20 billion depending on the sizes of the multiplier and crowding-out effects.** The multiplier effect makes the shift in aggregate demand greater than $20 billion. The crowding-out effect pushes the aggregate-demand curve in the opposite direction and can, if large enough, result in an aggregate-demand shift of less than $20 billion.

35-2f Changes in Taxes

The other important instrument of fiscal policy, besides government purchases, is taxation. When the government cuts personal income taxes, for instance, it increases households' take-home pay. Households will save some of this additional income, but they will also spend some on consumer goods. Because the tax cut increases consumer spending, it shifts the aggregate-demand curve to the right. Similarly, a tax increase depresses consumer spending and shifts the aggregate-demand curve to the left.

The size of the shift in aggregate demand resulting from a tax change is also affected by the multiplier and crowding-out effects. When the government cuts taxes and stimulates consumer spending, earnings and profits rise, stimulating additional spending. This is the multiplier effect. At the same time, the increase in income raises money demand, which, other things being equal, tends to increase interest rates. Higher interest rates mean a higher cost of borrowing and less investment spending. This is the crowding-out effect. Depending on the sizes of the multiplier and crowding-out effects, the shift in aggregate demand could be larger or smaller than the tax change that causes it.

In addition to the multiplier and crowding-out effects, there is another important determinant of the size of the shift in aggregate demand that results from a tax change: people's perceptions about whether the tax change is permanent or temporary. For example, suppose that the government announces a tax cut of $1,000 per person. In deciding how much of this $1,000 to spend, people must ask themselves how long this extra income will last. If they expect the tax cut to be permanent, they will view it as adding substantially to their financial resources and may, therefore, increase their spending by a large amount. In this case, the tax cut is likely to have a large impact on aggregate demand. But if people expect the tax change to be temporary, they will view it as adding only slightly to their financial resources and may increase their spending by only a small amount. In this case, the tax cut is likely to have a small impact on aggregate demand.

An extreme example of a temporary tax cut was the one announced in 1992. In that year, President George H. W. Bush faced a lingering recession and an upcoming reelection campaign. He responded to these circumstances by reducing the amount of income tax that the federal government would withhold from workers' paychecks. Because legislated income tax rates did not change, however, every dollar of reduced withholding in 1992 meant an extra dollar of taxes due on April 15, 1993, when income tax returns for 1992 were to be filed. This "tax cut" was really only a short-term loan from the government. Its impact on consumer spending and aggregate demand was relatively small.

FYI

How Fiscal Policy Might Affect Aggregate Supply

So far, this discussion has stressed how changes in government purchases and in taxes influence the quantity of goods and services demanded. Most economists believe that the short-run macroeconomic effects of fiscal policy work primarily through aggregate demand. Yet fiscal policy can potentially influence the quantity of goods and services supplied as well.

For instance, consider the effects of tax changes on aggregate supply. One of the **Ten Principles of Economics** in Chapter 1 is that people respond to incentives. When government policymakers cut tax rates, workers get to keep more of each dollar they earn, so they have a greater incentive to work and produce goods and services. If they respond to this incentive, the quantity of goods and services supplied will be greater at each price level, and the aggregate-supply curve will shift to the right.

Economists who stress the importance of tax policy for aggregate supply rather than aggregate demand are sometimes called **supply siders**. At times, some supply siders have argued that the influence of taxes on aggregate supply is so large that a cut in tax rates will stimulate enough additional production and income to increase tax revenue. This outcome is a theoretical possibility, but most economists do not consider it the normal case. While the supply-side effects of taxes are important to consider, they are rarely large enough to cause tax revenue to rise when tax rates fall.

Like changes in taxes, changes in government purchases can also affect aggregate supply. Suppose, for instance, that the government increases expenditure on a form of government-provided capital, such as roads or the nation's railroad network. Roads and railroads are used by private businesses to make deliveries to their customers, so an increase in the quantity or quality of this infrastructure increases these businesses' productivity. When the government spends more on roads and rails, it increases the quantity of goods and services supplied at any price level, shifting the aggregate-supply curve to the right. This effect on aggregate supply is probably more important in the long run than in the short run, however, because it takes time for the government to build infrastructure and put it into use. ■

QuickQuiz

4. If the government wants to expand aggregate demand, it can _____ government purchases or _____ taxes.
   a. increase; increase
   b. increase; decrease
   c. decrease; increase
   d. decrease, decrease

5. With the economy in a recession because of inadequate aggregate demand, the government increases its purchases by $1,200. Suppose the central bank adjusts the money supply to hold the interest rate constant, investment spending remains unchanged, and the marginal propensity to consume is $^2/_3$. How large is the increase in aggregate demand?
   a. $400
   b. $800
   c. $1,800
   d. $3,600

6. If the central bank in the preceding question had instead held the money supply constant and allowed the interest rate to adjust, the change in aggregate demand resulting from the increase in government purchases would have been
   a. larger.
   b. the same.
   c. smaller but still positive.
   d. negative.

Answers are at the end of the chapter.

35-3 Using Policy to Stabilize the Economy

The influence of monetary and fiscal policy on the aggregate demand for goods and services raises some important questions: Should policymakers use these instruments to control aggregate demand and stabilize the economy? If so, when? If not, why not?

35-3a The Case for Active Stabilization Policy

Let's return to the question that began this chapter: When the president and Congress raise taxes to reduce a budget deficit, how should the Federal Reserve respond? We have seen that the level of taxation is one determinant of the position of the aggregate-demand curve. When the government raises taxes, aggregate demand falls, depressing production and employment in the short run. If the Fed wants to prevent this adverse effect of the fiscal policy, it can expand aggregate demand by increasing the money supply. A monetary expansion would reduce interest rates, stimulate investment spending, and expand aggregate demand. If monetary policy is set appropriately, the combined changes in monetary and fiscal policy could leave the aggregate demand for goods and services unaffected.

This analysis is exactly the sort followed by members of the Federal Open Market Committee (FOMC). They know that monetary policy is an important determinant of aggregate demand. They also know that there are other important determinants as well, including fiscal policy set by the president and Congress. As a result, the FOMC watches the debates over fiscal policy with a keen eye.

This response of monetary policy to the change in fiscal policy is an example of a more general phenomenon: the use of policy instruments to stabilize aggregate demand and, in turn, production and employment. Economic stabilization has been an explicit goal of U.S. policy since the Employment Act of 1946. This act states that "it is the continuing policy and responsibility of the federal government to . . . promote full employment and production." In essence, this law and others hold the government accountable for short-run macroeconomic performance.

The Employment Act has two implications. First, the government should avoid being a cause of economic fluctuations. Unless a crisis requires it, most economists advise against large and sudden changes in monetary and fiscal policy because such changes are likely to cause sizable shifts in aggregate demand. Moreover, when large changes do occur, it is important that monetary and fiscal policymakers respond to each others' actions.

Second, the government should, as much as possible, respond to changes in the private economy to stabilize aggregate demand. The Employment Act was passed not long after the publication of Keynes's *The General Theory of Employment, Interest, and Money*, one of the most influential books ever written about economics. In it, Keynes emphasized the key role of aggregate demand in explaining short-run fluctuations. Keynes said the government should actively stimulate the economy when aggregate demand was insufficient to maintain production at its full-employment level.

Keynes (and his many followers) said that aggregate demand fluctuates because of largely irrational waves of pessimism and optimism. He used the term "animal spirits" to refer to these changes in attitude. When pessimism reigns, households reduce consumption spending, and firms reduce investment spending. The result is reduced aggregate demand, lower production, and higher unemployment. Conversely, when optimism reigns, households and firms increase spending. The result is higher aggregate demand, higher production, and inflationary pressure. Notice that these changes in attitude are, to some extent, self-fulfilling.

In principle, the government can adjust its monetary and fiscal policy in response to these waves of optimism and pessimism and, thereby, stabilize the economy. For example, when people are excessively pessimistic, the Fed can expand the money supply to lower interest rates and expand aggregate demand. When they are excessively optimistic, it can contract the money supply to raise interest rates and dampen aggregate demand. Former Fed Chairman William McChesney Martin

described this view of monetary policy very simply: "The Federal Reserve's job is to take away the punch bowl just as the party gets going."

Case Study

Keynesians in the White House

When a reporter in 1961 asked President John F. Kennedy why he advocated a tax cut, Kennedy replied, "To stimulate the economy. Don't you remember your Economics 101?" Kennedy's policy was, in fact, based on the theory of fiscal policy analyzed in this chapter. His goal was to raise consumer spending, expand aggregate demand, and increase the economy's production and employment.

In proposing a tax cut, Kennedy was relying on a group of eminent economic advisers. In addition to Paul Samuelson, who tutored Kennedy informally, they included James Tobin, Robert Solow, and Kenneth Arrow, all of whom would later win Nobel Prizes for their contributions to the field. As students in the 1940s, these economists had closely studied Keynes's *General Theory*, which was then only a few years old. Kennedy's proposal put Keynes's ideas into action.

While tax changes have a potent influence on aggregate demand, they can also alter the aggregate supply of goods and services, as discussed earlier in an FYI box. Part of the Kennedy proposal was an investment tax credit that gave a tax break to firms that invested in new capital. Higher investment would not only stimulate aggregate demand immediately but also increase the economy's productive capacity over the long run. When the tax cut Kennedy proposed was finally enacted in 1964, it helped usher in a period of robust growth.

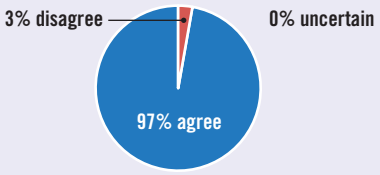
Since the 1964 tax cut, policymakers have, from time to time, used fiscal policy as a tool for influencing aggregate demand. For example, when President Barack Obama moved into the Oval Office in 2009, he faced an economy in the midst of a recession. One of his first policy initiatives was a stimulus bill called the American Recovery and Reinvestment Act (ARRA), which included substantial increases in government spending. A central aim of this legislation was to increase aggregate demand and speed up the recovery. And again, in 2020 and 2021, the federal government under both President Trump and President Biden engaged in fiscal stimulus in part to restore aggregate demand and mitigate the lingering effects of the coronavirus recession. ●

Ask the Experts

Economic Stimulus

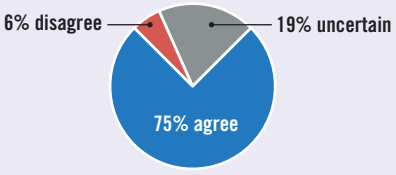
"Because of the American Recovery and Reinvestment Act of 2009, the U.S. unemployment rate was lower at the end of 2010 than it would have been without the stimulus bill."

What do economists say?



"Taking into account all of the ARRA's economic consequences—including the economic costs of raising taxes to pay for the spending, its effects on future spending, and any other likely future effects—the benefits of the stimulus will end up exceeding its costs."

What do economists say?



Source: IGM Economic Experts Panel, July 29, 2014.

35-3b The Case against Active Stabilization Policy

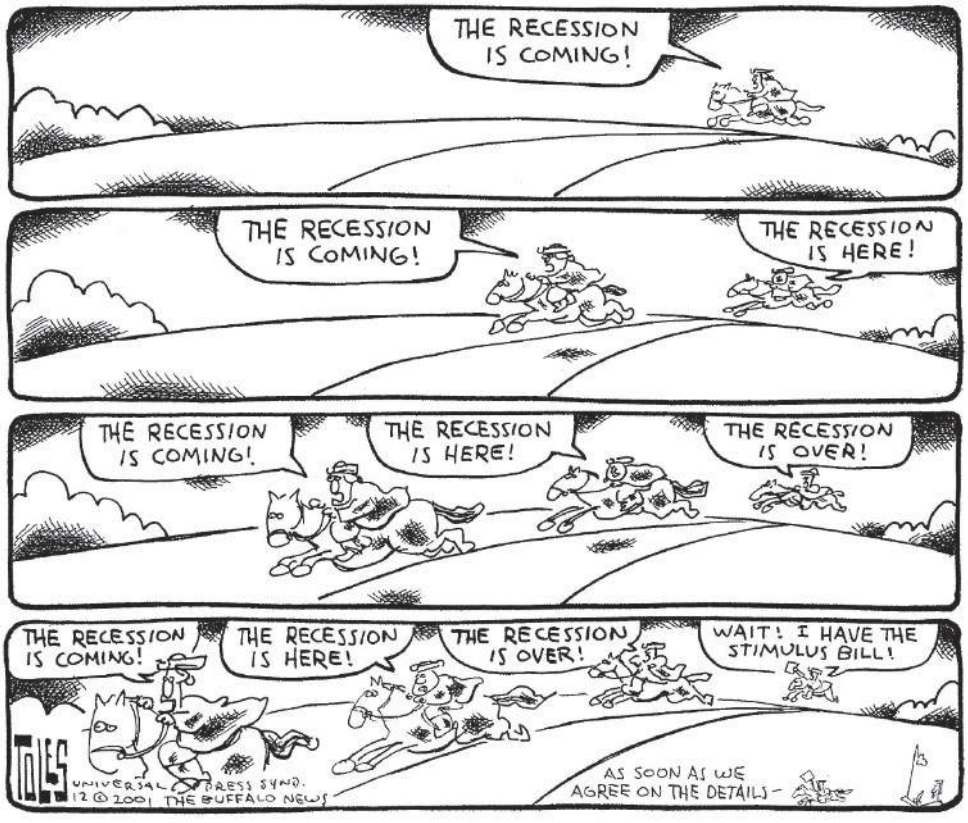
Some economists argue that the government should avoid active use of monetary and fiscal policy to try to stabilize the economy. They say that these policy instruments should be set to achieve long-run goals, such as robust economic growth and low inflation, and that the economy should deal with most short-run fluctuations on its own. These economists may admit that monetary and fiscal policy can stabilize the economy in theory, but they doubt whether it can do so in practice.

The main argument against active monetary and fiscal policy is that these policies operate with a long lag. As we have seen,

monetary policy works by changing interest rates, which, in turn, influence investment spending. But many firms make investment plans far in advance. Most economists believe that it takes at least six months for changes in monetary policy to have much effect on output and employment. Moreover, once these effects occur, they can last for several years. Critics of stabilization policy say that because of this lag, the Fed should not try to fine-tune the economy. They claim that the Fed often reacts to changing circumstances too late and, as a result, ends up causing rather than curing economic fluctuations. These critics advocate a passive monetary policy, such as slow and steady growth in the money supply.

Fiscal policy also works with a lag, one that is largely caused by the political process. In the United States, most big changes in government spending and taxes must go through congressional committees in both the House and the Senate, be passed by both legislative bodies, and then be signed by the president. Completing this process can take months or, in some cases, years. By the time the change in fiscal policy is ready to be implemented, the condition of the economy may have changed.

These lags in monetary and fiscal policy are problematic because economic forecasting is so imprecise. If it were possible to accurately predict the condition of the economy a year in advance, monetary and fiscal policymakers could look ahead when making policy decisions. With enough lead time, policymakers could choose the appropriate short-run policy despite the lags they face. In practice, however, recessions and depressions arrive without much advance warning. The best that policymakers can do is to respond to economic changes as they occur.



35-3c Automatic Stabilizers

All economists—both advocates and critics of stabilization policy—agree that the lags in implementation reduce the efficacy of policy as a tool for short-run stabilization. The economy would be more stable, therefore, if policymakers could find a way to avoid some of these lags. In fact, they have. **Automatic stabilizers** are changes in fiscal policy that stimulate aggregate demand when the economy goes into a recession but that occur without policymakers having to take any deliberate action.

automatic stabilizers changes in fiscal policy that stimulate aggregate demand when the economy goes into a recession but that occur without policymakers having to take any deliberate action

The most important automatic stabilizer is the tax system. When the economy enters a recession, the government's tax collections decline automatically because almost all taxes are closely tied to economic activity. The personal income tax depends on households' incomes, the payroll tax depends on workers' earnings, and the corporate income tax depends on firms' profits. Because incomes, earnings, and profits all fall in a recession, the government's tax revenue falls as well. This automatic tax cut stimulates aggregate demand and reduces the magnitude of economic fluctuations.

Some government spending also acts as an automatic stabilizer. When the economy goes into a recession and workers are laid off, more people become eligible for unemployment insurance benefits, welfare benefits, and other forms of income support. This automatic increase in government spending stimulates aggregate demand when demand is insufficient to maintain full employment. When the unemployment insurance system was first enacted in the 1930s, the economists who advocated this policy recognized its power as an automatic stabilizer.

The automatic stabilizers in the U.S. economy are not sufficiently strong to prevent recessions completely. Nonetheless, without them, output and employment would likely be more volatile than they are. For this reason, while some politicians have proposed a constitutional amendment requiring the federal government to always run a balanced budget, economists generally oppose it. When the economy experiences a recession, taxes fall, government spending rises, and the government's budget moves toward deficit. If the government faced a strict balanced-budget rule, it would be forced to look for ways to raise taxes or cut spending at the worst possible time. In other words, a strict balanced-budget rule would undercut the automatic stabilizers inherent in the current system of taxes and government spending.

QuickQuiz

7. Suppose a wave of negative "animal spirits" overruns the economy, and people become pessimistic about the future. To stabilize aggregate demand, the Fed could _____ its target for the federal funds rate or Congress could _____ taxes.
   a. increase; increase
   b. increase; decrease
   c. decrease; increase
   d. decrease; decrease
8. Monetary policy affects the economy with a lag mainly because it takes a long time
   a. for central banks to make policy changes.
   b. to change the money supply after a policy decision has been made.
   c. for a change in the money supply to affect interest rates.
   d. for a change in interest rates to affect investment spending.
9. Which of the following is an example of an automatic stabilizer? When the economy goes into a recession,
   a. more people become eligible for unemployment insurance benefits.
   b. stock prices decline, particularly for firms in cyclical industries.
   c. Congress begins hearings about a possible stimulus package.
   d. the Fed changes its target for the federal funds rate.

Answers are at the end of the chapter.

35-4 Conclusion

Before policymakers make any change in policy, they need to consider all the effects of their decisions. Earlier, this book examined classical models of the economy, which describe the long-run effects of monetary and fiscal policy. These models showed how fiscal policy influences saving, investment, and long-run growth and how monetary policy affects the price level and the inflation rate.

This chapter examined the short-run effects of monetary and fiscal policy. We saw how these policy instruments influence the aggregate demand for goods and services. Changes in aggregate demand can, in the short run, alter the economy's production and employment.

Both time horizons are important. When Congress alters government spending or taxes, it needs to consider both the long-run effects on growth and the short-run effects on employment. When the Fed changes the money supply and interest rates, it must recognize the long-run effect on inflation as well as the short-run effect on production. In all parts of government, policymakers must keep in mind both long-run and short-run goals.

Chapter in a Nutshell

- In his work on short-run economic fluctuations, Keynes proposed the theory of liquidity preference to explain the determinants of the interest rate. According to this theory, the interest rate adjusts to balance the supply and demand for money.
- An increase in the price level raises money demand and increases the interest rate that brings the money market into equilibrium. Because the interest rate represents the cost of borrowing, a higher interest rate reduces investment spending and thereby reduces the quantity of goods and services demanded. The downward-sloping aggregate-demand curve expresses this negative relationship between the price level and the quantity demanded.
- Policymakers can influence aggregate demand using monetary policy. An increase in the money supply reduces the equilibrium interest rate for any price level. Because a lower interest rate stimulates investment spending, the aggregate-demand curve shifts to the right. Conversely, a decrease in the money supply raises the equilibrium interest rate for any price level and shifts the aggregate-demand curve to the left.
- Policymakers can also influence aggregate demand using fiscal policy. An increase in government purchases or a cut in taxes shifts the aggregate-demand curve to the right. A decrease in government purchases or an increase in taxes shifts the aggregate-demand curve to the left.
- When the government alters spending or taxes, the resulting shift in aggregate demand can be larger or smaller than the fiscal change. The multiplier effect tends to amplify the effects of fiscal policy on aggregate demand. The crowding-out effect tends to dampen those effects.
- Because monetary and fiscal policy can influence aggregate demand, the government sometimes uses these policy instruments to stabilize the economy. Economists disagree about how actively the government should do this. Advocates of active stabilization policy say that changes in attitudes by households and firms shift aggregate demand and that, if the government does not respond, the result is undesirable and unnecessary fluctuations in output and employment. Advocates of more passive policy say that monetary and fiscal policy work with such long lags that attempts at stabilizing the economy often end up being destabilizing.

Key Concepts

theory of liquidity preference, p. 748
fiscal policy, p. 757
multiplier effect, p. 758
crowding-out effect, p. 761
automatic stabilizers, p. 767

Questions for Review

1. What is the theory of liquidity preference? How does it help explain the downward slope of the aggregate-demand curve?
2. Use the theory of liquidity preference to explain how a decrease in the money supply affects the aggregate-demand curve.
3. The government spends $3 billion to buy police cars. Explain why aggregate demand might increase by more or less than $3 billion.
4. Suppose that survey measures of consumer confidence indicate a wave of pessimism is sweeping the country. If policymakers do nothing, what will happen to aggregate demand? What should the Fed do if it wants to stabilize aggregate demand? If the Fed does nothing, what might Congress do to stabilize aggregate demand? Explain your reasoning.
5. Give an example of a government policy that acts as an automatic stabilizer. Explain why the policy has this effect.

Problems and Applications

1. Explain how each of the following developments would affect the supply of money, the demand for money, and the interest rate. Use diagrams to illustrate your answers.
   a. The Fed's bond traders buy bonds in open-market operations.
   b. An increase in credit-card availability reduces the amount of cash people want to hold.
   c. The Fed reduces the interest rate it pays on reserves.
   d. Households decide to hold more money to use for holiday shopping.
   e. A wave of optimism boosts business investment and expands aggregate demand.
2. The Fed expands the money supply by 5 percent.
   a. Use the theory of liquidity preference to illustrate in a graph the impact of this policy on the interest rate.
   b. Use the model of aggregate demand and aggregate supply to illustrate the impact of this change in the interest rate on output and the price level in the short run.
   c. When the economy makes the transition from its short-run equilibrium to its new long-run equilibrium, what happens to the price level?
   d. How does this change in the price level affect the demand for money and the equilibrium interest rate?
   e. Is this analysis consistent with the proposition that money has real effects in the short run but is neutral in the long run?
3. Suppose a computer virus disables automatic teller machines and disrupts online payment systems. As a result, people want to keep more cash on hand, increasing the demand for money.
   a. Assume the Fed does not change the money supply. According to the theory of liquidity preference, what happens to the interest rate? What happens to aggregate demand?
   b. If, instead, the Fed wants to stabilize aggregate demand, how should it change the money supply?
   c. If the Fed wants to accomplish this change in the money supply using open-market operations, what should it do?
4. Consider two policies—a tax cut that lasts for only one year and a tax cut that is expected to be permanent. Which policy will stimulate greater spending by consumers? Which policy will have the greater impact on aggregate demand? Explain.
5. The economy is in a recession with high unemployment and low output.
   a. Draw a graph of aggregate demand and aggregate supply to illustrate the current situation. Be sure to include the aggregate-demand curve, the short-run aggregate-supply curve, and the long-run aggregate-supply curve.
   b. Identify an open-market operation that would restore the economy to its natural rate.
   c. Draw a graph of the money market to illustrate the effect of this open-market operation. Show the resulting change in the interest rate.
   d. Draw a graph similar to the one in part (a) to show the effect of the open-market operation on output and the price level. Explain in words why the policy has the effect you have shown in the graph.
6. In the 1970s and early 1980s, new legislation allowed banks to pay interest on certain checking deposits, which they could not do previously.
   a. If we define money to include checking deposits, what effect did this legislation have on money demand? Explain.
   b. If the Fed had maintained a constant money supply in the face of this change, what would have happened to the interest rate? What would

have happened to aggregate demand and aggregate output?

c. If the Fed had maintained a constant market interest rate (the interest rate on nonmonetary assets) in the face of this change, what change in the money supply would have been necessary? What would have happened to aggregate demand and aggregate output?

7. Suppose economists observe that an increase in government spending of $10 billion raises the total demand for goods and services by $30 billion.
   a. If these economists ignore the possibility of crowding out, what would they estimate the marginal propensity to consume (*MPC*) to be?
   b. Now suppose the economists allow for crowding out. Would their new estimate of the *MPC* be larger or smaller than their initial one?

8. An economy is producing output $400 billion less than the natural level of output, and fiscal policymakers want to close this recessionary gap. The central bank agrees to adjust the money supply to hold the interest rate constant, so there is no crowding out. The marginal propensity to consume is ⅘, and the price level is completely fixed in the short run. In what direction and by how much must government spending change to close the recessionary gap? Explain your thinking.

9. Suppose government spending increases. Would the effect on aggregate demand be larger if the Fed held the money supply constant in response or if the Fed committed to maintaining a fixed interest rate? Explain.

10. Is expansionary fiscal policy more likely to lead to a short-run increase in investment
    a. when the investment accelerator is large or when it is small? Explain.
    b. when the interest sensitivity of investment is large or when it is small? Explain.

11. Consider an economy described by the following equations:

$$Y = C + I + G$$
$$C = 100 + 0.75(Y - T)$$
$$I = 500 - 50r$$
$$G = 125$$
$$T = 100$$

where Y is GDP, C is consumption, I is investment, G is government purchases, T is taxes, and r is the interest rate. If the economy were at full employment (that is, at its natural level of output), GDP would be 2,000.
    a. Explain the meaning of each of these equations.
    b. What is the marginal propensity to consume in this economy?
    c. Suppose the central bank adjusts the money supply to maintain the interest rate at 4 percent, so $r = 4$. Solve for GDP. How does it compare to the full-employment level?
    d. Assuming no change in monetary policy, what change in government purchases would restore full employment?
    e. Assuming no change in fiscal policy, what change in the interest rate would restore full employment?

QuickQuiz Answers

1. **b** 2. **c** 3. **b** 4. **b** 5. **d** 6. **c** 7. **d** 8. **d** 9. **a**

Chapter

36

The Short-Run Trade-Off between Inflation and Unemployment

Inflation and unemployment are two of the most closely watched indicators of economic performance. When the Bureau of Labor Statistics releases data on these variables each month, policymakers, stock and bond traders, and journalists often consider it big news. Some commentators have added together the inflation rate and the unemployment rate to produce a **misery index**, which they use to gauge the health of the economy.

How are these two measures of economic performance related? Earlier in the book, we discussed the long-run determinants of unemployment and inflation. The natural rate of unemployment depends on various features of the labor market, such as job search, minimum-wage laws, union power, and efficiency wages. By contrast, the inflation rate depends primarily on growth in the money supply, which a nation's central bank controls. In the long run, inflation and unemployment are largely unrelated problems.

In the short run, though, the opposite is true. One of the **Ten Principles of Economics** in Chapter 1 is that society faces a short-run trade-off between inflation and unemployment. If monetary and fiscal policymakers expand aggregate demand and move the

ROMAN SAMBORSKYI/SHUTTERSTOCK.COM; ARNO VAN ENGELAND/EYEEM/GETTY IMAGES

economy up along the short-run aggregate-supply curve, they can expand output and reduce unemployment for a while, but only at the cost of a more rapidly rising price level. If policymakers contract aggregate demand and move the economy down the short-run aggregate-supply curve, they can reduce inflation, but only at the cost of temporarily lower output and higher unemployment.

This chapter examines the inflation–unemployment trade-off more closely. The relationship between inflation and unemployment has attracted economists' attention for many years. The best way to understand this relationship is to see how their thinking about it has evolved. As we will see, the history of thought regarding inflation and unemployment since the 1950s is inextricably connected to the history of the U.S. economy. These two histories show why the trade-off between inflation and unemployment holds in the short run, why it does not hold in the long run, and what issues the trade-off raises for policymakers.

36-1 The Phillips Curve

Phillips curve
a curve that shows the short-run trade-off between inflation and unemployment

"Probably the single most important macroeconomic relationship is the Phillips curve." These are the words of the economist George Akerlof from the lecture he gave when he received the Nobel Prize in 2001. The **Phillips curve** is the short-run relationship between inflation and unemployment. Our story begins with the discovery of the Phillips curve and its migration to America.

36-1a Origins of the Phillips Curve

In 1958, the economist A. W. Phillips published an article in the British journal *Economica* that would make him famous. The article was titled "The Relationship between Unemployment and the Rate of Change of Money Wages in the United Kingdom, 1861–1957." For the period he studied, Phillips showed a negative correlation between the rate of unemployment and the rate of inflation. That is, Phillips showed that years with low unemployment tend to have high inflation, and years with high unemployment tend to have low inflation. (Phillips examined inflation in nominal wages rather than inflation in prices. For our purposes, the distinction is not important because these two measures of inflation usually move together.) Phillips concluded that two important macroeconomic variables—inflation and unemployment—were linked in a way that economists had not previously appreciated.

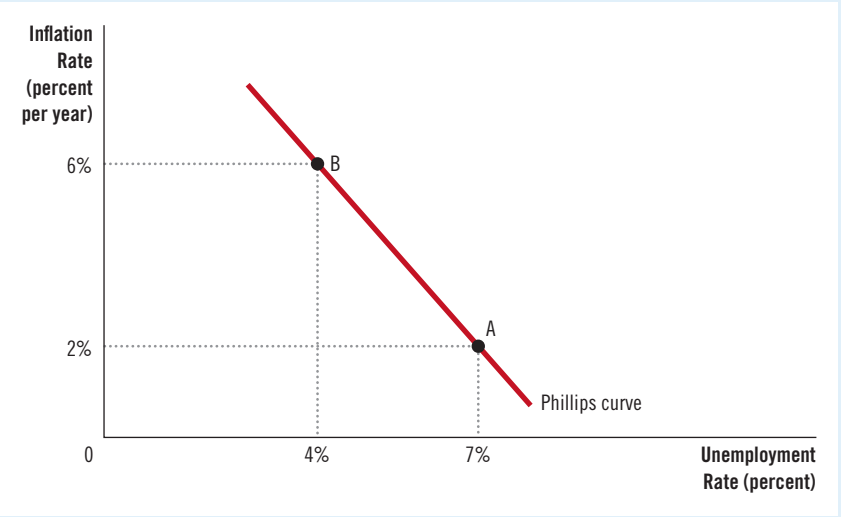
Phillips's discovery was based on data for the United Kingdom, but researchers quickly extended his finding to other countries. Two years after Phillips published his article, the economists Paul Samuelson and Robert Solow published an article in the *American Economic Review* called "Analytics of Anti-Inflation Policy" in which they showed a similar negative correlation between inflation and unemployment in data for the United States. They reasoned that this correlation arose because low unemployment was associated with high aggregate demand, which in turn put upward pressure on wages and prices throughout the economy. Samuelson and Solow dubbed the negative association between inflation and unemployment the **Phillips curve**. Figure 1 shows an example of a Phillips curve like the one found by Samuelson and Solow.

As the title of their paper suggests, Samuelson and Solow believed that Phillips's discovery held crucial lessons for policymakers. The Phillips curve, they suggested, was a menu of possible economic outcomes. By altering monetary and fiscal policy to influence aggregate demand, policymakers could choose any point on this curve.

Figure 1

The Phillips Curve

The Phillips curve illustrates a negative association between the inflation rate and the unemployment rate. At point A, inflation is low, and unemployment is high. At point B, inflation is high, and unemployment is low.



Point A offers high unemployment and low inflation. Point B offers low unemployment and high inflation. Policymakers might prefer both low inflation and low unemployment, but the historical data summarized by the Phillips curve indicated that this combination was unattainable. According to Samuelson and Solow, policymakers face a trade-off between inflation and unemployment, and the Phillips curve illustrates that trade-off.

36-1b Aggregate Demand, Aggregate Supply, and the Phillips Curve

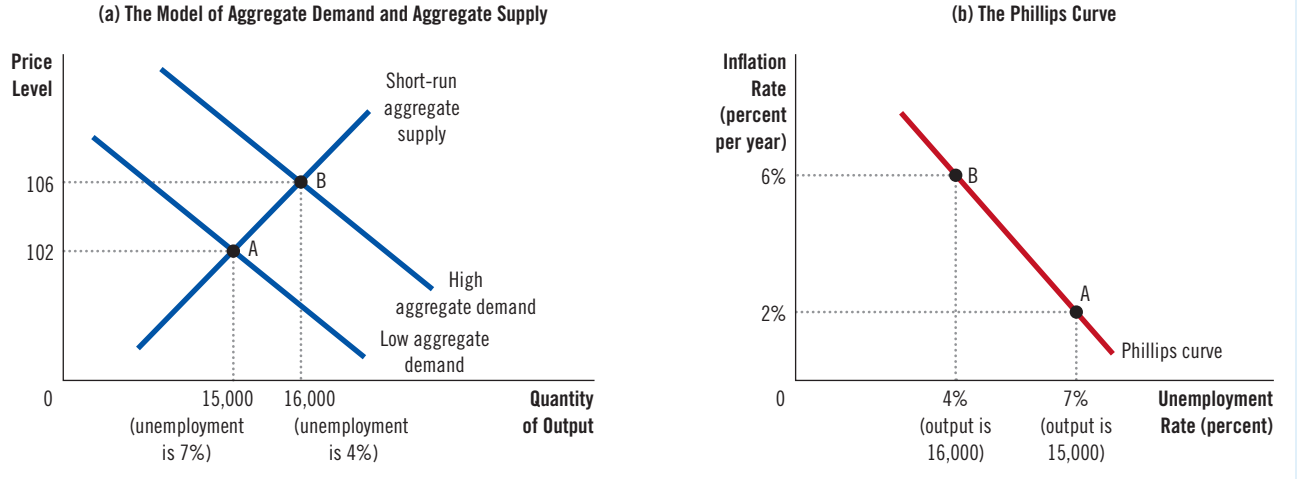
The menu of possible outcomes described by the Phillips curve can be explained by the model of aggregate demand and aggregate supply. **The Phillips curve shows the combinations of inflation and unemployment that arise in the short run as shifts in the aggregate-demand curve move the economy along the short-run aggregate-supply curve.** The previous two chapters showed that an increase in the aggregate demand for goods and services leads, in the short run, to a larger output of goods and services and a higher price level. Larger output means greater employment and a lower rate of unemployment. In addition, a higher price level translates into a higher rate of inflation. Thus, shifts in aggregate demand push inflation and unemployment in opposite directions in the short run—a relationship illustrated by the Phillips curve.

To see more fully how this works, consider an example. To keep the numbers simple, imagine that the price level (as measured, for instance, by the consumer price index) equals 100 in the year 2025. Figure 2 shows two possible outcomes that might occur in the year 2026 depending on the strength of aggregate demand. One outcome occurs if aggregate demand is high, and the other occurs if aggregate demand is low. Panel (a) shows these two outcomes using the model of aggregate demand and aggregate supply. Panel (b) shows the same two outcomes using the Phillips curve.

Figure 2

How the Phillips Curve Is Related to the Model of Aggregate Demand and Aggregate Supply

This figure assumes a price level of 100 for the year 2025 and charts possible outcomes for the year 2026. Panel (a) shows the model of aggregate demand and aggregate supply. If aggregate demand is low, the economy is at point A. Output is low (15,000), and the price level is low (102). If aggregate demand is high, the economy is at point B. Output is high (16,000), and the price level is high (106). Panel (b) shows the implications for the Phillips curve. Point A, which arises when aggregate demand is low, has high unemployment (7 percent) and low inflation (2 percent). Point B, which arises when aggregate demand is high, has low unemployment (4 percent) and high inflation (6 percent).



Panel (a) of the figure shows what happens to output and the price level in the year 2026. If the aggregate demand for goods and services is low, the economy experiences outcome A. The economy produces output of 15,000, and the price level is 102. By contrast, if aggregate demand is high, the economy experiences outcome B. Output is 16,000, and the price level is 106. This is an example of a familiar conclusion: Higher aggregate demand moves the economy to an equilibrium with higher output and a higher price level.

Panel (b) shows what these two possible outcomes mean for unemployment and inflation. Because firms need more workers when they produce a greater output of goods and services, unemployment is lower in outcome B than in outcome A. In this example, when output rises from 15,000 to 16,000, unemployment falls from 7 percent to 4 percent. Moreover, because the price level is higher at outcome B than at outcome A, the inflation rate (the percentage change in the price level from the previous year) is also higher. In this case, the price level was 100 in the year 2025, so outcome A has an inflation rate of 2 percent, and outcome B has an inflation rate of 6 percent. The two possible outcomes for the economy can be compared either in terms of output and the price level (using the model of aggregate demand and aggregate supply) or in terms of unemployment and inflation (using the Phillips curve).

Because monetary and fiscal policy can shift the aggregate-demand curve, they can move an economy along the Phillips curve. Increases in the money supply,

increases in government spending, or cuts in taxes expand aggregate demand and move the economy to a point on the Phillips curve with higher inflation and lower unemployment. Decreases in the money supply, cuts in government spending, or increases in taxes contract aggregate demand and move the economy to a point on the Phillips curve with lower inflation and higher unemployment. In this sense, the Phillips curve offers policymakers a menu of combinations of inflation and unemployment.

QuickQuiz

1. The Phillips curve started as an observed _____ correlation between the inflation rate and the _____.
   a. positive; nominal interest rate
   b. positive; unemployment rate
   c. negative; nominal interest rate
   d. negative; unemployment rate

2. When the Federal Reserve increases the money supply and expands aggregate demand, it moves the economy along the Phillips curve to a point with _____ inflation and _____ unemployment.
   a. higher; higher
   b. higher; lower
   c. lower; higher
   d. lower; lower

Answers are at the end of the chapter.

36-2 Shifts in the Phillips Curve: The Role of Expectations

The Phillips curve seems to offer policymakers a menu of inflation–unemployment outcomes, but this raises a question: Does this set of possible choices remain constant over time? In other words, is the downward-sloping Phillips curve a stable relationship on which policymakers can rely? Economists took up this issue in the late 1960s, shortly after Samuelson and Solow introduced the Phillips curve into the macroeconomic policy debate.

36-2a The Long-Run Phillips Curve

In 1968, the economist Milton Friedman published a paper in the *American Economic Review* based on an address he had recently given as president of the American Economic Association. The paper, titled "The Role of Monetary Policy," contained sections on "What Monetary Policy Can Do" and "What Monetary Policy Cannot Do." Friedman argued that one thing monetary policy cannot do, other than for a short time, is lower unemployment by raising inflation. At about the same time, another economist, Edmund Phelps, reached the same conclusion. Like Friedman, Phelps published a paper denying the existence of a long-run trade-off between inflation and unemployment.

Both Friedman and Phelps based their conclusions on classical principles of macroeconomics. Classical theory points to growth in the money supply as the primary determinant of inflation. But classical theory also states that monetary growth does not affect real variables such as output and employment; it merely alters all prices and nominal incomes proportionately. In particular, monetary growth does not influence the factors that determine the economy's unemployment rate, such as job search and efficiency wages. Friedman and Phelps concluded that, in the long run, the rate of inflation and the rate of unemployment would not be related.

Here, in his own words, is Friedman's view about what the Federal Reserve can hope to accomplish for the economy in the long run:

> The monetary authority controls nominal quantities—directly, the quantity of its own liabilities [currency plus bank reserves]. In principle, it can use this control to peg a nominal quantity—an exchange rate, the price level, the nominal level of national income, the quantity of money by one definition or another—or to peg the change in a nominal quantity—the rate of inflation or deflation, the rate of growth or decline in nominal national income, the rate of growth of the quantity of money. It cannot use its control over nominal quantities to peg a real quantity—the real rate of interest, the rate of unemployment, the level of real national income, the real quantity of money, the rate of growth of real national income, or the rate of growth of the real quantity of money.

According to Friedman, monetary policymakers face a long-run Phillips curve that is vertical, as in Figure 3. If the Fed increases the money supply slowly, the inflation rate is low, and the economy finds itself at point A. If the Fed increases the money supply quickly, the inflation rate is high, and the economy finds itself at point B. In either case, the unemployment rate tends toward its normal level, called the **natural rate of unemployment**. The vertical long-run Phillips curve illustrates the conclusion that unemployment does not depend on money growth and inflation in the long run.

The vertical long-run Phillips curve is, in essence, an expression of the classical idea of monetary neutrality. Previously, we expressed monetary neutrality with a vertical long-run aggregate-supply curve. Figure 4 shows that the vertical long-run Phillips curve and the vertical long-run aggregate-supply curve are two sides of the same coin. In panel (a) of this figure, an increase in the money supply shifts the aggregate-demand curve to the right from AD_1 to AD_2. Because of this shift, the long-run equilibrium moves from point A to point B. The price level rises from P_1 to P_2, but because the aggregate-supply curve is vertical, output remains the same. In panel (b), more rapid growth in the money supply raises the inflation rate by

Figure 3

The Long-Run Phillips Curve

According to Friedman and Phelps, there is no trade-off between inflation and unemployment in the long run. Growth in the money supply determines the inflation rate. Regardless of the inflation rate, the unemployment rate gravitates toward its natural rate. As a result, the long-run Phillips curve is vertical.

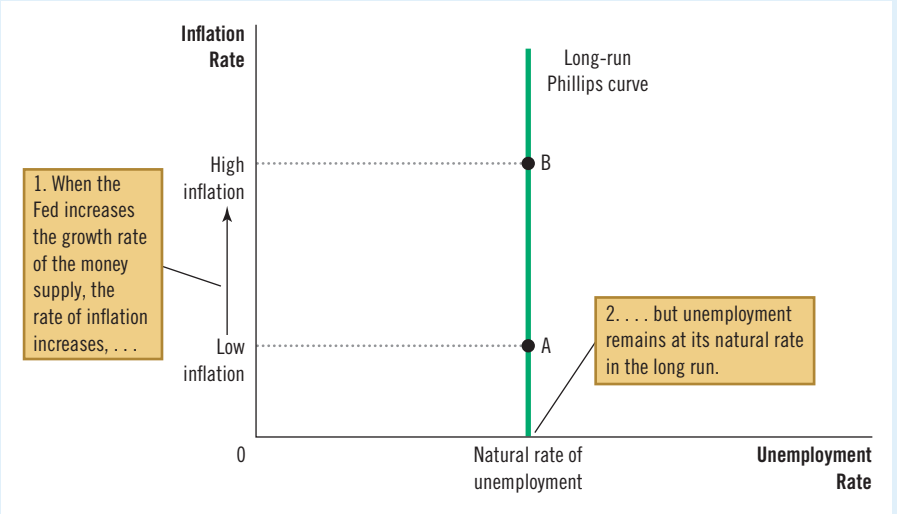
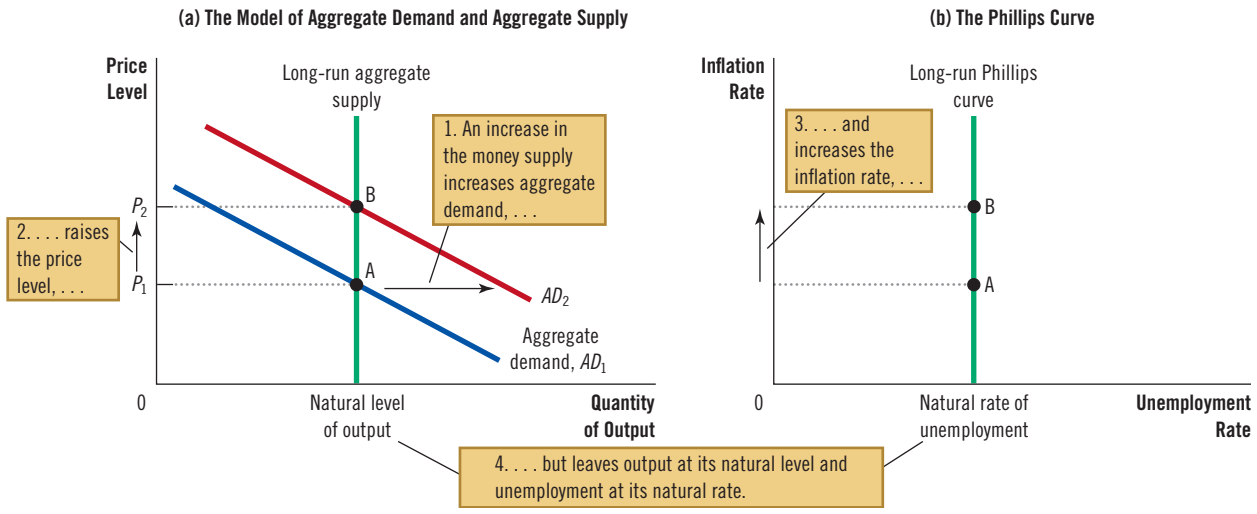


Figure 4

How the Long-Run Phillips Curve Is Related to the Model of Aggregate Demand and Aggregate Supply

Panel (a) shows the model of aggregate demand and aggregate supply with a vertical aggregate-supply curve. When expansionary monetary policy shifts the aggregate-demand curve to the right from AD_1 to AD_2, the equilibrium moves from point A to point B. The price level rises from P_1 to P_2, while output remains the same. Panel (b) shows the long-run Phillips curve, which is vertical at the natural rate of unemployment. In the long run, expansionary monetary policy moves the economy from lower inflation (point A) to higher inflation (point B) without changing the rate of unemployment.



moving the economy from point A to point B. But because the Phillips curve is vertical, the rate of unemployment is the same at these two points. Both the vertical long-run aggregate-supply curve and the vertical long-run Phillips curve imply that monetary policy influences nominal variables (the price level and the inflation rate) but not real ones (output and unemployment). In the long run, regardless of the monetary policy pursued by the Fed, output is at its natural level, and unemployment is at its natural rate.

36-2b The Meaning of "Natural"

What is so "natural" about the natural rate of unemployment? Friedman and Phelps used this adjective to describe the unemployment rate toward which the economy gravitates in the long run. Yet the natural rate of unemployment is not necessarily the socially desirable rate of unemployment. Nor is the natural rate of unemployment constant over time.

For example, suppose that a newly formed union uses its market power to raise the real wages of some workers above the equilibrium level. The result is an excess supply of workers and a higher natural rate of unemployment. This unemployment is natural not because it is good but because it is beyond the influence of monetary policy. More rapid money growth would reduce neither the market power of the union nor the level of unemployment; it would lead only to more inflation.

Although monetary policy cannot influence the natural rate of unemployment, other types of policy can. To reduce the natural rate of unemployment, policymakers might look to policies that improve the functioning of the labor market. Earlier in the book, we discussed how various labor-market policies, such as minimum-wage laws, collective-bargaining laws, unemployment insurance, and job-training programs, can affect the natural rate of unemployment. A policy change that reduces the natural rate of unemployment would shift the long-run Phillips curve to the left. In addition, because lower unemployment means more workers are producing goods and services, the quantity of goods and services supplied would be larger at any given price level, and the long-run aggregate-supply curve would shift to the right. The economy could then enjoy lower unemployment and higher output for any given rate of money growth and inflation.

36-2c Reconciling Theory and Evidence

At first, Friedman and Phelps's conclusion that there is no long-run trade-off between inflation and unemployment might not seem convincing. Their argument was based on an appeal to **theory**, specifically classical theory's prediction of monetary neutrality. By contrast, the negative correlation between inflation and unemployment documented by Phillips, Samuelson, and Solow was based on actual **evidence** from the real world. Why should anyone believe that policymakers faced a vertical Phillips curve when the world seemed to offer a downward-sloping one? Shouldn't the findings of Phillips, Samuelson, and Solow lead us to reject monetary neutrality?

Friedman and Phelps were aware of these questions, and they offered a way to reconcile classical macroeconomic theory with the finding of a downward-sloping Phillips curve in data from the United Kingdom and the United States. They claimed that a negative relationship between inflation and unemployment exists in the short run but that it cannot be used by policymakers as a menu of outcomes in the long run. Policymakers can pursue expansionary monetary policy to achieve lower unemployment for a while, but eventually, unemployment will return to its natural rate. In the long run, more expansionary monetary policy leads only to higher inflation.

Friedman and Phelps's work was the basis of our discussion of the difference between the short-run and long-run aggregate-supply curves in Chapter 34. As you may recall, the long-run aggregate-supply curve is vertical, indicating that the price level does not influence quantity supplied in the long run. But the short-run aggregate-supply curve slopes upward, indicating that an increase in the price level raises the quantity of goods and services that firms supply. According to the sticky-wage theory of aggregate supply, for instance, nominal wages are set in advance based on the price level that workers and firms expect to prevail. When prices turn out to be higher than expected, firms have an incentive to increase production and employment; when prices are lower than expected, firms reduce production and employment. Yet because the expected price level and nominal wages will eventually adjust, the positive relationship between the actual price level and quantity supplied exists only in the short run.

Friedman and Phelps applied this same logic to the Phillips curve. Just as the aggregate-supply curve slopes upward only in the short run, the trade-off between inflation and unemployment holds only in the short run. And just as the long-run aggregate-supply curve is vertical, the long-run Phillips curve is also vertical. Once again, expectations are the key to understanding how the short run and the long run are related.

Friedman and Phelps introduced a new variable into the analysis of the inflation–unemployment trade-off: **expected inflation**. Expected inflation measures how much people expect the overall price level to change. Because the expected price level affects nominal wages, expected inflation is one factor that determines the position of the short-run aggregate-supply curve. In the short run, the Fed can take expected inflation (and the short-run aggregate-supply curve) as already determined. When the money supply changes, the aggregate-demand curve shifts, and the economy moves along a given short-run aggregate-supply curve. In the short run, therefore, monetary changes lead to unexpected fluctuations in output, prices, unemployment, and inflation. In this way, Friedman and Phelps explained the downward-sloping Phillips curve that Phillips, Samuelson, and Solow had documented.

The Fed's ability to create unexpected inflation by increasing the money supply exists only in the short run. In the long run, people come to expect whatever inflation rate the Fed chooses to produce, and nominal wages will adjust to keep pace with inflation. As a result, the long-run aggregate-supply curve is vertical. Changes in aggregate demand, such as those caused by changes in monetary policy, affect neither the economy's output of goods and services nor the number of workers that firms need to hire to produce those goods and services. Friedman and Phelps concluded that unemployment returns to its natural rate in the long run.

36-2d The Short-Run Phillips Curve

The analysis of Friedman and Phelps can be summarized by the following equation:

$$\begin{pmatrix}\text{Unemployment}\\ \text{rate}\end{pmatrix} = \begin{pmatrix}\text{Natural rate of}\\ \text{unemployment}\end{pmatrix} - a\left(\begin{pmatrix}\text{Actual}\\ \text{inflation}\end{pmatrix} - \begin{pmatrix}\text{Expected}\\ \text{inflation}\end{pmatrix}\right).$$

This equation relates the unemployment rate to the natural rate of unemployment, actual inflation, and expected inflation. It is, in essence, another expression of the aggregate-supply equation we have seen previously. In the short run, expected inflation is given, so higher actual inflation is associated with lower unemployment. (The variable a is a parameter that measures how much unemployment responds to unexpected inflation.) In the long run, people come to expect whatever inflation the Fed produces, so actual inflation equals expected inflation, and unemployment is at its natural rate.

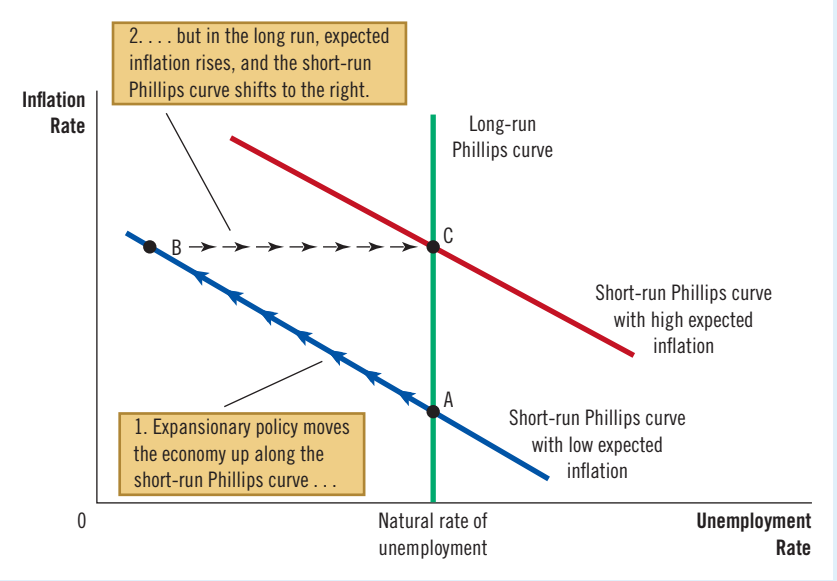
This equation implies there can be no stable short-run Phillips curve. Each short-run Phillips curve reflects a particular expected rate of inflation. (To be precise, if you graph the equation, you'll find that the downward-sloping short-run Phillips curve intersects the vertical long-run Phillips curve at the expected rate of inflation.) When expected inflation changes, the short-run Phillips curve shifts.

According to Friedman and Phelps, it is dangerous to view the Phillips curve as a menu of options available to policymakers. To see why, imagine an economy that starts with low inflation, with an equally low rate of expected inflation, and with unemployment at its natural rate. In Figure 5, the economy is at point A. Now suppose that policymakers try to take advantage of the trade-off between inflation and unemployment by using monetary or fiscal policy to expand aggregate demand. In the short run, when expected inflation is given, the economy goes from point A to point B. Unemployment falls below its natural rate, and the actual inflation rate rises above expected inflation. As the economy moves from

Figure 5

How Expected Inflation Shifts the Short-Run Phillips Curve

The higher the expected rate of inflation, the higher the curve representing the short-run trade-off between inflation and unemployment will be. At point A, expected inflation and actual inflation are equal at a low rate, and unemployment is at its natural rate. If the Fed pursues an expansionary monetary policy, the economy moves from point A to point B in the short run. At point B, expected inflation is still low, but actual inflation is high. Unemployment is below its natural rate. In the long run, expected inflation rises, and the economy moves to point C. At point C, expected inflation and actual inflation are both high, and unemployment is back to its natural rate.



point A to point B, policymakers might think they have achieved permanently lower unemployment at the cost of higher inflation—a bargain that, if possible, might be worth making.

This situation, however, will not persist. Over time, people get used to this higher inflation rate and raise their expectations of inflation. When expected inflation rises, firms and workers start taking higher inflation into account when setting wages and prices. The short-run Phillips curve then shifts to the right, as shown in the figure. The economy ends up at point C, with higher inflation than at point A but with the same level of unemployment. Thus, Friedman and Phelps concluded that policymakers face only a temporary trade-off between inflation and unemployment. In the long run, expanding aggregate demand more rapidly will yield higher inflation without any reduction in unemployment.

36-2e The Natural Experiment for the Natural-Rate Hypothesis

Friedman and Phelps made a bold prediction in 1968: If policymakers try to take advantage of the Phillips curve by choosing higher inflation to reduce unemployment, they will succeed at reducing unemployment only temporarily. This view—that unemployment eventually returns to its natural rate regardless of the rate of inflation—is called the **natural-rate hypothesis**. A few years after Friedman and Phelps proposed this hypothesis, monetary and fiscal policymakers inadvertently tested it in a natural experiment. Their laboratory was the U.S. economy.

natural-rate hypothesis the claim that unemployment eventually returns to its normal, or natural, rate, regardless of the rate of inflation

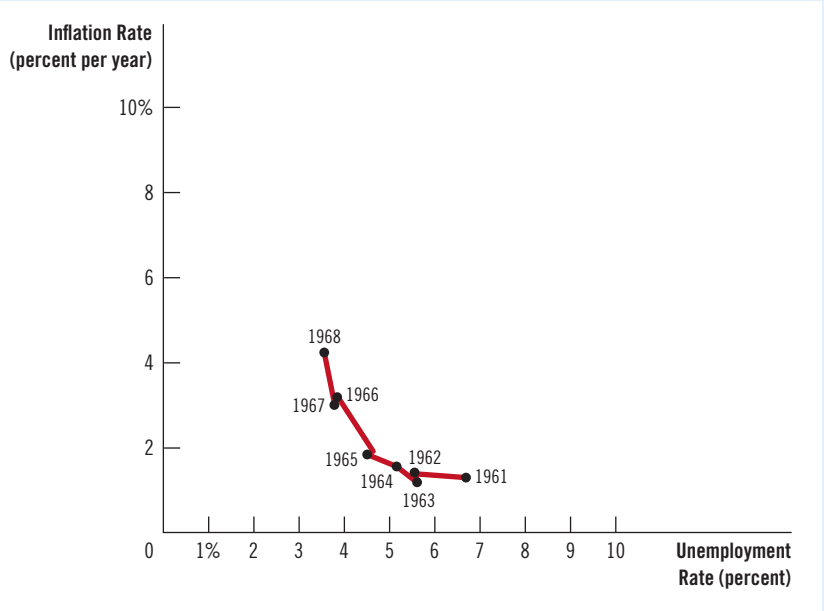
Before examining the outcome of this test, let's look at the data that Friedman and Phelps were seeing when they made their prediction in 1968. Figure 6 shows the unemployment and inflation rates from 1961 to 1968. These data trace out an almost perfect Phillips curve. As inflation rose over these eight years, unemployment fell. The economic data from this era seemed to confirm that policymakers faced a trade-off between inflation and unemployment.

Figure 6

The Phillips Curve in the 1960s

This figure uses annual data from 1961 to 1968 on the unemployment rate and on the inflation rate (as measured by the GDP deflator) to show the negative relationship between inflation and unemployment.

Source: U.S. Department of Labor; U.S. Department of Commerce.



The apparent success of the Phillips curve in the 1960s made the prediction of Friedman and Phelps all the bolder. In 1958, Phillips suggested a negative association between inflation and unemployment. In 1960, Samuelson and Solow showed that it existed in U.S. data. Another decade of data confirmed the relationship. To some economists at the time, it seemed ridiculous to claim that the reliable Phillips curve would start shifting once policymakers tried to take advantage of it.

But that is exactly what happened. Beginning in the late 1960s, government policies expanded the aggregate demand for goods and services. In part, this expansion was due to fiscal policy: Government spending rose as the Vietnam War heated up. In part, it was due to monetary policy: Because the Fed tried to hold down interest rates in the face of expansionary fiscal policy, the money supply (as measured by M2) rose about 13 percent per year from 1970 to 1972, compared with 7 percent per year in the early 1960s. As a result, inflation stayed high (about 5 to 6 percent per year in the late 1960s and early 1970s, compared with about 1 to 2 percent per year in the early 1960s). But as Friedman and Phelps had predicted, unemployment did not stay low.

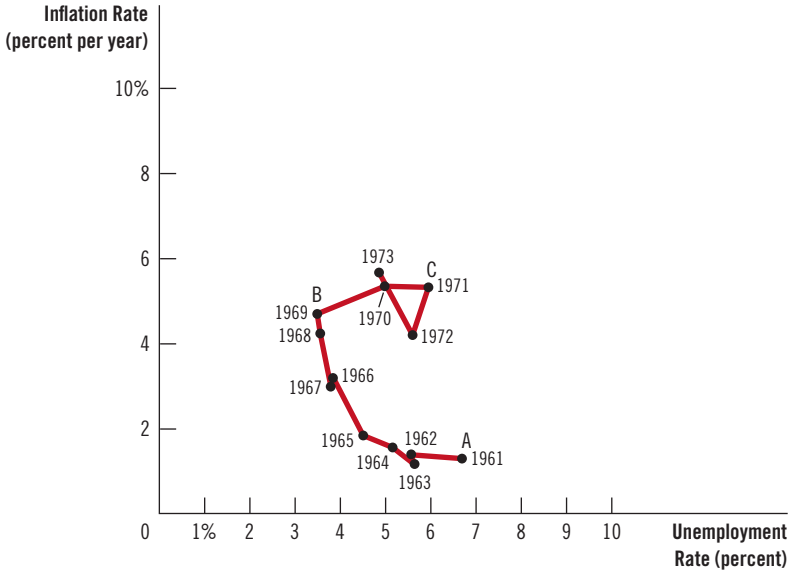
Figure 7 displays the history of inflation and unemployment from 1961 to 1973. It shows that the simple negative relationship between these two variables started to break down around 1970. As inflation remained high in the early 1970s, people's expectations of inflation caught up with reality, and the unemployment rate reverted to the 5 percent to 6 percent range that had prevailed in the early 1960s. Notice that the history illustrated in Figure 7 resembles the theory of a shifting short-run Phillips curve shown in Figure 5. By 1973, policymakers learned that Friedman and Phelps were right: There is no trade-off between inflation and unemployment in the long run.

Figure 7

The Breakdown of the Phillips Curve

This figure shows annual data from 1961 to 1973 on the unemployment rate and on the inflation rate (as measured by the GDP deflator). The Phillips curve of the 1960s breaks down in the early 1970s, just as Friedman and Phelps had predicted. Notice that the points labeled A, B, and C in this figure correspond roughly to the points in Figure 5.

Source: U.S. Department of Labor; U.S. Department of Commerce.



QuickQuiz

3. The natural rate of unemployment is
   a. the socially optimal level of joblessness.
   b. the level of joblessness the economy reaches in the short run.
   c. the amount of joblessness that cannot be reduced by public policies.
   d. the normal level of joblessness, regardless of inflation.

4. If the Federal Reserve reduces the rate of money growth and maintains it at the new lower rate, eventually, expected inflation will _____, and the short-run Phillips curve will shift _____.
   a. decrease; downward
   b. decrease; upward
   c. increase; downward
   d. increase; upward

Answers are at the end of the chapter.

36-3 Shifts in the Phillips Curve: The Role of Supply Shocks

Friedman and Phelps suggested in 1968 that changes in expected inflation shift the short-run Phillips curve, and the experience of the early 1970s convinced most economists that Friedman and Phelps were right. Within a few years, however, the economics profession would turn its attention to a different source of shifts in the short-run Phillips curve: shocks to aggregate supply.

This time, the impetus for the change came not from two economics professors but from an oil cartel. In 1974, the Organization of Petroleum Exporting Countries (OPEC) began to exert its market power. The OPEC countries, including Saudi Arabia, Kuwait, and Iraq, restricted the amount of crude oil they pumped and sold on world markets. Within a few years, this reduction in supply caused the world price of oil to almost double.

A large increase in the world price of oil is an example of a **supply shock**, an event that directly affects firms' costs of production and the prices they charge. Supply shocks shift the economy's aggregate-supply curve and, as a result, the Phillips curve. For example, when a cartel increases the price of crude oil, it raises the cost of producing gasoline, heating oil, tires, and other products, reducing the quantity of goods and services supplied at any given price level. As panel (a) of Figure 8 shows, this reduction in supply is represented by the leftward shift in the aggregate-supply curve from AS_1 to AS_2. Output falls from Y_1 to Y_2, and the price level rises from P_1 to P_2. The economy experiences **stagflation**—a pernicious combination of falling output (stagnation) and rising prices (inflation).

supply shock
an event that directly alters firms' costs and prices, shifting the economy's aggregate-supply curve and thus the Phillips curve

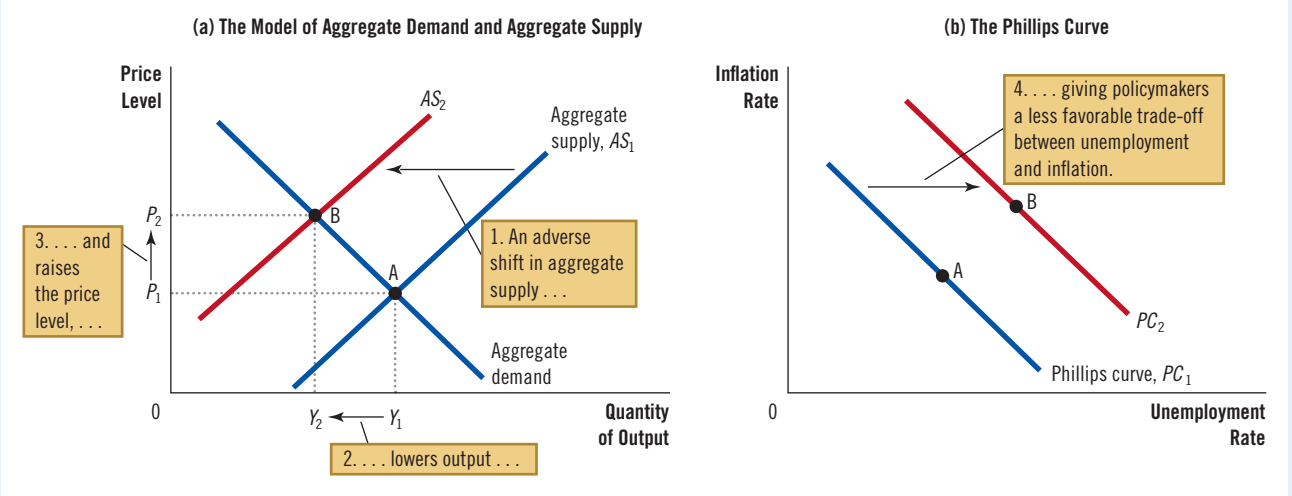
This shift in aggregate supply is associated with a similar shift in the short-run Phillips curve, shown in panel (b). Because firms need fewer workers to produce the smaller output, employment falls, and unemployment rises. And because the price level is higher, the inflation rate—the percentage change in the price level from the previous year—is also higher. Thus, the shift in aggregate supply leads to higher unemployment and higher inflation. The short-run trade-off between inflation and unemployment shifts to the right from PC_1 to PC_2.

Confronted with an adverse shift in aggregate supply, policymakers face a difficult choice between fighting inflation and fighting unemployment. If they contract aggregate demand to fight inflation, they will raise unemployment further. If they expand aggregate demand to fight unemployment, they will raise inflation further. In other words, policymakers face a less favorable trade-off between inflation and unemployment than they did before the shift in aggregate supply: They have

Figure 8

An Adverse Shock to Aggregate Supply

Panel (a) shows the model of aggregate demand and aggregate supply. When the aggregate-supply curve shifts to the left from AS_1 to AS_2, the equilibrium moves from point A to point B. Output falls from Y_1 to Y_2, and the price level rises from P_1 to P_2. Panel (b) shows the short-run trade-off between inflation and unemployment. The adverse shift in aggregate supply moves the economy from a point with lower unemployment and lower inflation (point A) to a point with higher unemployment and higher inflation (point B). The short-run Phillips curve shifts to the right from PC_1 to PC_2. Policymakers now face a worse set of options for inflation and unemployment.



to live with a higher rate of inflation for a given rate of unemployment, a higher rate of unemployment for a given rate of inflation, or some combination of higher unemployment and higher inflation.

Faced with such an adverse shift in the Phillips curve, policymakers will ask whether the shift is temporary or permanent. The answer depends on how people adjust their expectations of inflation. If people view the rise in inflation from the supply shock as a temporary aberration, expected inflation will not change, and the Phillips curve will soon revert to its former position. But if people believe the shock will lead to a new era of higher inflation, then expected inflation will rise, and the Phillips curve will remain at its new, less desirable position.

In the United States during the 1970s, expected inflation did rise substantially. This rise in expected inflation was partly attributable to the Fed's decision to accommodate the supply shock with higher money growth. (Recall that policymakers are said to **accommodate** an adverse supply shock when they respond to it by increasing aggregate demand in an effort to keep output from falling.) Because of this policy decision, the recession that resulted from the supply shock was smaller than it otherwise might have been, but the U.S. economy faced an unfavorable trade-off between inflation and unemployment for many years. The problem was compounded in 1979 when OPEC once again started to exert its market power, more than doubling the price of oil. Figure 9 shows inflation and unemployment in the U.S. economy during this period.

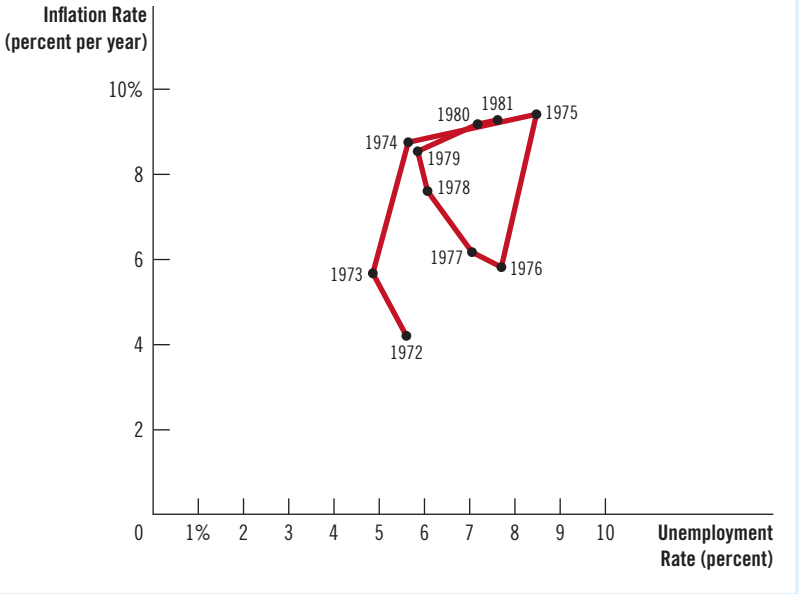
In 1980, after two OPEC supply shocks, the U.S. economy had an inflation rate of more than 9 percent and an unemployment rate of about 7 percent. This combination of inflation and unemployment was not at all near the trade-off that seemed possible in the 1960s. Back then, the Phillips curve suggested that an unemployment rate of 7 percent would be associated with an inflation rate of only 1 percent; inflation of more than 9 percent was unthinkable. With the misery index in 1980 near a historic high, many people were unhappy with the economy's performance. Something had to be done, and soon it would be.

Figure 9

The Supply Shocks of the 1970s

This figure shows annual data from 1972 to 1981 on the unemployment rate and on the inflation rate (as measured by the GDP deflator). In the periods 1973–1975 and 1978–1981, increases in world oil prices led to higher inflation and higher unemployment.

Source: U.S. Department of Labor; U.S. Department of Commerce.



QuickQuiz

5. When an adverse supply shock shifts the short-run aggregate-supply curve to the left, it also
   a. moves the economy along the short-run Phillips curve to a point with higher inflation and lower unemployment.
   b. moves the economy along the short-run Phillips curve to a point with lower inflation and higher unemployment.
   c. shifts the short-run Phillips curve to the right.
   d. shifts the short-run Phillips curve to the left.
6. From one year to the next, inflation falls from 5 to 4 percent, while unemployment rises from 6 to 7 percent. Which of the following events could be responsible for this change?
   a. The central bank increases the growth rate of the money supply.
   b. The government cuts spending and raises taxes to reduce the budget deficit.
   c. Newly discovered oil reserves cause world oil prices to plummet.
   d. The appointment of a new Fed chair increases expected inflation.
7. From one year to the next, inflation falls from 5 to 4 percent, while unemployment falls from 7 to 6 percent. Which of the following events could be responsible for this change?
   a. The central bank increases the growth rate of the money supply.
   b. The government cuts spending and raises taxes to reduce the budget deficit.
   c. Newly discovered oil reserves cause world oil prices to plummet.
   d. The appointment of a new Fed chair increases expected inflation.

Answers are at the end of the chapter.

36-4 The Cost of Reducing Inflation

In October 1979, as OPEC was imposing adverse supply shocks on the world's economies for the second time in a decade, Paul Volcker, then the Federal Reserve chair, decided that the time for action had come. Volcker had been appointed by President Carter only two months earlier, and he had taken the job knowing that inflation had reached unacceptable levels. As guardian of the nation's monetary system, he felt he had little choice but to pursue a policy of disinflation. **Disinflation** is a reduction in the rate of inflation, and it should not be confused with **deflation**, a reduction in the price level. To draw an analogy to a car's motion, disinflation is like slowing down, whereas deflation is like going in reverse. Volcker wanted the economy's rising level of prices to slow down.

He had no doubt that the Fed could use the tools of monetary policy to reduce inflation. But what would be the short-run cost? The answer to this question was much less certain.

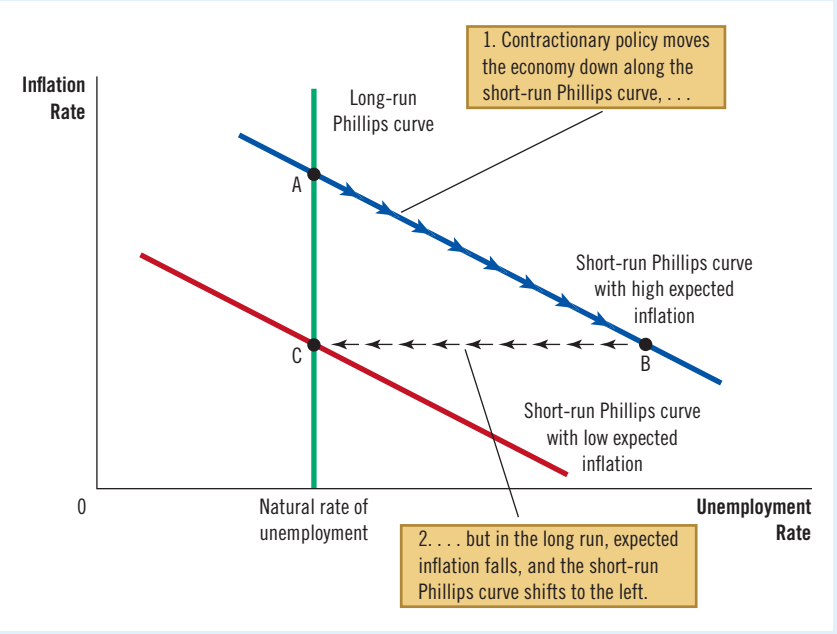
36-4a The Sacrifice Ratio

To reduce the inflation rate, the Fed has to pursue contractionary monetary policy. Figure 10 shows some of the effects of such a decision. Other things being equal, when the Fed slows growth in the money supply, it contracts aggregate demand. A fall in aggregate demand, in turn, reduces the quantity of goods and services that firms produce, leading to a rise in unemployment. The economy begins at point A in the figure and moves along the short-run Phillips curve to point B, which has lower inflation and higher unemployment. Over time, as people come to understand that prices are rising more slowly, expected inflation falls, and the short-run Phillips curve shifts downward. The economy moves from point B to point C. Inflation is lower than it was initially at point A, and unemployment is back at its natural rate.

Figure 10

Disinflationary Monetary Policy in the Short Run and Long Run

When the Fed pursues contractionary monetary policy to reduce inflation, the economy moves along a short-run Phillips curve from point A to point B. Over time, expected inflation falls, and the short-run Phillips curve shifts downward. When the economy reaches point C, unemployment is back at its natural rate.



Thus, if a nation wants to reduce inflation, it must endure a period of high unemployment and low output. In Figure 10, this cost is represented by the movement of the economy through point B as it travels from point A to point C. The size of this cost depends on the slope of the Phillips curve and how quickly expectations of inflation adjust to the new monetary policy.

Many studies have examined the data on inflation and unemployment to estimate the cost of reducing inflation. The findings of these studies are often summarized in a statistic called the **sacrifice ratio**. The sacrifice ratio is the number of percentage points of annual output lost in the process of reducing inflation by 1 percentage point. A typical estimate of the sacrifice ratio is 5. That is, for each percentage point that inflation is reduced, 5 percent of annual output must be sacrificed in the transition.

sacrifice ratio
the number of percentage points of annual output lost in the process of reducing inflation by 1 percentage point

Such estimates surely must have made Volcker apprehensive as he confronted the task of reducing inflation. Inflation was running at almost 10 percent per year. To reach moderate inflation of, say, 4 percent per year would mean reducing inflation by 6 percentage points. If each percentage point costs 5 percent of the economy's annual output, then reducing inflation by 6 percentage points would require sacrificing 30 percent of annual output.

According to studies of the Phillips curve and the cost of disinflation, this sacrifice could be paid in various ways. An immediate reduction in inflation would depress output by 30 percent for a single year, an outcome that was surely too harsh even for an inflation hawk like Volcker. It would be better, many argued, to spread out the cost over several years. If the reduction in inflation took place over five years, for instance, then output would have to average only 6 percent below trend during that period to add up to a sacrifice of 30 percent. An even more gradual approach would be to reduce inflation slowly over a decade so that output would have to be only 3 percent below trend. Whatever path was chosen, however, it seemed that reducing inflation would not be easy.

36-4b Rational Expectations and the Possibility of Costless Disinflation

Just as Volcker was pondering how costly reducing inflation might be, a group of economics professors was leading an intellectual revolution that would challenge the conventional wisdom on the sacrifice ratio. This group included such prominent economists as Robert Lucas, Thomas Sargent, and Robert Barro. Their revolution was based on a new approach to economic theory and policy called **rational expectations**. According to the theory of rational expectations, people optimally use all the information they have, including information about government policies, when forecasting the future.

rational expectations the theory that people optimally use all the knowledge they have, including information about government policies, when forecasting the future

This approach has had profound implications for many areas of macroeconomics, but none is more important than its application to the trade-off between inflation and unemployment. As Friedman and Phelps had first emphasized, expected inflation is an important variable that explains why there is a trade-off between inflation and unemployment in the short run but not in the long run. How quickly the short-run trade-off disappears depends on how quickly people adjust their expectations of inflation. Proponents of rational expectations expanded upon the Friedman-Phelps analysis to argue that when economic policies change, people adjust their expectations of inflation accordingly. The studies of inflation and unemployment that had tried to estimate the sacrifice ratio had failed to take account of the direct effect of the policy regime on expectations. As a result, estimates of the sacrifice ratio were, according to the rational-expectations theorists, unreliable guides for policy.

In a 1981 paper titled "The End of Four Big Inflations," Sargent described this new view as follows:

> An alternative "rational expectations" view denies that there is any inherent momentum to the present process of inflation. This view maintains that firms and workers have now come to expect high rates of inflation in the future and that they strike inflationary bargains in light of these expectations. However, it is held that people expect high rates of inflation in the future precisely because the government's current and prospective monetary and fiscal policies warrant those expectations. . . . An implication of this view is that inflation can be stopped much more quickly than advocates of the "momentum" view have indicated and that their estimates of the length of time and the costs of stopping inflation in terms of forgone output are erroneous. . . . This is not to say that it would be easy to eradicate inflation. On the contrary, it would require more than a few temporary restrictive fiscal and monetary actions. It would require a change in the policy regime. . . . How costly such a move would be in terms of forgone output and how long it would be in taking effect would depend partly on how resolute and evident the government's commitment was.

According to Sargent, the sacrifice ratio could be much smaller than suggested by previous estimates. Indeed, in the most extreme case, it could be zero. The key question was the credibility of the government and central bank. If policymakers were sufficiently credible in their commitment to a policy of low inflation, people would lower their expectations of inflation immediately. The short-run Phillips curve would shift downward, and the economy would reach low inflation quickly without the cost of temporarily high unemployment and low output.

36-4c The Volcker Disinflation

As we have seen, when Volcker faced the prospect of reducing inflation from its peak of about 10 percent, the economics profession offered two conflicting predictions. One group of economists offered estimates of the sacrifice ratio and concluded that

reducing inflation would have great costs in terms of lost output and high unemployment. Another group offered the theory of rational expectations and concluded that reducing inflation could be much less costly and, perhaps, could even have no cost at all. Who was right?

Figure 11 shows inflation and unemployment from 1979 to 1987. As you can see, Volcker did succeed at reducing inflation. It came down from almost 10 percent in 1980 and 1981 to about 4 percent in 1983 and 1984. Credit for this reduction in inflation goes completely to monetary policy. Fiscal policy at this time was acting in the opposite direction: The increases in the budget deficit during the Reagan administration expanded aggregate demand, which tends to raise inflation. The fall in inflation from 1981 to 1984 is attributable to the tough anti-inflation policies of the dogged Fed chair.

The figure shows that the Volcker disinflation came at the cost of high unemployment. In 1982 and 1983, the unemployment rate was about 10 percent—about 4 percentage points above its level when Volcker was appointed Fed chair. The Volcker disinflation produced a recession that was then the deepest the United States had experienced since the Great Depression of the 1930s.

Does this episode refute the rational-expectations theory of costless disinflation? Some economists have argued that the answer to this question is a resounding yes. Indeed, the pattern of disinflation shown in Figure 11 is like the pattern predicted in Figure 10. As the economy transitioned from high inflation (point A in both figures) to low inflation (point C), it experienced a painful period of high unemployment (point B).

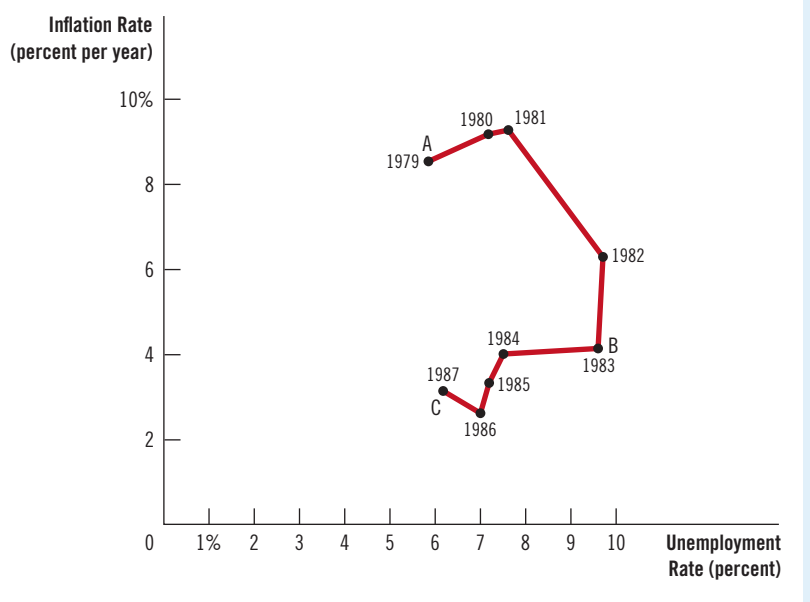
Yet there are two reasons not to reject the conclusions of the rational-expectations theorists so quickly. First, even though the Volcker disinflation did impose a cost of temporarily high unemployment, the cost was not as large as many economists had predicted. Most estimates of the sacrifice ratio based on the Volcker disinflation

Figure 11

The Volcker Disinflation

This figure shows unemployment and inflation from 1979 to 1987. The reduction in inflation (as measured by the GDP deflator) came at the cost of very high unemployment in 1982 and 1983. Note that the points labeled A, B, and C in this figure correspond roughly to the points in Figure 10.

Source: U.S. Department of Labor; U.S. Department of Commerce.



are smaller than estimates that had been obtained from previous data. Perhaps Volcker's tough stand on inflation did have a direct effect on expectations, as the rational-expectations theorists claimed.

Second, and more importantly, even though Volcker announced that he would aim monetary policy to lower inflation, much of the public did not believe him. Because few people thought Volcker would reduce inflation as quickly as he did, expected inflation did not fall immediately; as a result, the short-run Phillips curve did not shift down as quickly as it might have. Some evidence for this hypothesis comes from the forecasts made by commercial forecasting firms: Their forecasts of inflation fell more slowly in the 1980s than did actual inflation. For that reason, the Volcker disinflation does not necessarily refute the rational-expectations view that credible disinflation can be costless. It does show, however, that policymakers cannot count on people to immediately believe them when they announce a policy of disinflation, especially if it marks a major departure from the policy that preceded it. Rational people may doubt policymakers' commitment to low inflation without concrete evidence, which Volcker ultimately provided with his relentless determination.

QuickQuiz

8. Reducing inflation will tend to be costly if
   a. policymakers are credibly committed to low inflation.
   b. wages and prices are not very sticky.
   c. expectations of inflation are slow to adjust.
   d. central bankers exhibit a strong dislike of inflation.

9. Advocates of the theory of rational expectations believe that
   a. the sacrifice ratio can be much smaller if policymakers make a credible commitment to low inflation.
   b. if disinflation catches people by surprise, it will have minimal impact on unemployment.
   c. wage and price setters never expect the central bank to follow through on its announcements.
   d. expected inflation depends on the rates of inflation that people have recently observed.

Answers are at the end of the chapter.

36-5 Recent History

More recent history can be divided into three periods: the Greenspan era, the Great Recession, and the pandemic.

36-5a The Greenspan Era

After the OPEC inflation of the 1970s and the Volcker disinflation of the 1980s, the U.S. economy experienced relatively mild fluctuations in inflation and unemployment. Figure 12 shows inflation and unemployment from 1984 to 2005. This can be called the Greenspan era after Alan Greenspan, who, in 1987, followed Paul Volcker as chair of the Federal Reserve and remained in that position for more than 18 years.

This period began with a favorable supply shock. In 1986, OPEC members argued over production levels, and their agreement to restrict supply broke down. Oil prices fell by about half. As the figure shows, this favorable supply shock led to a fall in inflation and unemployment from 1984 to 1986.

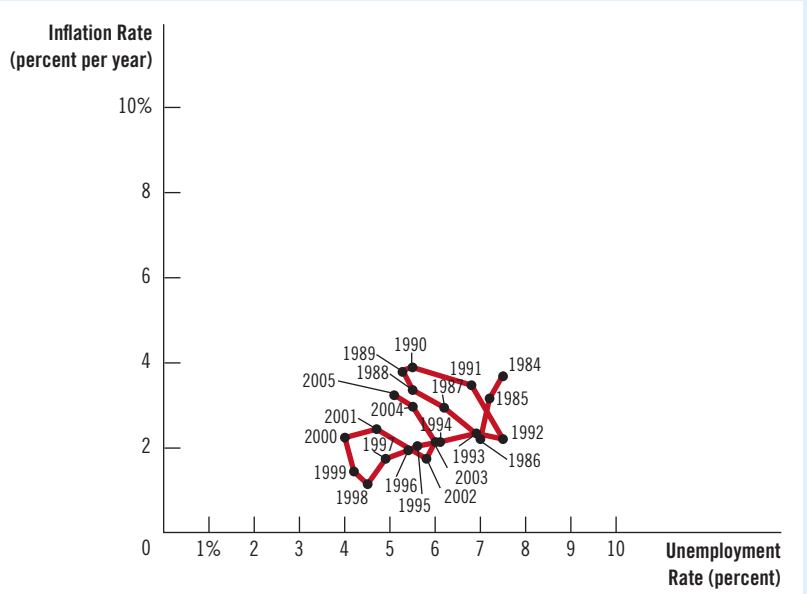
Throughout the Greenspan era, the Fed was careful to avoid repeating the mistakes of the 1960s, when excessive aggregate demand pushed unemployment below

Figure 12

The Greenspan Era

This figure shows annual data from 1984 to 2005 on the unemployment rate and on the inflation rate (as measured by the GDP deflator). During most of this period, Alan Greenspan was chair of the Federal Reserve. Fluctuations in inflation and unemployment were relatively small.

Source: U.S. Department of Labor; U.S. Department of Commerce.



the natural rate and raised inflation. When unemployment fell and inflation rose in 1989 and 1990, the Fed raised interest rates and contracted aggregate demand, leading to a small recession in 1991 and 1992. Unemployment rose above most estimates of the natural rate, and inflation fell once again.

The rest of the 1990s witnessed a technological boom and a period of economic prosperity. Inflation gradually drifted downward, approaching zero by the end of the decade. Unemployment also fell, leading many observers to believe that the natural rate of unemployment had fallen. Part of the credit for this good economic performance went to Greenspan and his colleagues at the Fed, for low inflation can be achieved only with prudent monetary policy. But good luck in the form of favorable supply shocks was also part of the story.

In 2001, however, the economy ran into problems. The end of the dot-com stock market bubble, the 9/11 terrorist attacks, and corporate accounting scandals all depressed aggregate demand. Unemployment rose during the first recession in a decade. But a combination of expansionary monetary and fiscal policies helped end the downturn, and by early 2005, unemployment was close to most estimates of the natural rate.

In 2005, President Bush nominated Ben Bernanke to succeed Alan Greenspan as Fed chair. Bernanke was sworn in on February 1, 2006. In 2009, Bernanke was reappointed by President Obama. At the time of his initial nomination, Bernanke said, "My first priority will be to maintain continuity with the policies and policy strategies established during the Greenspan years."

36-5b The Great Recession

Bernanke may have hoped to continue the policies of the Greenspan era and to enjoy the relative calm of those years, but his wishes would not be fulfilled. During his first few years on the job, the new Fed chair faced some daunting challenges.

As previous chapters discussed, the main problems arose in the housing market and financial system. From 1995 to 2006, the U.S. housing market boomed, and average U.S. house prices more than doubled. But the boom was unsustainable, and from 2006 to 2009, house prices plummeted by about one-third. This large fall led to declines in household wealth and difficulties for many financial institutions that had bet (through the purchase of mortgage-backed securities) that house prices would continue to rise. It set off a financial crisis that resulted in a large decline in aggregate demand and a steep increase in unemployment.

We have previously looked at the story of the crisis and the policy responses to it, but Figure 13 shows what these events meant for inflation and unemployment. From 2007 to 2010, as the decline in aggregate demand raised unemployment from less than 5 percent to about 10 percent, it also reduced the rate of inflation from 3 percent in 2006 to less than 1 percent in 2009, the lowest inflation in more than a half-century. In essence, the economy rode down the short-run Phillips curve.

After 2010, the economy slowly recovered from the Great Recession, and the job of Fed chair passed from Bernanke to Janet Yellen in 2014 and then to Jerome Powell in 2018. Unemployment gradually declined, while the rate of inflation remained between 1 and 2 percent. By 2018 and 2019, the unemployment rate had fallen to less than 4 percent, and inflation had risen slightly to about 2 percent—the Fed's announced target.

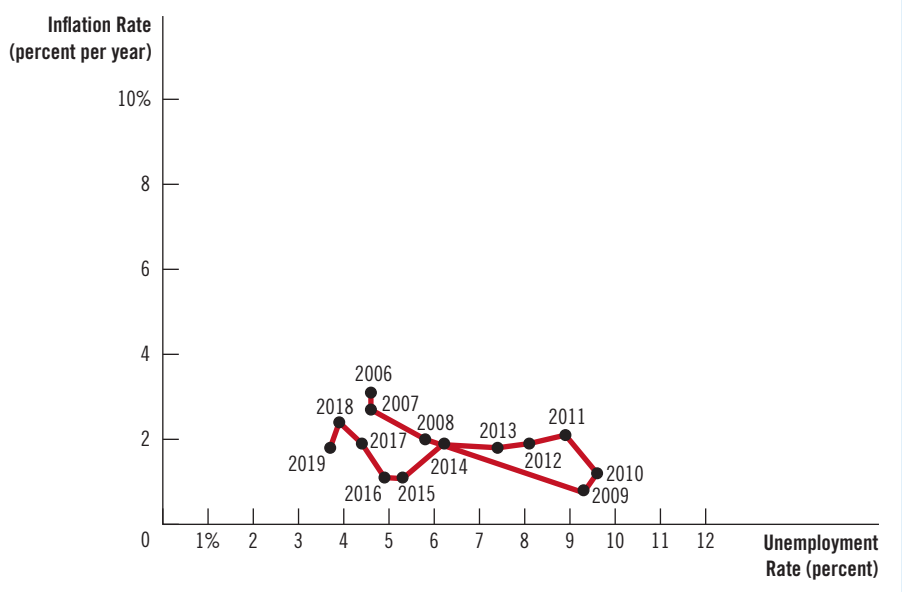
One notable feature of this period was the absence of sizeable shifts in the short-run Phillips curve. The very low inflation of 2009 and 2010 appears not to have substantially reduced expected inflation, which might have shifted the short-run Phillips curve downward. Similarly, the low unemployment during 2018 and 2019 increased inflation somewhat, but it did not increase expected inflation, which could have shifted the short-run Phillips curve upward. Instead, expected inflation remained steady at about 2 percent, keeping the short-run Phillips curve relatively stable.

Figure 13

The Phillips Curve during and after the Recession of 2008–2009

This figure shows annual data from 2006 to 2019 on the unemployment rate and on the inflation rate (as measured by the GDP deflator). A financial crisis caused aggregate demand to plummet, leading to much higher unemployment and pushing inflation down to a very low level.

Source: U.S. Department of Labor; U.S. Department of Commerce.

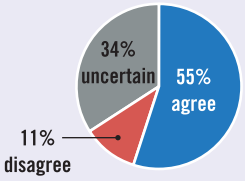


Ask the Experts

The Inflation of 2021–2022

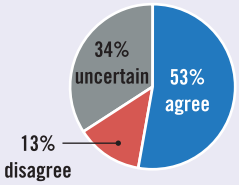
"The supply bottlenecks that are currently contributing to rising prices can be reasonably expected to abate without causing inflation over the longer term to be above the Fed's target."

What do economists say?



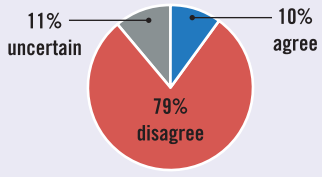
"The current combination of U.S. fiscal and monetary policy poses a serious risk of prolonged higher inflation."

What do economists say?



"A significant factor behind today's higher US inflation is dominant corporations in uncompetitive markets taking advantage of their market power to raise prices in order to increase their profit margins."

What do economists say?



Source: IGM Economic Experts Panel, November 23, 2021, January 11, 2022.

A common explanation for this stability is that since Volcker's success in fighting inflation, the Federal Reserve had established substantial credibility in its commitment to keep inflation at about 2 percent. This credibility kept expected inflation well-anchored. As a result, the position of the short-run Phillips curve reacted less to short-run events.

36-5c The Pandemic

The long expansion after the recession of 2008–2009 was cut short in 2020 when the coronavirus pandemic led to a sharp recession, as discussed earlier in the book. The pandemic affected both aggregate demand and aggregate supply as well as both inflation and unemployment.

The initial impact on aggregate demand was contractionary. People were told to avoid shopping, eating at restaurants, flying for business or personal reasons, and other economic activities that require close personal contact. From February to April of 2020, the unemployment rate jumped from 3.5 to 14.8 percent. The consumer price index fell by 1 percent over those two months, indicating a brief period of deflation.

But aggregate demand recovered quickly. The recovery was partly attributable to the relaxation of restrictions. It was also partly attributable to expansionary monetary and fiscal policies. The Fed cut interest rates to about zero, and Congress passed a series of relief bills under Presidents Trump and Biden.

The pandemic's impact on aggregate supply was consistently adverse. Early in the pandemic, many non-essential businesses were told to close. Even after the restrictions were relaxed, many people were reluctant to return to work. For example, the number of retirements in 2020 was twice that in 2019. The labor-force participation rate for those 55 and older declined by about 2 percentage points early in the pandemic and had not recovered by the end of 2021. At the same time, the pandemic disrupted global supply chains. The inability of some businesses to obtain critical inputs further contracted aggregate supply.

By the end of 2021, the unemployment rate was back below 5 percent, but the combination of the adverse supply shocks and the expansion in aggregate demand from monetary and fiscal policy increased inflation. The 12-month inflation rate, as measured by the consumer price index, rose to 7.5 percent in January 2022, the highest rate in 40 years.

Policymakers initially said that the inflation surge would be transitory. They thought that inflation would subside as supply-chain disruptions were fixed. Some economists, however, worried that the monetary and fiscal expansion had been excessive and that inflation would not soon return to the Fed's target of 2 percent.

The outcome was unclear as this book was going to press. Part of the resolution would come from the behavior of expected inflation. If the Fed maintained the credibility it had earned since Volcker, the high inflation might indeed be short-lived. But if expectations became unanchored, the Phillips curve might shift, leading to a persistently less favorable trade-off between inflation and unemployment.

QuickQuiz

10. The financial crisis of 2008–2009 that led to the Great Recession reduced aggregate _____, which tends to _____ the Phillips curve.
    a. supply; shift
    b. supply; move the economy along
    c. demand; shift
    d. demand; move the economy along

11. The increase in retirements during the pandemic of 2020 reduced aggregate _____, which tends to _____ inflation.
    a. supply; increase
    b. supply; decrease
    c. demand; increase
    d. demand; decrease

Answers are at the end of the chapter.

36-6 Conclusion

This chapter has examined how economists' thinking about inflation and unemployment has evolved. We have discussed the ideas of many of the best economists of the 20th century: from the Phillips curve of Phillips, Samuelson, and Solow, to the natural-rate hypothesis of Friedman and Phelps, to the rational-expectations theory of Lucas, Sargent, and Barro. Six members of this group won Nobel Prizes for their work in economics.

The trade-off between inflation and unemployment has generated much intellectual turmoil, but certain principles command consensus. Here is how Milton Friedman expressed the relationship between inflation and unemployment in 1968:

> There is always a temporary tradeoff between inflation and unemployment; there is no permanent tradeoff. The temporary tradeoff comes not from inflation per se, but from unanticipated inflation, which generally means, from a rising rate of inflation. The widespread belief that there is a permanent tradeoff is a sophisticated version of the confusion between "high" and "rising" that we all recognize in simpler forms. A rising rate of inflation may reduce unemployment, a high rate will not.
>
> But how long, you will say, is "temporary"? . . . I can at most venture a personal judgment, based on some examination of the historical evidence, that the initial effects of a higher and unanticipated rate of inflation last for something like two to five years.

Today, more than a half-century later, this statement still reflects the view of most macroeconomists.

Chapter in a Nutshell

- The Phillips curve describes a negative relationship between inflation and unemployment. By expanding aggregate demand, policymakers can choose a point on the Phillips curve with higher inflation and lower unemployment. By contracting aggregate demand, policymakers can choose a point with lower inflation and higher unemployment.
- The trade-off between inflation and unemployment described by the Phillips curve holds only in the short run. In the long run, expected inflation adjusts to changes in actual inflation, and the short-run Phillips curve shifts. As a result, the long-run Phillips curve is vertical at the natural rate of unemployment.
- The short-run Phillips curve also shifts because of shocks to aggregate supply. An adverse supply shock, such as an increase in world oil prices, gives policymakers a less favorable trade-off between inflation and unemployment. That is, after an adverse supply shock,

policymakers have to accept a higher rate of inflation for any given rate of unemployment or a higher rate of unemployment for any given rate of inflation.

- When the Fed contracts growth in the money supply to reduce inflation, it moves the economy along the short-run Phillips curve, resulting in temporarily high unemployment. The cost of disinflation depends on how quickly expectations of inflation fall. Some economists argue that a credible commitment to low inflation can reduce the cost of disinflation by inducing a quick adjustment of expectations.

Key Concepts

Phillips curve, p. 772
natural-rate hypothesis, p. 780
supply shock, p. 783
sacrifice ratio, p. 786
rational expectations, p. 787

Questions for Review

1. Draw the short-run trade-off between inflation and unemployment. How might the Fed move the economy from one point on this curve to another?
2. Draw the long-run trade-off between inflation and unemployment. Explain how the short-run and long-run trade-offs are related.
3. What is "natural" about the natural rate of unemployment? Why might the natural rate of unemployment differ across countries?
4. Suppose a drought destroys farm crops and drives up the price of food. What is the effect on the short-run trade-off between inflation and unemployment?
5. The Fed decides to reduce inflation. Use the Phillips curve to show the short-run and long-run effects of this policy. How might the short-run costs be reduced?

Problems and Applications

1. Suppose the natural rate of unemployment is 6 percent. On one graph, draw two Phillips curves that describe the four situations listed here. Label the point that shows the position of the economy in each case.
   a. Actual inflation is 5 percent, and expected inflation is 3 percent.
   b. Actual inflation is 3 percent, and expected inflation is 5 percent.
   c. Actual inflation is 5 percent, and expected inflation is 5 percent.
   d. Actual inflation is 3 percent, and expected inflation is 3 percent.
2. Illustrate the effects of the following developments on both the short-run and long-run Phillips curves. Give the economic reasoning underlying your answers.
   a. a rise in the natural rate of unemployment
   b. a decline in the price of imported oil
   c. a rise in government spending
   d. a decline in expected inflation
3. Suppose that a fall in consumer spending causes a recession.
   a. Illustrate the immediate change in the economy using both an aggregate-supply/aggregate-demand diagram and a Phillips-curve diagram. On both graphs, label the initial long-run equilibrium as point A and the resulting short-run equilibrium as point B. What happens to inflation and unemployment in the short run?
   b. Now suppose that over time, expected inflation changes in the same direction that actual inflation changes. What happens to the position of the short-run Phillips curve? After the recession is

over, does the economy face a better or worse set of inflation–unemployment combinations? Explain.

4. Suppose the economy is in a long-run equilibrium.
   a. Draw the economy's short-run and long-run Phillips curves.
   b. Suppose a wave of business pessimism reduces aggregate demand. Show the effect of this shock on your diagram from part (a). If the Fed undertakes expansionary monetary policy, can it return the economy to its original inflation rate and original unemployment rate?
   c. Now suppose the economy is back in long-run equilibrium, and then the price of imported oil rises. Show the effect of this shock with a new diagram like that in part (a). If the Fed undertakes expansionary monetary policy, can it return the economy to its original inflation rate and original unemployment rate? If the Fed undertakes contractionary monetary policy, can it return the economy to its original inflation rate and original unemployment rate? Explain why this situation differs from that in part (b).
5. The inflation rate is 10 percent, and the central bank is considering slowing the rate of money growth to reduce inflation to 5 percent. Economist Milton believes that expectations of inflation change quickly in response to new policies, whereas economist James believes that expectations are very sluggish. Which economist is more likely to favor the proposed change in monetary policy? Why?
6. Suppose the Federal Reserve's policy is to maintain low and stable inflation by keeping unemployment at its natural rate. However, the Fed believes that the natural rate of unemployment is 4 percent when the actual natural rate is 5 percent. If the Fed based its policy decisions on its belief, what would happen to the economy? How might the Fed come to realize that its belief about the natural rate was mistaken?
7. Suppose the Federal Reserve announced that it would pursue contractionary monetary policy to reduce inflation. For each of the following conditions, explain whether it would make the ensuing recession more or less severe.
   a. Wage contracts have short durations.
   b. There is little confidence in the Fed's determination to reduce inflation.
   c. Expectations of inflation adjust quickly to actual inflation.
8. In 2008, the Federal Reserve faced a decrease in aggregate demand caused by the housing and financial crises and a decrease in short-run aggregate supply caused by rising commodity prices.
   a. Starting from a long-run equilibrium, illustrate the effects of these two changes using both an aggregate-supply/aggregate-demand diagram and a Phillips-curve diagram. On both diagrams, label the initial long-run equilibrium as point A and the resulting short-run equilibrium as point B. For each of the following variables, state whether it rises or falls or whether the impact is ambiguous: output, unemployment, the price level, the inflation rate.
   b. Suppose the Fed responds quickly to these shocks and adjusts monetary policy to keep unemployment and output at their natural rates. What action would it take? On the same set of graphs from part (a), show the results. Label the new equilibrium as point C.
   c. Why might the Fed choose not to pursue the course of action described in part (b)?

QuickQuiz Answers

1. **d** 2. **b** 3. **d** 4. **a** 5. **c** 6. **b** 7. **c** 8. **c** 9. **a** 10. **d** 11. **a**

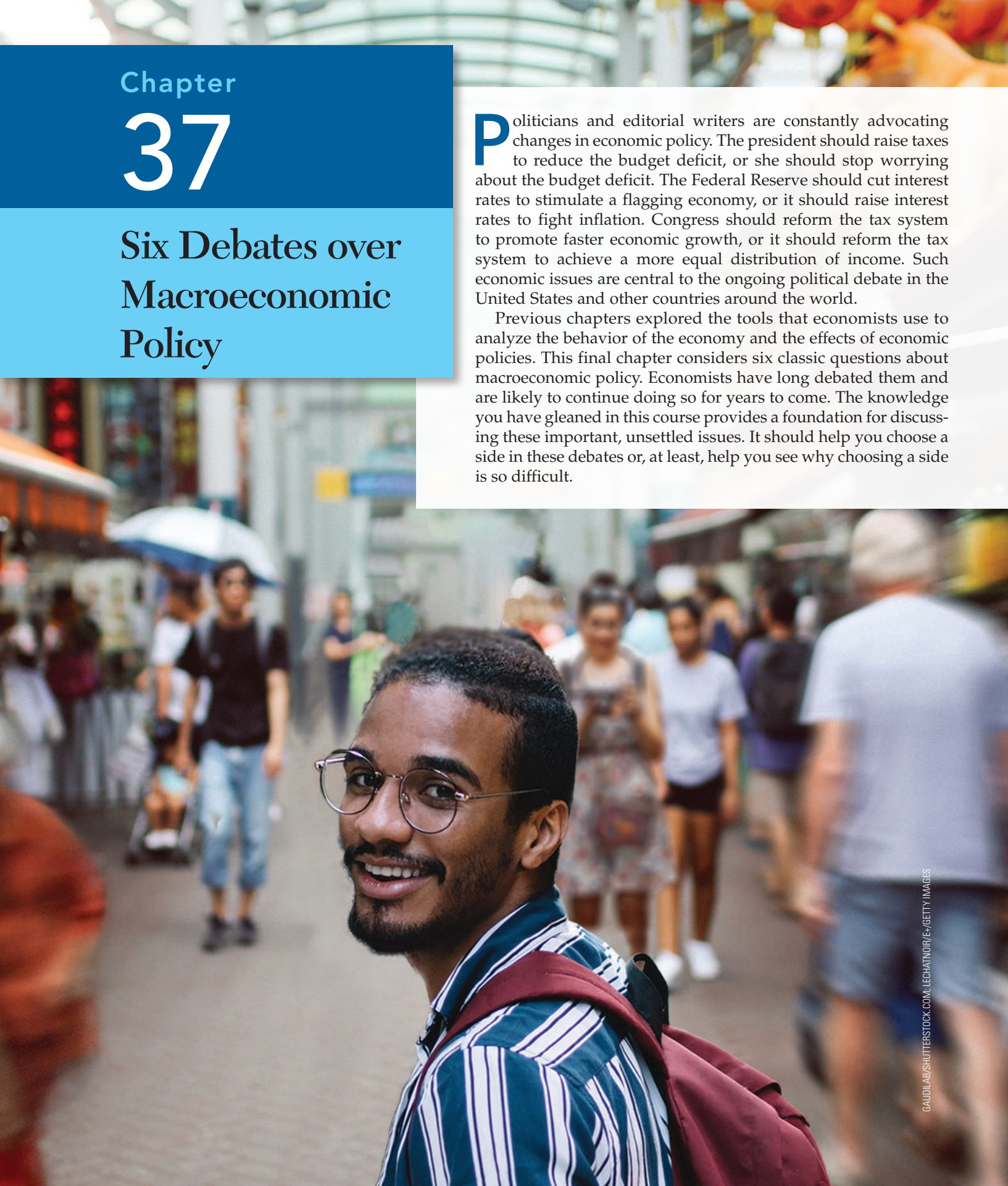
Chapter

37

Six Debates over Macroeconomic Policy

Politicians and editorial writers are constantly advocating changes in economic policy. The president should raise taxes to reduce the budget deficit, or she should stop worrying about the budget deficit. The Federal Reserve should cut interest rates to stimulate a flagging economy, or it should raise interest rates to fight inflation. Congress should reform the tax system to promote faster economic growth, or it should reform the tax system to achieve a more equal distribution of income. Such economic issues are central to the ongoing political debate in the United States and other countries around the world.

Previous chapters explored the tools that economists use to analyze the behavior of the economy and the effects of economic policies. This final chapter considers six classic questions about macroeconomic policy. Economists have long debated them and are likely to continue doing so for years to come. The knowledge you have gleaned in this course provides a foundation for discussing these important, unsettled issues. It should help you choose a side in these debates or, at least, help you see why choosing a side is so difficult.



GAUDILAB/SHUTTERSTOCK.COM; LECHATNOIR/E+/GETTY IMAGES

37-1 How Actively Should Policymakers Try to Stabilize the Economy?

The preceding three chapters discussed how changes in aggregate demand and aggregate supply can lead to short-run fluctuations in production and employment and how monetary and fiscal policy can shift aggregate demand and influence these fluctuations. But even if policymakers **can** influence short-run economic fluctuations, does that mean they **should**? Our first debate concerns how active monetary and fiscal policymakers should be in attempting to smooth the ups and downs of the business cycle.

37-1a The Case for Robust Stabilization Policy

Left on their own, economies fluctuate. When households and firms become pessimistic, for instance, they cut back on spending, thereby reducing the aggregate demand for goods and services. The decline in aggregate demand, in turn, reduces the production of goods and services. Firms lay off workers, and the unemployment rate rises. Real GDP and other measures of income fall. Rising unemployment and falling income help confirm the pessimism that initially generated the economic downturn.

Such a recession has no benefit for society—it represents a sheer waste of resources. Workers who lose their jobs because of declining aggregate demand would rather be working. Business owners whose factories are idle would rather be producing goods and services and selling them at a profit.

There is no good reason for society to suffer through the booms and busts of the business cycle. Macroeconomic theory shows policymakers how to reduce the severity of economic fluctuations. By "leaning against the wind" of economic change, monetary and fiscal policy can stabilize aggregate demand, production, and employment. When aggregate demand is inadequate to ensure full employment, policymakers should boost government spending, cut taxes, expand the money supply, and reduce interest rates. When aggregate demand is excessive, risking higher inflation, policymakers should cut government spending, raise taxes, reduce the money supply, and increase interest rates. Such policy actions put macroeconomic theory to its best use by leading to a more stable economy, which benefits everyone.

37-1b The Case for Modest Stabilization Policy

Monetary and fiscal policy can be used to stabilize the economy in theory, but there are substantial obstacles in practice.

One problem is that monetary and fiscal policies do not affect the economy immediately but instead work with a long lag. Monetary policy affects aggregate demand primarily by changing interest rates, which in turn affect spending, particularly on residential and business investment. But because many households and firms set their most important spending plans far in advance, it takes time for changes in interest rates to alter aggregate demand. Studies indicate that changes in monetary policy have little effect on aggregate demand until about six months after the change is made.

Fiscal policy works with a lag because of the long political process for changing taxes and government spending. To make a substantial change in fiscal policy in

FRANK MODELL/THE NEW YORKER COLLECTION/THE CARTOON BANK

the United States, a bill must go through congressional committees, pass both the House and the Senate, and be signed by the president. It can take years to propose, pass, and implement a major shift in fiscal policy.

Because of these long lags, policymakers who want to stabilize the economy need to look ahead to economic conditions that will prevail when their actions take effect. Unfortunately, economic forecasting is imprecise, in part because macroeconomics is such a primitive science and in part because the shocks that cause economic fluctuations are intrinsically unpredictable. This means that when policymakers change monetary or fiscal policy, they must rely on no more than educated guesses about future conditions.

Too often, policymakers who try to stabilize the economy end up having the opposite effect. Conditions can easily change from the time a policy action begins to the time it takes effect. As a result, policymakers can inadvertently aggravate the business cycle rather than smoothing it. Some economists claim that many major downturns, including the Great Depression of the 1930s, can be traced to destabilizing policy actions.

Early in their training, physicians learn the rule "first, do no harm." The human body has natural restorative powers. Confronted with a sick patient and an uncertain diagnosis, often a doctor should do nothing but leave the patient alone. Intervening without reliable knowledge risks making matters worse.

The same can be said about treating an ailing economy. It might be desirable for policymakers to eliminate all fluctuations, but such a goal is not realistic given the limits of macroeconomic knowledge and the inherent unpredictability of world events. It is reasonable for monetary and fiscal policymakers to intervene when an economic downturn is deep or protracted, but they cannot be expected to prevent every recession. In most cases, they should be content if they do no harm.

QuickQuiz

1. Approximately how long does it take for a change in monetary policy to influence aggregate demand?
   a. one month
   b. six months
   c. two years
   d. five years

2. Fiscal policy has a long lag mainly because
   a. policymakers at the Federal Reserve do not meet frequently.
   b. firms making investments are slow to respond to changes in interest rates.
   c. the political process is slow to enact changes in government spending or taxes.
   d. consumers are slow to respond to changes in their after-tax incomes.

Answers are at the end of the chapter.

37-2 Should the Government Fight Recessions with Spending Hikes or Tax Cuts?

When George W. Bush became president in 2001, the economy was slipping into a recession. He responded by cutting tax rates. When Barack Obama became president in 2009, the economy was in the middle of the Great Recession, the worst economic downturn in many decades. He responded with a stimulus package that offered some tax reductions but also included substantial increases in government spending. The contrast between these two policies illustrates a classic question of macroeconomics: Which instrument of fiscal policy—government spending or taxes—is a better tool for reducing the severity of economic downturns?

37-2a The Case for Fighting Recessions with Spending Hikes

John Maynard Keynes transformed economics when he wrote *The General Theory of Employment, Interest, and Money* during the Great Depression of the 1930s, the worst downturn in U.S. history. Since then, economists have understood that the fundamental problem during recessions is inadequate aggregate demand. When firms can't sell enough goods and services, they reduce production and employment. The key to ending recessions is restoring aggregate demand to a level consistent with full employment.

Monetary policy is often the first line of defense against economic downturns. By increasing the money supply, the central bank reduces interest rates. Lower interest rates reduce the cost of borrowing to finance investment projects, such as new factories and new housing. Increased investment spending adds to aggregate demand and helps to restore normal levels of production and employment.

Fiscal policy provides an additional tool to combat recessions. When the government cuts taxes, it increases households' disposable income, encouraging them to increase spending on consumption. When the government buys goods and services, it adds directly to aggregate demand. Moreover, these fiscal actions can have multiplier effects: Higher aggregate demand leads to higher incomes, higher incomes lead to additional consumer spending, and additional consumer spending leads to further increases in aggregate demand.

Fiscal policy is particularly useful when monetary policy loses its effectiveness. For example, during the Great Recession of 2008 and 2009 and the coronavirus recession of 2020, the Federal Reserve cut its target interest rate to about zero. Though

some central banks have targeted interest rates that are slightly negative, interest rates can't go much below zero because, at some point, people would rather hold onto their cash than lend it out. When the target interest rate hits this effective lower bound, the Fed loses its best tool for stimulating the economy. Unconventional monetary policy, such as quantitative easing and forward guidance, may be useful, but it is arguably less effective. In this circumstance, it is natural for the government to turn to fiscal policy—government spending and taxes—to prop up aggregate demand.

Traditional Keynesian analysis indicates that increases in government purchases are a more potent tool than decreases in taxes. When households get extra disposable income from a tax cut, they will likely save some of that additional income rather than spend it all (especially if households view the tax reduction as temporary rather than permanent). The fraction of the extra income saved does not contribute to the aggregate demand for goods and services. By contrast, when the government spends a dollar buying a good or service, that dollar immediately and fully adds to aggregate demand.

In 2009, economists in the Obama administration used a conventional macroeconomic model to calculate the magnitude of these effects. According to their model simulations, each dollar of tax cuts increases GDP by $0.99, while each dollar of government purchases increases GDP by $1.59. Because increases in government spending offer a bigger "bang for the buck" than decreases in taxes, the policy response in 2009 featured fewer federal tax cuts and more spending.

Policymakers focused on three kinds of spending. First, there was spending on "shovel-ready" projects. These were public works projects such as repairs to highways and bridges on which construction could begin immediately, putting the unemployed back to work. Second, there was federal aid to state and local governments. Because many of these governments are constitutionally required to run balanced budgets, falling tax revenues during recessions can make it necessary for them to lay off teachers, police, and other public workers; federal aid prevented that outcome or, at least, reduced its severity. Third, there were increased payments to the jobless through the unemployment insurance system. Because the unemployed are often financially stretched, they were thought to be likely to spend rather than save this extra income. These transfer payments were thought to contribute more to aggregate demand—and, in turn, to production and employment—than tax cuts would. According to the macroeconomic model used by the Obama administration, the $800 billion stimulus package would create or save more than 3 million jobs by the end of the president's second year in office.

It is impossible to know for sure what effect the stimulus, in fact, had. Because there is only one run at history, no one knows what would have happened without the stimulus package. Yet one thing is clear: While the economic downturn of 2008–2009 was severe, it could have been worse. In the Great Depression of the 1930s, real GDP fell by 27 percent, and unemployment reached 25 percent. In the Great Recession, real GDP fell by only 4 percent, and unemployment reached only 10 percent. As judged by either GDP or unemployment, the Great Recession did not approach the magnitude of the Great Depression.

37-2b The Case for Fighting Recessions with Tax Cuts

There is a long tradition of using tax policy to stimulate a moribund economy. President Kennedy proposed a tax reduction as one of his major economic initiatives; it eventually passed under President Johnson in 1964. President Reagan also signed into law significant tax cuts when he became president in 1981. Both of these tax reductions were soon followed by strong economic growth.

Tax cuts have a powerful influence on both aggregate demand and aggregate supply. They increase aggregate demand by increasing households' disposable income, as emphasized in traditional Keynesian analysis. But they can also increase aggregate demand by altering incentives. For example, if the tax reductions take the form of an expanded investment tax credit, they can induce increased spending on investment goods. Because investment spending is the most volatile component of GDP over the business cycle, stimulating investment is a key to ending recessions. Policymakers can target investment using well-designed tax policy.

While tax cuts increase aggregate demand, they can also increase aggregate supply. When the government reduces marginal tax rates, workers keep a higher fraction of any income they earn. That gives the unemployed a greater incentive to search for jobs and the employed a greater incentive to work longer hours. Increased aggregate supply, along with the increased aggregate demand, means that the production of goods and services can expand without putting upward pressure on inflation.

Increasing government spending during recessions entails various problems. First of all, rational consumers understand that higher government spending, together with the government borrowing needed to finance it, will likely lead to higher future taxes. The anticipation of future taxes induces those consumers to reduce spending today. Moreover, like most taxes, future taxes are likely to cause deadweight losses. As businesses look ahead to a more highly distorted future economy, they may reduce their expectations of future profits and reduce investment spending. Because of these effects, government-spending multipliers may be smaller than is conventionally believed.

It is also far from clear whether the government can spend money both wisely and quickly. Large government spending projects often require years of planning, as policymakers and voters weigh the projects' costs and benefits. But when unemployment soars during recessions, the need for additional aggregate demand is immediate. If the government tries to be careful in planning its expenditures, it may fail to increase aggregate demand in a timely fashion. If, however, the government increases spending quickly, it may end with ill-conceived public projects that create temporary employment but produce little lasting value and leave future taxpayers with a legacy of government debt.

Tax cuts have the advantage of decentralizing spending decisions rather than relying on a centralized and cumbersome political process. Households spend their disposable income on things they value, and firms spend their investment dollars on projects they expect to be profitable. This suggests that tax cuts may be the better weapon to combat economic downturns.

QuickQuiz

3. According to traditional Keynesian analysis, which of the following increases aggregate demand the most?
   a. $100 billion increase in taxes
   b. $100 billion decrease in taxes
   c. $100 billion increase in government purchases
   d. $100 billion decrease in government purchases

4. A cut in income tax rates tends to _____ aggregate demand and _____ aggregate supply.
   a. increase; increase
   b. increase; decrease
   c. decrease; increase
   d. decrease; decrease

Answers are at the end of the chapter.

37-3 Should Monetary Policy Be Made by Rule or Discretion?

As we learned in the chapter on the monetary system, the Federal Open Market Committee sets monetary policy in the United States. The committee meets about every six weeks to evaluate the state of the economy and decide whether to raise, lower, or leave unchanged its target for the federal funds rate. In some circumstances, it may also opt for unconventional monetary actions, such as quantitative easing and forward guidance.

The Federal Open Market Committee operates with substantial discretion over how to conduct monetary policy. The laws that created the Fed gave the institution limited guidance about what goals it should pursue. A 1977 amendment to the 1913 Federal Reserve Act said the Fed "shall maintain long run growth of the monetary and credit aggregates commensurate with the economy's long run potential to increase production, so as to promote effectively the goals of maximum employment, stable prices, and moderate long-term interest rates." But the act does not specify how to weigh these various goals, nor does it tell the Fed how to pursue whatever objective it might choose.

Some economists are critical of this institutional design. Our next debate over macroeconomic policy focuses on whether the Fed should have its discretionary powers reduced and, instead, be committed to following a rule for how it conducts monetary policy.

37-3a The Case for Rule-based Monetary Policy

Discretion in the conduct of monetary policy entails two problems. First, it does not limit incompetence and abuse of power. When the government sends police into a community to maintain civic order, it gives them strict guidelines about how to carry out their job. Because police have great power, allowing them to exercise that power however they wanted would be dangerous. (And, even so, disputes over policing are all too common.) Yet when the government gives central bankers the authority to maintain economic order, it gives them few guidelines. Monetary policymakers are allowed undisciplined discretion.

One example of the abuse of power is that central bankers are sometimes tempted to use monetary policy to affect the outcome of elections. Suppose that the vote for the incumbent president is based on economic conditions at the time of reelection. A central banker sympathetic to the incumbent might be tempted to pursue expansionary policies just before the election to stimulate production and employment, knowing that the resulting inflation will not show up until after the election. When central bankers ally themselves with politicians, discretionary policy can lead to economic fluctuations that reflect the electoral calendar. Economists call such fluctuations the **political business cycle**. Before the election of 1972, for instance, President Richard Nixon pressured Fed Chair Arthur Burns to pursue a more expansionary monetary policy, presumably to bolster Nixon's reelection chances.

The second and subtler problem with discretionary monetary policy is that it might lead to higher inflation than is desirable. Consider the issue from the perspective of Frida, a hypothetical Fed chair. Frida knows that there is no long-run trade-off between inflation and unemployment, and she thinks inflation is costly, so she announces that the Fed will aim for zero inflation. Yet once the public has

formed inflation expectations, Frida faces a short-run trade-off between inflation and unemployment. She is tempted to renege on her promise of price stability to reduce unemployment. This discrepancy between announcements (what policymakers say they will do) and actions (what they actually do) is called the **time inconsistency of policy**. Because policymakers can be time inconsistent, people are skeptical when central bankers say they will reduce inflation. In other words, people may expect higher inflation than central bankers claim they are trying to achieve. Higher expectations of inflation, in turn, shift the short-run Phillips curve upward, making the short-run trade-off between inflation and unemployment less favorable than it otherwise might be.

One way to avoid these two problems with discretionary policy is to commit the central bank to a policy rule. For example, suppose that Congress passed a law requiring the Fed to increase the money supply by exactly 3 percent per year. (Why 3 percent? Because real GDP grows, on average, about 3 percent per year, and because money demand grows with real GDP, 3 percent growth in the money supply is roughly the rate necessary to produce long-run price stability.) Such a law would eliminate incompetence and abuse of power on the part of the Fed, and it would make the political business cycle impossible. In addition, policy could no longer be time inconsistent. People would now believe the Fed's announcement of low inflation because the Fed would be legally required to pursue a low-inflation monetary policy. With low expected inflation, the economy would face a more favorable short-run trade-off between inflation and unemployment.

Other rules for monetary policy are also possible. A more active rule might allow some feedback from the state of the economy to changes in monetary policy. For example, a more active rule might require the Fed to increase monetary growth by 1 percentage point for every percentage point that unemployment rises above its natural rate. Regardless of the precise form of the rule, committing the Fed to some rule would yield advantages by limiting incompetence, abuse of power, and time inconsistency in the conduct of monetary policy.

37-3b The Case for Discretionary Monetary Policy

Discretionary monetary policy may have some drawbacks, but it has an important advantage: flexibility. The Fed confronts ever-changing circumstances, not all of which can be foreseen. In the 1930s, banks failed in record numbers. In the 1970s, the world price of oil skyrocketed. In October 1987, the stock market fell by 22 percent in a single day. From 2007 to 2009, house prices dropped, and home foreclosures soared, sending the financial system into disarray. In 2020, a pandemic closed large segments of the economy. The Fed must decide how to respond to these shocks. A designer of a policy rule could not possibly imagine all the contingencies and specify in advance the right policy response. It is better to appoint good people to conduct monetary policy and give them the freedom to do the best they can.

Moreover, the alleged problems with discretion are largely hypothetical. The practical importance of the political business cycle, for instance, is far from clear. While Nixon did pressure Burns in 1972, it is not clear that he succeeded in altering monetary policy: Interest rates rose during the election year. Moreover, in some cases, just the opposite seems to occur. President Jimmy Carter appointed Paul Volcker to head the Federal Reserve in 1979. Nonetheless, in October of that year, Volcker switched to a contractionary monetary policy to combat the high inflation

that he had inherited from his predecessor. The predictable result of Volcker's decision was a recession, and the predictable result of the recession was a decline in Carter's popularity. Rather than using monetary policy to help the president who had appointed him, Volcker took actions he thought were in the national interest, even though they contributed to Carter's defeat by Ronald Reagan in the November 1980 election.

The practical importance of time inconsistency is also far from clear. Although people are often skeptical of central-bank announcements, central bankers can achieve credibility over time by backing up their words with actions. In the 1990s and 2000s, the Fed achieved and maintained a low rate of inflation, despite the ever-present temptation to take advantage of the short-run trade-off between inflation and unemployment. This experience shows that low inflation does not require that the Fed be committed to a policy rule.

Any attempt to replace discretion with a rule must confront the difficult task of specifying a precise rule. Despite much research examining the costs and benefits of alternative rules, economists have not reached consensus about what a good rule would be. Until there is consensus, society has little choice but to give central bankers discretion to conduct monetary policy as they see fit.

Inflation Targeting

Over the past few decades, many central banks around the world have adopted a policy called **inflation targeting**. Sometimes, a central bank announces its intentions regarding the inflation rate over the next few years. At other times, this policy takes the form of a national law that specifies an inflation goal for the central bank.

Inflation targeting is not a commitment to an ironclad rule. In all the countries that have adopted it, central banks still have a fair amount of discretion. Inflation targets are often set as a range—an inflation rate of 1 to 3 percent, for example—rather than a single number, so the central bank can choose where in the range it wants to be. Moreover, the central bank is sometimes allowed to adjust its target for inflation, at least temporarily, if some event (such as a shock to world oil prices) pushes inflation outside the target range.

Although inflation targeting leaves the central bank with some discretion, the policy does constrain how that discretion is used. When a central bank is told simply to "do the right thing," it is hard to hold the central bank accountable because people can argue forever about what is right. By contrast, when a central bank has an inflation target, it is easier to judge whether the central bank is meeting its goals. Inflation targeting does not tie the bank's hands, but it increases the transparency and accountability of monetary policy. In a sense, inflation targeting is a compromise in the debate over rules versus discretion.

Compared with other central banks around the world, the Federal Reserve was slow to adopt a policy of inflation targeting, although some commentators had long suggested that the Fed had an implicit inflation target of about 2 percent. In January 2012, the Federal Open Market Committee made the policy more explicit. Its press release read as follows:

> The inflation rate over the longer run is primarily determined by monetary policy, and hence the Committee has the ability to specify a longer-run goal for inflation. The Committee judges that inflation at the rate of 2 percent, as measured by the annual change in the price index for personal consumption expenditures, is most consistent over the longer run with the Federal Reserve's statutory mandate. Communicating this inflation goal clearly to the public helps keep longer-term inflation expectations firmly anchored, thereby fostering price stability and moderate long-term interest rates and enhancing the Committee's ability to promote maximum employment in the face of significant economic disturbances.

Inflation targeting in the United States is an evolving policy. In August 2020, the Fed said it sought an "average" inflation rate of 2 percent rather than a firm target of 2 percent. It did not say over what time horizon the average would be calculated. This change built greater leeway into the Fed's policymaking. ■

QuickQuiz

5. Advocates for setting monetary policy by rule rather than discretion often argue that
   a. central bankers with discretion are tempted to renege on their announced commitments to low inflation.
   b. central bankers following a rule will be more responsive to the needs of the political process.
   c. fiscal policy is better than monetary policy as a tool for economic stabilization.
   d. it is sometimes useful to give the economy a burst of surprise inflation.

6. A policy of inflation targeting
   a. removes the need for discretionary decision making by central bankers.
   b. frees central bankers from having to respond to shocks to aggregate demand.
   c. makes central bank policy more transparent and accountable.
   d. has been abandoned by most central banks around the world.

Answers are at the end of the chapter.

37-4 Should the Central Bank Aim for an Inflation Rate Near Zero?

One of the **Ten Principles of Economics** in Chapter 1 is that prices rise when the government prints too much money. Another is that society faces a short-run trade-off between inflation and unemployment. These principles were developed more fully over the course of this book. Together, they raise a question for policymakers: What rate of inflation should a central bank aim for?

37-4a The Case for Near-Zero Inflation

Inflation confers no benefit on society but imposes several costs. Economists have identified six of them:

- Shoeleather costs from reduced money holdings
- Menu costs from more frequent adjustment of prices
- Increased variability of relative prices
- Unintended changes in tax liabilities due to non-indexation of the tax code
- Confusion and inconvenience resulting from a changing unit of account
- Arbitrary redistributions of wealth associated with dollar-denominated debts

Some economists argue that these costs are small, but others claim they can be substantial, even during periods of moderate inflation. Moreover, many people undoubtedly dislike inflation. In 2021, for example, when inflation rose to about 6 percent, opinion polls identified inflation as the nation's leading economic problem.

The benefits of low inflation must be weighed against the costs of achieving it. Reducing inflation often requires a period of high unemployment and low output, as illustrated by the short-run Phillips curve. But this disinflationary recession is only temporary. Once people come to understand that policymakers are aiming for low inflation, expectations of inflation will fall, and the short-run trade-off will improve. Because expectations adjust, there is no trade-off between inflation and unemployment in the long run.

Reducing inflation is, therefore, a policy with temporary costs and permanent benefits. Once the disinflationary recession is over, the benefits of low inflation

persist into the future. If policymakers are farsighted, they should be willing to incur the temporary costs for the permanent benefits. This was precisely the calculation made by Paul Volcker in the early 1980s when he tightened monetary policy and reduced inflation from about 10 percent in 1980 to about 4 percent in 1983. Although unemployment in 1982 reached its highest level since the Great Depression, the economy eventually recovered from the recession, leaving a legacy of low inflation. Today, Volcker is considered a hero among central bankers.

Moreover, the costs of reducing inflation need not be as large as some economists claim. If the Fed announces a credible commitment to low inflation, it can directly influence expectations of inflation. Such a change in expectations can improve the short-run trade-off between inflation and unemployment, allowing the economy to reach lower inflation at a reduced cost. The key to this strategy is credibility: People must believe that the Fed is going to carry through on its announced policy. Congress could help by passing legislation that makes price stability the Fed's primary goal. Such a law would decrease the cost of achieving low inflation without reducing any of the resulting benefits.

37-4b The Case for Living with Moderate Inflation

The benefits of achieving an inflation rate near zero—rather than, say, 4 percent per year—are small, while the costs can be large. Estimates of the sacrifice ratio suggest that reducing inflation by 1 percentage point requires giving up about 5 percent of one year's output. Reducing inflation from 4 percent to zero requires a loss of 20 percent of a year's output. People might dislike inflation of 4 percent, but it is not clear that they would (or should) be willing to pay 20 percent of a year's income to get rid of it.

The social costs of disinflation are even larger than this 20 percent figure suggests because the lost income is not spread equitably throughout the population. When the economy goes into recession, all incomes do not fall proportionately. Instead, the fall in aggregate income is concentrated on those who lose their jobs. The vulnerable workers are often those with the least skills and experience. Hence, much of the cost of reducing inflation is borne by those who can least afford to pay it.

Economists list several costs of inflation, but there is no consensus that these costs are substantial. The shoeleather costs, menu costs, and others that economists have identified do not seem great, at least for moderate rates of inflation. Yes, many people dislike inflation, but they may be misled into believing the inflation fallacy—the view that inflation inevitably erodes living standards. Economists understand that long-run living standards depend on productivity, not monetary policy. Because inflation in nominal incomes typically goes hand in hand with inflation in prices, reducing inflation with tighter monetary policy would not cause real incomes to rise more rapidly.

Moreover, policymakers can lessen many of the costs of inflation without actually reducing inflation. They can eliminate the problems associated with the non-indexed tax system by rewriting the tax laws to account for the effects of inflation. They can reduce the arbitrary redistributions of wealth between creditors and debtors caused by unexpected inflation by issuing indexed government bonds, which the U.S. Treasury has done since the Clinton administration introduced them in 1997. Promoting more widespread use of such bonds would insulate holders of government debt from inflation.

Reducing inflation might be desirable if it could be done at no cost, as some economists argue is possible. Yet this trick is hard to carry out in practice. When nations

In the News

The Goals of Monetary Policy

Two economists argue that the Federal Reserve should limit its focus.

The Fed's Duty Is to the Economy, Not 'Equity'

By Michael T. Belongia and Peter N. Ireland

The Federal Reserve Bank of St. Louis is in the early stages of creating an Institute for Economic Equity "to support an economy in which everyone can benefit regardless of race, ethnicity, gender, or where they live," with an emphasis on "economic outcomes experienced by historically marginalized groups." The Federal Reserve Banks of Boston, Atlanta and Minneapolis have their own initiatives in the works. Sen. Pat Toomey (R., Pa.) has rightly asked regional presidents: Should banks consider social justice and equity in monetary policy decisions?

To put these new initiatives in context, consider the Fed's mission and independence. The Federal Reserve Act directs the Fed to conduct monetary policy in a manner that will "promote effectively the goals of maximum employment, stable prices, and moderate long-term interest rates." Each goal is expressed as an outcome for the economy, rather than the benefit of a particular segment of the economy or prices of individual goods.

Put another way, the tools of monetary policy don't exist to reduce unemployment in the auto or steel industries. The goal isn't to keep gasoline prices low but to maintain a modest, stable rate of increase in the average value of a basket of commodities.

All central banks have confronted this tension between the economy's general performance and the welfare of particular interests. The long recent period of historically low interest rates, for instance, reduces the rate of return earned by savers. At the same time, lower interest rates reduce the cost of purchasing a home, car and other durable goods.

A low value of the dollar benefits U.S. exporters by making their goods less expensive to foreign buyers. But it also harms U.S. consumers by increasing the prices of imported goods. If the Fed considered these trade-offs rather than

reduce inflation, they almost always experience a period of high unemployment and low output. It is risky to believe that the central bank could achieve credibility so quickly that disinflation would be painless.

Indeed, a disinflationary recession can leave permanent scars on the economy. Firms in all industries reduce their spending on new plant and equipment substantially during recessions, making investment the most volatile component of GDP. Even after the recession is over, the smaller capital stock reduces productivity, incomes, and living standards below the levels they otherwise would have achieved. In addition, when workers become unemployed in downturns, they lose job skills, permanently reducing their value as workers.

A little bit of inflation may even be a good thing. Some economists believe that inflation "greases the wheels" of the labor market. Because workers resist cuts in nominal wages, a fall in real wages is more easily accomplished when the price level is rising. Inflation thus makes it easier for real wages to adjust to changes in labor-market conditions.

In addition, inflation allows for the possibility of negative real interest rates. Nominal interest rates cannot fall much below zero because lenders can always hold on to their money rather than lending it out at a negative return. If inflation is zero, real interest rates can also never be negative. But if inflation is, say, 4 percent, then a cut in nominal interest rates can produce real interest rates of negative 4 percent. Sometimes, the economy may need deeply negative real interest rates to provide sufficient stimulus to aggregate demand—an option ruled out by zero inflation.

In light of all these arguments, why should policymakers put the economy through a costly and inequitable disinflationary recession to achieve zero inflation?

general economic performance, would it choose to ease or tighten monetary policy knowing full well that any action will benefit some and harm others? Regrettable or not, the conduct of monetary policy inevitably will have distributional consequences that are secondary to pursuing the best outcome for the whole.

This tension between the general and the particular helps explain the argument that central banks should be "independent" institutions so their policies are insulated from political pressure. But independence doesn't mean a central bank is free to set its own goals, which are usually established by a legislative body.

The Fed isn't free to choose its own monetary-policy objectives, but it has wide latitude and independence to choose a strategy to achieve those goals. While the Fed might choose to set a target for an interest rate or growth of the money supply, these are merely functional ways to pursue the goals set by Congress. Lawmakers can then hold the Fed accountable for its performance in achieving its mandate.

But how can the Fed pursue new goals for "equity" while staying committed to its legislated mandate to achieve "maximum employment, stable prices, and moderate long-term interest rates?"

In pursuit of equity, does the Fed plan to become more involved in the allocation of credit? What tools does the Fed plan to use? Will the new goal of equity for "historically marginalized groups" require changes in the distributional effects monetary policy has on savers, borrowers, exporters and consumers? Perhaps most fundamental, how will the Fed define "equity?"

The Fed's new initiatives expose the central bank to political pressures that are inconsistent with an independent institution. If the Fed wants to lose its independence, it is taking steps to assure that outcome. ■

Questions to Discuss

1. Do you agree with the authors that the Fed should not concern itself with issues of social equity? Why or why not?
2. If the Fed follows the authors' advice, what other institutions should focus on social equity? Why might these institutions be better equipped for the task?

Mr. Belongia is a professor of economics at the University of Mississippi. Mr. Ireland is a professor of economics at Boston College.

Source: *The Wall Street Journal*, June 10, 2021.

The economist Alan Blinder, who was once vice chairman of the Fed, argued in his book *Hard Heads, Soft Hearts* that policymakers should not make this choice:

> The costs that attend the low and moderate inflation rates experienced in the United States and in other industrial countries appear to be quite modest—more like a bad cold than a cancer on society. . . . As rational individuals, we do not volunteer for a lobotomy to cure a head cold. Yet, as a collectivity, we routinely prescribe the economic equivalent of lobotomy (high unemployment) as a cure for the inflationary cold.

Blinder concludes that it is better to learn to live with moderate inflation.

QuickQuiz

7. Which of the following is NOT an argument for a zero rate of inflation?
   a. It eliminates distortions from a non-indexed tax code.
   b. It encourages people to hold a greater quantity of money.
   c. It reduces the menu costs that firms have to incur.
   d. It stops real wages from falling if nominal wages cannot be cut.

8. Which of the following is NOT an argument for a positive rate of inflation?
   a. It permits real interest rates to be negative.
   b. It increases the variability of relative prices.
   c. It allows real wages to fall without cuts in nominal wages.
   d. It would be costly to reduce inflation to zero.

Answers are at the end of the chapter.

37-5 Should the Government Balance Its Budget?

Government debt is the subject of a persistent macroeconomic debate. Whenever the government spends more than it collects in tax revenue, it finances this budget deficit by issuing government debt. Our study of financial markets showed how budget deficits affect saving, investment, and interest rates. But how big a problem are budget deficits? The next debate concerns whether fiscal policymakers should make balancing the government's budget a high priority.

37-5a The Case for a Balanced Budget

The U.S. federal government is far more indebted today than it was four decades ago. In 1980, the federal debt was $712 billion; in 2021, it was $22.4 trillion. If we divide today's debt by the size of the population, we learn that each person's share of the government debt is about $68,000.

The government debt places a burden on future taxpayers. When these debts and accumulated interest come due, these people will face difficult choices. They can embrace a combination of higher taxes and less government spending to make resources available to pay off the debt and accumulated interest. Or they can delay the day of reckoning and borrow once again to pay off the old debt and interest, making the debt burden even larger. In essence, when the government runs a budget deficit and issues debt, it allows current taxpayers to pass the bill for some of their government spending on to future taxpayers. Inheriting such a large debt will lower the living standard of future generations.

In addition, budget deficits have macroeconomic effects. Because these deficits represent **negative** public saving, they lower national saving (the sum of private and public saving), causing real interest rates to rise and investment to fall. Reduced investment leads to a smaller stock of capital, which reduces labor productivity, real wages, and the production of goods and services. In short, when the government increases its debt, future generations are born into an economy with lower incomes as well as higher taxes.

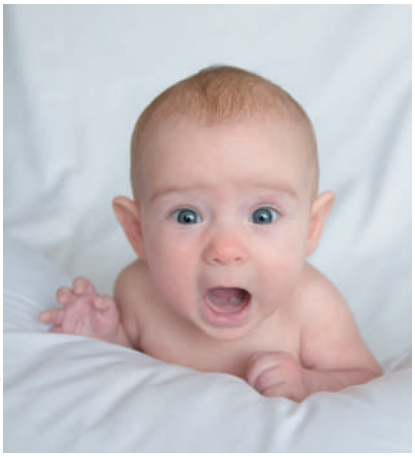
There are, nevertheless, situations in which running a budget deficit is justifiable. Historically, war is the most common cause of large increases in government debt. Financing short-term military spending by borrowing makes sense because otherwise, taxes would rise precipitously in wartime. Such high tax rates would lead to large deadweight losses. They would also be unfair to current citizens who are making the sacrifice of fighting the war to ensure security and freedom not only for themselves but also for future generations.

Similarly, it is reasonable to allow a budget deficit during economic downturns. Tax revenue automatically declines during recessions because payroll and income taxes are levied on measures of income. And government spending automatically rises as more people become eligible for programs such as unemployment insurance. If the government tried to balance its budget during recessions, it would have to raise taxes or cut spending when unemployment is high. Such a policy would tend to depress aggregate demand when it needs to be stimulated and, therefore, could make downturns more severe.

Yet not all budget deficits can be justified because of war or recession. In 2021, the Congressional Budget Office (CBO) projected that, if current policies are maintained, U.S. government debt as a percentage of GDP would increase from 102 percent in 2021 to 195 percent in 2040. Moreover, the CBO made the optimistic assumption the nation will experience neither a major military conflict nor a major economic downturn during this period. The government was projected to run sizable

budget deficits simply because presidents and Congresses had committed the federal government to a variety of spending programs without passing the taxes necessary to fund them.

This projected policy is unsustainable. Eventually, the government will need to enact measures to bring spending in line with tax revenue. The open question is whether the fiscal adjustment should take the form of reduced spending, increased taxes, or a combination of the two. Compared with the alternative of ongoing budget deficits, a balanced budget means greater national saving, increased capital accumulation, and faster economic growth. It means that people in the future will live in a more prosperous economy.



DMISFOTOS/SHUTTERSTOCK.COM

"What?!? My share of the government debt is $68,000?"

37-5b The Case against a Balanced Budget

The problem of government debt is often exaggerated. Although it represents a tax burden on younger generations, that burden is not large compared with the average person's lifetime income. The debt of the U.S. federal government is about $68,000 per person. Someone who works 40 years for $50,000 a year will earn $2 million. In other words, the government debt represents only 3.4 percent of a typical person's lifetime resources.

It is also misleading to consider a budget deficit in isolation. The deficit is just one piece of a larger picture of how the government raises and spends money. When setting fiscal policy, policymakers affect different generations of taxpayers in many ways. The government's budget deficit or surplus should be evaluated along with these policies.

For example, suppose the government reduces the budget deficit by cutting spending on public investments, such as education. Does this policy make younger generations better off? The government debt will be smaller, reducing their tax burden. Yet if they are less educated than they otherwise would be, their productivity and incomes will be lower. Many studies find that the return to schooling (the wage increase that results from an additional year in school) is large. Reducing the budget deficit by cutting education spending could, all things considered, hurt future generations.

Single-minded concern about the budget deficit is dangerous because it draws attention away from other policies that redistribute income across generations. For example, in the 1960s and 1970s, the U.S. federal government raised Social Security benefits for the elderly. It financed this higher spending by increasing the payroll tax on the working-age population. This policy redistributed income away from younger generations toward older generations, even though it did not affect the government debt. Whether this was good or bad is the subject of another debate. What's relevant here is that the budget deficit is only a small part of the larger issue of how government policy affects the welfare of different generations.

To some extent, forward-looking parents can reverse the adverse effects of government debt. Suppose the government cuts taxes and runs a budget deficit. Parents can offset the impact of the deficit simply by saving their tax cut and investing in their children, perhaps by leaving them a bequest. Doing so would enhance the children's ability to bear the burden of the future taxes implied by the greater government debt. If everyone behaved this way, higher private saving by parents would offset the public dissaving of budget deficits, and deficits would not affect the economy. Most economists doubt that parents are so farsighted, but some people may act this way, and anyone could. The point is that deficits give people the opportunity to consume at their children's expense but don't require parents to do so. If the government debt were actually a great problem facing future generations, some parents would help to solve it.

Critics of budget deficits sometimes assert that the government debt cannot continue to rise forever, but in fact, it can. Just as a bank evaluating a loan application would compare a person's debts to his income, we should judge the burden of the government debt relative to the size of the nation's income. Because of population growth and technological progress, the total income of the U.S. economy grows over time. As a result, the nation's ability to pay the interest on the government debt grows as well. As long as the government debt grows more slowly than the nation's income, there is nothing to prevent the government debt from growing forever.

Some numbers can put this into perspective. The CBO projects that the output of the U.S. economy will grow by about 2 percent per year. If inflation also averages 2 percent per year, as the Fed is targeting, nominal income will grow at about 4 percent per year. The government debt can, therefore, rise by 4 percent per year without increasing the ratio of debt to income. In 2021, the federal government debt was $22.4 trillion; 4 percent of this figure is about $900 billion. As long as the federal budget deficit is smaller than $900 billion, the policy is sustainable.

To be sure, large budget deficits cannot persist forever. The $2.7 trillion budget deficits experienced in 2020 and 2021 during the coronavirus pandemic would not be sustainable over the long term. But zero is the wrong target for fiscal policymakers. As long as the deficit is moderate in size, there will never be a day of reckoning that forces government borrowing to end or the economy to collapse.

QuickQuiz

9. Throughout U.S. history, what has been the most common cause of substantial increases in government debt?
   a. pandemics
   b. wars
   c. infrastructure spending
   d. tax cuts

10. Other things being equal, when the government runs a large budget deficit, it _____ national saving and thereby _____ capital formation and productivity growth.
    a. increases; increases
    b. increases; decreases
    c. decreases; increases
    d. decreases; decreases

Answers are at the end of the chapter.

37-6 Should the Tax Laws Be Reformed to Encourage Saving?

A nation's standard of living depends on its ability to produce goods and services. This was one of the **Ten Principles of Economics** in Chapter 1. As we saw in the chapter on production and growth, a nation's productive capability is determined partly by how much it saves and invests. Our last debate is whether policymakers should reform the tax laws to encourage saving.

37-6a The Case for Promoting Saving through Tax Reform

A nation's saving rate is a key determinant of its long-run prosperity. When the saving rate is higher, more resources are available for investment in new plant and equipment, which in turn raises the capital stock, labor productivity, wages, and incomes. Another of the **Ten Principles of Economics** in Chapter 1 is that people respond to incentives. This lesson applies to people's decisions about how much to save. If a nation's laws make saving attractive, people will save more of their incomes, and this higher saving will lead to a more prosperous future.

Unfortunately, the U.S. tax system discourages saving by heavily taxing the return to saving. For example, consider Willa Worker, a 25-year-old who saves $1,000 of her income to have a more comfortable retirement at the age of 70. If she buys a bond that pays an interest rate of 6 percent, the $1,000 will accumulate at the end of 45 years to $13,800 in the absence of taxes on interest. But suppose Willa faces a marginal tax rate on interest income of 40 percent, which is typical for many workers once federal and state income taxes are added together. In this case, her after-tax interest rate is only 3.6 percent, and the $1,000 will accumulate at the end of 45 years to only $4,900. That is, accumulated over this long span of time, the tax rate on interest income reduces the benefit of saving $1,000 from $13,800 to $4,900—or by about 64 percent.

The tax code further discourages saving by taxing some forms of capital income twice. Suppose Willa uses some of her saving to buy stock in a corporation. When the corporation earns a profit from its capital investments, it first pay tax on this profit in the form of the corporate income tax. If the corporation pays out the rest of the profit to Willa and other stockholders in the form of dividends, they pay tax on this income a second time in the form of the individual income tax. This double taxation substantially reduces the return to stockholders, reducing the incentive to save.

The tax laws again discourage saving if Willa wants to leave her accumulated wealth to her children or other relatives rather than consuming it during her lifetime. People can bequeath some money tax-free, but for large bequests, the estate tax rate (federal and state combined) can be as high as 50 percent. Concern about national saving is mainly motivated by a desire to ensure prosperity for future generations. It is odd, therefore, that the tax laws impede the most direct way in which one generation can help the next.

In addition to the tax code, many other policies and institutions reduce the incentive for Willa to save. Some government benefits, such as welfare and Medicaid, are means-tested: They are reduced for those who have saved in the past and accumulated some wealth. Similarly, colleges and universities grant financial aid based on the wealth of students and their parents. Such a policy, like a tax on wealth, deters saving.

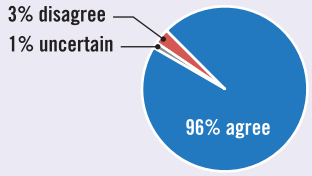
There are many ways in which tax reform could encourage saving, or at least reduce the disincentive that Willa now faces. Already, the tax laws give preferential treatment to some types of saving. When a taxpayer puts income into an Individual Retirement Account (IRA), for instance, that income and the interest it earns are not taxed until the funds are withdrawn at retirement. A similar tax advantage is given to retirement accounts with other names, such as 401(k), 403(b), and profit-sharing plans. Tax-favored accounts are also available for saving earmarked for college tuition. There are, however, limits on who is eligible for these plans and on how much an eligible person can put into them. Moreover, these plans provide little incentive for other types of saving, such as saving to build an emergency fund, buy a house or car, or pay for a vacation or home improvement. A small step toward encouraging saving would be to expand the ability of households to use tax-advantaged accounts for a broader range of goals.

A more comprehensive approach would be to reconsider the entire basis by which the government collects revenue. The centerpiece of the U.S. tax system is the income tax. A dollar earned is taxed at the same rate whether it is spent or saved. An alternative endorsed by many economists is a consumption tax, under which a household pays taxes based on what it spends. Income that is saved is exempt from taxation until the saving is later withdrawn and spent on consumption goods. In essence, a consumption tax automatically puts all saving into a tax-advantaged savings account, much like an IRA. A switch from income taxation to consumption taxation would increase Willa's incentive to save.

Ask the Experts: Taxing Capital and Labor

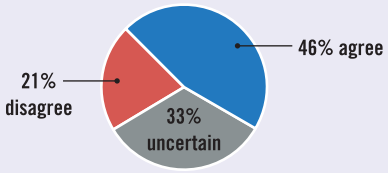
"One drawback of taxing capital income at a lower rate than labor income is that it gives people incentives to relabel income that policymakers find hard to categorize as 'capital' rather than 'labor'."

What do economists say?



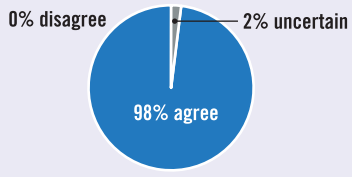
"Despite relabeling concerns, taxing capital income at a permanently lower rate than labor income would result in higher average long-term prosperity, relative to an alternative that generated the same amount of tax revenue by permanently taxing capital and labor income at equal rates instead."

What do economists say?



"Although they do not always agree about the precise likely effects of different tax policies, another reason why economists often give disparate advice on tax policy is because they hold differing views about choices between raising average prosperity and redistributing income."

What do economists say?



Source: IGM Economic Experts Panel, October 9, 2012.

37-6b The Case against Promoting Saving through Tax Reform

Greater saving may be desirable, but it is not the only goal of tax policy. Policymakers also must be sure to distribute the tax burden fairly. The problem with proposals to increase the incentive to save is that they increase the tax burden on those who can least afford it.

It is undeniable that households with high incomes save a larger fraction of their income than households with low incomes. As a result, any tax change that favors people who save will tend to favor people with high incomes. Policies such as tax-advantaged retirement accounts may seem appealing, but they lead to a less egalitarian society. By reducing the tax burden on the wealthy who can take advantage of these accounts, they force the government to raise the tax burden on families with lower incomes.

Moreover, tax policies designed to encourage saving may not be effective at achieving that goal. Economic theory does not give a clear prediction about whether a higher rate of return would increase saving. The outcome depends on the relative size of two conflicting forces called the **substitution effect** and the **income effect**. On the one hand, a higher rate of return raises the benefit of saving: Each dollar saved today produces more consumption in the future. This substitution effect tends to increase saving. On the other hand, a higher rate of return lowers the need for saving: A household has to save less to achieve any target level of consumption in the future. This income effect tends to reduce saving. If the substitution and income effects approximately cancel each other, as some studies suggest, then saving will not change when lower taxation of capital income raises the rate of return.

There are ways to increase national saving other than by giving tax breaks to the rich. National saving is the sum of private and public saving. Instead of trying to alter the tax code to encourage greater private saving, policymakers can simply raise public saving by reducing the budget deficit, perhaps by raising taxes on the wealthy. This approach offers a direct way of raising national saving and increasing prosperity for future generations.

Indeed, once public saving is taken into account, tax provisions to encourage saving might backfire. Tax changes that reduce the taxation of capital income reduce government revenue and thereby lead to a larger budget deficit. To increase national saving, such changes in the tax code must increase private saving by more than they decrease public saving. If they fail to do so, so-called saving incentives can potentially make matters worse.

QuickQuiz

11. Advocates of taxing consumption rather than income argue that
    a. the current tax code discourages people from saving.
    b. the rich consume a higher fraction of income than the middle class.
    c. a consumption tax is a better automatic stabilizer.
    d. taxing consumption does not cause any deadweight losses.

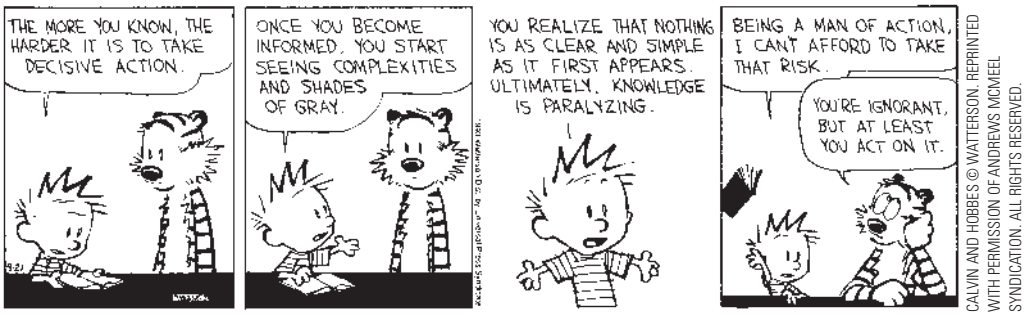
12. Critics of taxing consumption rather than income argue that
    a. switching to a consumption tax would benefit the rich.
    b. private saving does not respond much to tax incentives.
    c. reducing the budget deficit is a better way to raise national saving.
    d. All of the above.

Answers are at the end of the chapter.

37-7 Conclusion: Economic Policy and Shades of Gray

This chapter discussed six classic debates over macroeconomic policy. For each, it began with a policy question and then offered the arguments on both sides. If you find it hard to pick a side, you may find some comfort in the fact that you are not alone. The study of economics does not always make it easy to choose among alternative policies. Indeed, by clarifying the trade-offs that policymakers face, it can make choices more difficult.

Calvin, the boy in the following cartoon strip, could well have been reading this textbook:



Hobbes, the tiger, was wise enough to distrust Calvin's unschooled confidence.

Difficult choices have no right to seem easy. When you hear politicians and other men and women of action proposing something that sounds too good to be true, it probably is. If they appear to be offering you a free lunch, look for the hidden price tag. Few policies come with benefits and no costs. By helping you see through the fog of rhetoric so common in political discourse, the study of economics should make you a better participant in our national debates.

Chapter in a Nutshell

- Advocates of active monetary and fiscal policy view the economy as inherently unstable and believe that policy can manage aggregate demand to offset the instability. Critics of active monetary and fiscal policy emphasize that policy affects the economy with a lag and that our ability to forecast future economic conditions is poor. As a result, attempts to stabilize the economy can end up being destabilizing.
- Advocates of increased government spending to fight recessions argue that because the extra income from tax cuts may be saved rather than spent, direct government spending provides a greater boost to aggregate demand, which is key to promoting production and employment. Critics of spending hikes argue that tax cuts can expand both aggregate demand and aggregate supply and that hasty increases in government spending may lead to wasteful public projects.
- Advocates of rules for monetary policy argue that discretionary policy can suffer from incompetence, the abuse of power, and time inconsistency. Critics of rules for monetary policy argue that discretionary policy is more flexible in responding to changing economic circumstances.
- Advocates of a zero-inflation target emphasize that inflation has many costs and few benefits. Moreover, the cost of eliminating inflation—depressed output and increased unemployment—is only temporary. Even this cost can be reduced if the central bank credibly commits to reduce inflation, thereby directly lowering inflation expectations. Critics of a zero-inflation target claim that moderate inflation imposes only small costs on society and that the recession necessary to reduce inflation to zero is quite costly. The critics also point out two benefits to moderate inflation: It facilitates real-wage adjustment, and it allows real interest rates to be negative when necessary.
- Advocates of a balanced government budget argue that budget deficits impose an unjustifiable burden on future generations by raising their taxes and lowering their incomes. Critics of a balanced government budget argue that the deficit is only one small piece of fiscal policy. Single-minded concern about the budget deficit can obscure the many ways in which policy, including various spending programs, affects different generations.
- Advocates of tax incentives for saving point out that society discourages saving in many ways, such as by heavily taxing capital income and reducing benefits for those who have accumulated wealth. They endorse reforming the tax laws to encourage saving, perhaps by switching from an income tax to a consumption tax. Critics of tax incentives for saving argue that many proposed changes to stimulate saving would primarily benefit the wealthy, who do not need a tax break. They also argue that such changes might have only a small effect on private saving. Raising public saving by reducing the government's budget deficit would provide a more direct and equitable way to increase national saving.

Questions for Review

1. What causes the lags in the effect of monetary and fiscal policy on aggregate demand? What are the implications of these lags for the debate over active versus passive policy?
2. According to traditional Keynesian analysis, why does a tax cut have a smaller effect on GDP than a similarly sized increase in government spending? Why might the opposite be the case?
3. What might motivate a central banker to cause a political business cycle? What does the political business cycle imply for the debate over policy rules?
4. Explain how credibility might affect the cost of reducing inflation.
5. Why are some economists against a target of zero inflation?
6. Explain two ways in which a government budget deficit hurts a future worker.
7. What are two situations in which most economists view a budget deficit as justifiable?
8. Some economists say that the government can continue running a budget deficit forever. How is that possible?
9. Some income from capital is taxed twice. Explain.
10. What adverse effect might be caused by tax incentives to increase saving?

Problems and Applications

1. The chapter suggests that the economy, like the human body, has "natural restorative powers."
   a. Illustrate the short-run effect of a fall in aggregate demand using an aggregate-demand/aggregate-supply diagram. What happens to total output, income, and employment?
   b. If the government does not use stabilization policy, what happens to the economy over time? Illustrate this adjustment on your diagram. Does it generally occur in a matter of months or a matter of years?
   c. Do you think the "natural restorative powers" of the economy mean that policymakers should be passive in response to the business cycle?
2. Policymakers who want to stabilize the economy must decide how much to change the money supply, government spending, or taxes. Why is it difficult for policymakers to choose the appropriate strength of their actions?
3. The problem of time inconsistency applies to fiscal policy as well as to monetary policy. Suppose the government announced a reduction in taxes on income from capital investments, like new factories.
   a. If investors believed that capital taxes would remain low, how would the government's action affect the level of investment?
   b. After investors have responded to the announced tax reduction, does the government have an incentive to renege on its policy? Explain.
   c. Given your answer to part (b), would investors believe the government's announcement? What can the government do to increase the credibility of announced policy changes?
   d. Explain why this situation is similar to the time-inconsistency problem faced by monetary policymakers.
4. Chapter 2 explains the difference between positive analysis and normative analysis. In the debate about whether the central bank should aim for zero inflation, which areas of disagreement involve positive statements, and which involve normative judgments?
5. Why are the benefits of reducing inflation permanent and the costs temporary? Why are the costs of increasing inflation permanent and the benefits temporary? Use Phillips-curve diagrams in your answer.
6. Suppose the federal government cuts taxes and increases spending, raising the budget deficit to 12 percent of GDP. If nominal GDP is rising 5 percent per year, are such budget deficits sustainable forever? Explain. If budget deficits of this size are maintained for 20 years, what is likely to happen to your taxes and your children's taxes in the future? Can you personally do something today to offset this future effect?
7. Explain how each of the following policies redistributes income across generations. Is the redistribution from young to old or from old to young?
   a. an increase in the budget deficit
   b. more generous subsidies for education loans
   c. greater investments in highways and bridges
   d. an increase in Social Security benefits
8. What is the fundamental trade-off that society faces if it chooses to save more? How might the government increase national saving?

QuickQuiz Answers

1. **b** 2. **c** 3. **c** 4. **a** 5. **a** 6. **c** 7. **d** 8. **b** 9. **b** 10. **d** 11. **a** 12. **d**

Chapter

38

Appendix: How Economists Use Data

"Data! Data! Data!" Sherlock Holmes once cried. "I can't make bricks without clay." As usual, the fictional detective was right: To solve a mystery, or to understand any other aspect of the world, we need data. Theories and principles are critically important, but only after observing what's happening around us can we be sure we know what is true and what is not.

The term **data** refers to factual information that provides the basis for reasoning and discussion. In economics, data are often quantitative, such as a person's income, a firm's profit, the market price of ice cream, the amount of ice cream sold, or a nation's gross domestic product. Data allow us to attach real numbers to the conceptual variables found in economic theory.

Data analysis is increasingly central to modern economics. Over the past half century, advances in computing power have enabled economists to analyze ever larger data sets, and research based on data has become increasingly important. Compared with economists of the past, modern economists base their beliefs and policy advice less on pure theory and casual observation and more on the hardheaded analysis of data.

data
factual information, often quantitative, that provides the basis for reasoning and discussion

BLACKCAT/E+/GETTY IMAGES; GOWITHSTOCK/SHUTTERSTOCK.COM

econometrics
the subfield of economics that develops tools to analyze data

A subfield of economics, called **econometrics**, is devoted to developing tools for data analysis. In essence, econometrics is the study of the statistical methods that are useful for understanding the economy. Many colleges offer courses in econometrics, and students majoring in economics are often required to take one as part of their training. This appendix chapter offers a brief introduction.

We address three issues. First, we consider the kinds of data that economists use. Second, we discuss what economists aim to achieve through data analysis. Third, we examine some of the challenges that arise when drawing inferences from data and the methods that econometricians have devised to meet those challenges.

38-1 The Data That Economists Gather and Study

Let's begin by discussing the sources and types of data that economists use most often.

38-1a Experimental Data

randomized controlled trial
an experiment in which a researcher randomly divides subjects into groups, treats the groups differently, and compares their outcomes

Sometimes data come from randomized controlled trials. A **randomized controlled trial** is an experiment in which a researcher randomly divides subjects into groups, treats the groups differently, and compares how the groups respond to their treatments.

For example, suppose a pharmaceutical company comes up with a new drug to treat a disease. Before regulators allow the company to market the drug, it must prove that the drug is safe and effective. The company's researchers start by recruiting a sample of, say, 200 people who have the disease. Half of the patients are randomly assigned to the **treatment group** and given the drug. The other half are assigned to the **control group** and given a placebo (a harmless but ineffective pill that looks like the actual drug). The researchers then follow the health of the two groups. If the patients in the treatment group fare better than those in the control group, the drug is deemed safe and effective. Otherwise, the drug is declared unsafe, ineffective, or both.

experimental data
data that come from a researcher running a randomized controlled trial

Data gathered from randomized controlled trials are called **experimental data**. In many cases, controlled trials are the most reliable way to draw inferences about things we want to learn. If the number of trial participants is large enough and the assignment to the treatment and control groups is truly random, we can be sure that the only important difference between the two groups is their exposure to the treatment.

Randomized controlled trials are sometimes used in the social sciences. (The case study below presents an example.) But their usefulness is limited in economics. The problem is often one of feasibility. Experiments can be expensive to run, and policymakers may object to the unfairness of treating people differently. And sometimes, the economic cost of running the experiment would be too large. For example, to study the effects of monetary policy, a central bank could set its policy randomly from year to year and then observe the consequences. This experiment might advance the cause of social science, but it would have such an adverse impact on a nation's welfare that no one would seriously consider conducting it.

Case Study

The Moving to Opportunity Program

An important example of experimental data comes from the Moving to Opportunity Program, which the U.S. Department of Housing and Urban Development ran in the 1990s. The goal was to study the effects of living in high-poverty neighborhoods.

The researchers recruited several thousand low-income families living in neighborhoods experiencing high levels of poverty to participate in the experiment. By lottery, the families were divided into a treatment group and two control groups. Families in the treatment group received vouchers that subsidized rent if they moved to more affluent neighborhoods. In one control group, families received rental vouchers without any restriction on where they could live. In the second control group, families received nothing at all. The researchers compared the subsequent life outcomes, such as earnings and educational attainment, for the family members in the three groups.

Some of the results were disappointing. Even though many in the treatment group used the vouchers to leave high-poverty neighborhoods, the adults in the three groups did not exhibit significant differences in economic outcomes. The average incomes of the adult family members were about the same in the treatment and control groups, though health outcomes were somewhat better for those in the treatment group. Similarly, there were no significant differences in the measured life outcomes for older children (ages 13 to 18) in the treatment group.

Yet the program had a significant, positive impact on children who were below the age of 13 when their families received the vouchers. Younger children in the treatment group did not perform better in school than those in the control groups, as measured by test scores in reading and math. But later in their lives, they had significantly higher rates of college attendance, lower rates of single motherhood, and higher incomes as adults. These results show that young children enjoy long-term benefits if their families leave high-poverty neighborhoods. ●

38-1b Observational Data

Because experimental data are not always available, economists often rely on **observational data**, which are obtained not from conducting an experiment but from simply observing the world as it is. Observational data can come from surveys of households and firms and from administrative records, such as tax returns. Compared with experimental data, observational data have the advantage of being more easily produced and more widely available, but they present two challenges to data analysts.

observational data
data that come from a researcher observing the world as it presents itself

The first is the problem of confounding variables. A **confounding variable** is a variable that is omitted from the analysis but, because it is related to the variables being measured and studied, can lead the researcher to an incorrect conclusion.

confounding variable
an omitted variable that can mislead the researcher because it is related to the variables of interest

For example, suppose you want to know whether reducing class size in grade school improves learning. You might be tempted to estimate the impact of class size by comparing the average test scores of students in large and small classes. That strategy would be fine if students and teachers were placed into the classes randomly, as is the case with experimental data. But with observational data, because the placement of students and teachers is probably not random, other variables related to class size may enter the picture and bias the results. Small classes, for instance, might be more common in towns with better-educated and higher-income populations. If parental education affects student performance, it is a confounding variable that makes the impact of small classes seem larger than

it actually is. The benefits of parental education could be incorrectly attributed to class size. Or perhaps school principals assign less-experienced teachers to smaller classes. If teacher experience affects student performance, it is a confounding variable that makes the impact of small classes appear smaller than it actually is. The disadvantages of having a less experienced teacher could mask the benefits of small classes. Because many variables can be correlated with one another in observational data, researchers need to carefully distinguish the effects of one variable from the effects of another.

reverse causality
a situation in which a researcher confuses the direction of influence between two variables

The second challenge presented by observational data is the problem of reverse causality. **Reverse causality** describes a situation in which a researcher believes that one variable influences a second variable, when, in fact, it is the second variable that influences the first.

For example, suppose you observe that the quantity consumed of some food is positively correlated with a person's body mass index (BMI), an indicator of obesity. Should you conclude that the consumption of that food causes an increase in BMI? That inference would be correct if food consumption were set randomly, as in a controlled trial, but problems can arise with observational data. If the food in question were ice cream, the direction of causality might indeed run from food consumption to BMI: Eating a lot of ice cream may cause weight gain. The positive correlation does not prove this, but the hypothesis is at least plausible. On the other hand, if the food were diet soft drinks, a different interpretation might be in order. Maybe people with a high BMI are trying to lose weight and therefore choose to consume diet soft drinks. That is, rather than the consumption of diet soft drinks causing a high BMI, a high BMI may cause the consumption of diet soft drinks. This example illustrates a general lesson: Sorting out what is cause and what is effect is often tricky when using observational data.

Despite these problems, observational data can be useful if the data analyst is careful. Later, this chapter introduces some of the methods econometricians have devised to deal with confounding variables and to determine causal effects.

38-1c Three Types of Data

Whether data are experimental or observational, they can come in three types: cross-sectional, time-series, and panel.

cross-sectional data
data that present information about multiple subjects (such as people, firms, or nations) at a given time

Cross-sectional data show the characteristics of multiple subjects (such as people, firms, or nations) at a given time. For example, we might survey a group of workers and ask each of them to report their wage, education, age, experience, profession, race, gender, and so on. We can use these data to see how these variables are related to one another. For example, we can examine how much wages differ by race or gender after adjusting for differences in education, age, experience, and profession.

time-series data
data that present information about a single subject (such as a person, firm, or nation) at various times

Time-series data show the characteristics of a single subject (one person, firm, or nation) at various times. For example, we might measure a nation's unemployment rate (the percent of the labor force that is jobless) and its GDP (gross domestic product, a measure of production and income) every year over a 60-year period. We can use these data to study how fluctuations in unemployment and GDP are related.

panel data
data that present information about multiple subjects (such as people, firms, or nations) at various times

Panel data combine the elements of cross-sectional and time-series data to show the characteristics of multiple subjects (such as people, firms, or nations) at various times. This type of data, also known as **longitudinal data**, is useful for examining how changes in one variable affect another. For example, we could study how winning the lottery affects a person's labor-force participation by comparing the changing behavior over time of lottery winners and losers.

QuickQuiz

1. In a randomized controlled trial, subjects are placed into treatment and control groups based on
   a. willingness to pay.
   b. income.
   c. estimated benefit from treatment.
   d. chance.

2. Observational data has the advantage of
   a. solving the problem of reverse causality.
   b. being widely available.
   c. avoiding confounding variables.
   d. coming from randomized controlled trials.

Answers are at the end of the chapter.

38-2 What Economists Do with Data

Having seen how economists gather data (from experiments or observations) and the types of data they gather (cross-sectional, time-series, or panel), let's consider what economists hope to achieve through data analysis.

38-2a Describing the Economy

Economic data are often interesting in themselves as quantitative descriptions of the world. For example:

- You may have heard that most people devote a large fraction of their spending to housing, but you might not know how large it is. The data show that average consumers in the United States spend 42 percent of their budgets on housing.
- You know that, by definition, a household at the 90th percentile of the U.S. income distribution has higher income than a household at the 10th percentile, but you might not know how different their incomes are. The data show that the rich household has about 12 times the income of the poor household.
- You probably know that the United States has higher income per person than Mexico, but you might not know how much higher it is. The data show that average income per person in the United States is about three times that in Mexico.
- You may have read that spending on healthcare has risen as a share of total spending in the economy, but you might not know by how much. The data show that U.S. healthcare spending has risen from 5 percent of GDP in 1960 to 18 percent in 2019.

These kinds of facts are useful to know. As we develop theories to understand how the world works and consider policies to improve it, keeping an eye on the data gives us a better sense of the world as it is.

38-2b Quantifying Relationships

Economic theory often suggests that certain variables are related, but it rarely tells us how strongly they are related. Often, we need a sense of those magnitudes. That is, we need estimates of a model's **parameters**, the numerical values that govern the strength of the relationships among variables.

parameters
the numerical values that govern the strength of the relationships among variables in a model

Consider an example. Suppose policymakers are considering a tax on luxury cars. They might want to know whether the burden of the tax will fall more on the buyers or the sellers of the cars. The incidence of a tax depends on the price

elasticities of supply and demand, which measure the responsiveness of quantity supplied and quantity demanded to changes in the good's price. If demand is more elastic than supply, sellers bear most of the burden; if supply is more elastic than demand, buyers bear most of the burden.

This theoretical conclusion takes us only so far. To answer the policymakers' question, we need estimates of the parameters, which here are the price elasticities of supply and demand. To come up with those estimates, researchers would collect data on the luxury car market. A careful analysis of the data can establish quantitatively the determinants of quantity supplied and quantity demanded. In particular, it would yield estimates of the price elasticities, which can be used to project the incidence of the proposed tax.

38-2c Testing Hypotheses

Economic theories attempt to describe the world in which we live. Like all theories in science, an economic theory is just a hypothesis, an educated guess about how the world works. To confirm or refute the hypothesis, we need to turn to data.

For example, consider the impact of schooling on wages. Economist Betsey thinks that education is a great way to increase a worker's wage. She believes that the human capital produced in school makes workers more productive, and more productive workers are paid more. Economist Justin thinks education is a waste of time. He believes that most things taught in school are useless in most jobs and that people are better off getting job experience than wasting time in the classroom.

No amount of theorizing can settle the debate between Betsey and Justin. Their disagreement is **empirical**: It can be addressed only by the facts, not by sheer logic. In this case, we need to turn to data on wages, education, and job experience to decide which of their hypotheses is correct. (Spoiler alert: Most economists side with Betsey.)

38-2d Predicting the Future

"It's tough to make predictions, especially about the future." Yogi Berra was a wise man, yet economists are often asked to predict the future. A microeconomist might be asked how the upcoming merger between two firms will affect prices in the market for their products. A macroeconomist might be asked how quickly a sudden spike in inflation will subside.

Sometimes, you can make forecasts simply by finding patterns in the data and extrapolating them into the future. For example, suppose you observe that when married couples trade in their compact cars for minivans, they usually have a new baby a few months later. If your neighbors come home with a minivan one day, you might reasonably predict that a baby is on the way. Economists call this relationship between minivans and babies an **empirical regularity**, and it might be useful for forecasting for a while. But the relationship is not reliably stable. If a car company introduced a new line of family-friendly SUVs, for instance, minivan purchases might become a less useful tool for predicting births.

To make reliable predictions, economists often turn to models, which are mathematical representations of the forces at work in a given situation. For a model to be useful in making quantitative predictions, economists need to quantify each relationship within it. They do so using the relevant data to estimate the model's parameters. Once the estimated model is in hand, they can use it to make predictions.

The FRB/US Model

An important model in economic policymaking is the Federal Reserve Board's model of the U.S. economy, abbreviated as FRB/US (and pronounced as "fur-bus"). The FRB/US model tries to describe the main macroeconomic elements of the U.S. economy, including the relationships among key variables such as GDP, inflation, unemployment, and interest rates. The central bank uses the model for forecasting and policy analysis.

The FRB/US model includes hundreds of equations, each describing a piece of the economy. Many of these equations are **identities**—equations that must be true because of how the variables in the equation are defined. An identity does not have any parameters that need to be estimated. (An example is the national income accounts identity $Y = C + I + G + NX$, which states that GDP is the sum of consumption, investment, government purchases, and net exports.) But about 60 equations in the FRB/US model are equations that describe how households or firms respond to economic conditions, and these equations include crucial parameters. For example, an equation for consumption would show how households' spending on consumer goods and services depends on their current income, expected future income, wealth, interest rates, and so on. The relative importance of these determinants of consumption is reflected in the parameters of the consumption equation. Economists at the Federal Reserve estimate these parameters by applying econometric techniques to the time-series data on the U.S. economy.

With the estimated FRB/US model in hand, Fed economists use the model for two purposes. The first is forecasting. Based on current policy and economic conditions, they project the most likely outcome for the future. The projection is based on simultaneously solving the hundreds of equations in the model. This task might seem impossible, and it would be if they had to rely on pencil and paper. Fortunately, computer algorithms are available to solve such large-scale models.

The second purpose for which Fed economists use the FRB/US model is policy analysis. They ask how the future would be different from their baseline projection if the Fed changed monetary policy in some way. The result is a set of alternative policy scenarios. They show what would happen to key economic variables—GDP, unemployment, inflation, and so on—if the Fed tightened or loosened monetary policy. The members of the Federal Open Market Committee, who set monetary policy, can use this menu of scenarios as a guide for choosing a direction for policy.

How reliable are the forecasts presented by the Fed economists? Studies find that they are as good as or better than those provided by most private economic forecasters, but they are far from perfect. Because the reliability of the forecasts depends on the accuracy of the FRB/US model, Fed economists are always looking for ways to improve it. Some improvements come from new conceptual insights, such as better economic theory. Others come from more data that arrive as time passes and from new statistical insights as econometricians devise better ways to use data. ●

QuickQuiz

3. Economists use data to
   a. describe the economy.
   b. estimate parameters.
   c. test hypotheses.
   d. All of the above.

Answers are at the end of the chapter.

38-3 The Methods of Data Analysis

Having discussed the kinds of data that economists use and what they hope to achieve using the data, let's consider some of the methods that econometricians have developed for data analysis.

38-3a Finding the Best Estimate

According to the theory of human capital, when workers become more educated, they become more productive, and their wages increase. That statement is **qualitative**: It addresses the nature of the relationship between education and wages but not the strength of it. Suppose you wanted to go beyond this qualitative statement and ask the **quantitative** question, "By how much does an extra year of schooling increase a worker's wage?" This question is empirical. You can only answer it using data.

You begin by surveying a number of workers and collecting data on their wages and education. It would look like Table 1. (You would likely want your sample to include more than seven workers, but for purposes of illustration, seven is sufficient.) The data in Table 1 are an example of cross-sectional data.

As you look over the data, you see that more educated workers do, in fact, tend to earn more. The two workers with 12 years of schooling, presumably high school graduates, earn an average of $25 an hour. The two workers with 16 years of schooling, perhaps college graduates, earn an average of $40 an hour. And the two workers with 18 years of schooling, who may have done some study in graduate school, earn an average of $45 an hour.

Yet wages do not always rise with education. Chloe has four more years of schooling than Brooke, but they earn the same wage. Emma has two more years of schooling than Flynn, but she earns $10 per hour less than he does. Education may be one determinant of a worker's wage, but there must be other important factors as well.

One way to start understanding these data is to graph them, as in Figure 1. Each point represents one observation. The figure shows the positive correlation between wages and education: Points to the right (indicating more years of schooling) tend to also be higher in the graph (indicating higher wages). Yet the points do not lie along a straight line or even a simple curve. They resemble a cloud, suggesting that there are factors beyond years of schooling that affect wages.

Table 1

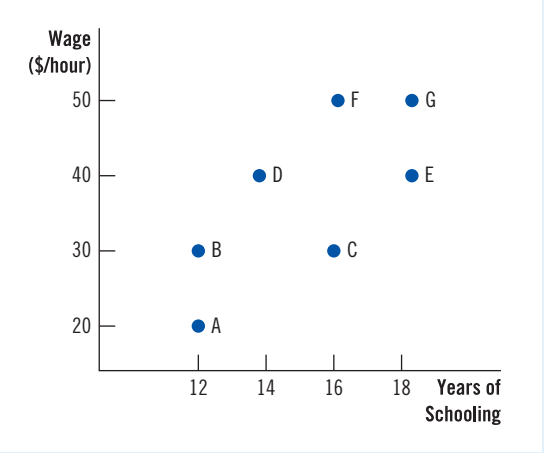
Data on Wages and Education

Worker	Wage ($/hour)	Years of Schooling
Andy	20	12
Brooke	30	12
Chloe	30	16
Diego	40	14
Emma	40	18
Flynn	50	16
Gina	50	18

Figure 1

A Scatterplot of the Data

Plotting data on wages and education shows the positive correlation between these two variables. That is, they tend to move in the same direction.



To determine how much each year of schooling increases a worker's wage, an economist will turn to a **statistical model**—a mathematical representation of the process that generates the data. The simplest such model is the following:

$$\text{WAGE}_i = \beta_0 + \beta_1 \times \text{SCHOOL}_i + \varepsilon_i$$

where β_0 and β_1 are parameters that measure how the variables are related. According to this model, person i's wage (WAGE_i) depends on years of schooling (SCHOOL_i) and is influenced by a random variable (ε_i). The variable on the left-hand side, WAGE_i, is called the **dependent variable**; it is the variable being explained. The measured variable on the right-hand side, SCHOOL_i, is called the **independent variable**; it is the variable taken as given. The term ε_i is called the **residual**. It represents the many forces, such as experience and cognitive ability, that influence wages but are excluded from the model. The residual is assumed to be zero on average and uncorrelated with the independent variable. (We will discuss the role of this assumption later, but for now, let's just go with it.)

This statistical model is called a **linear regression**. In essence, this model draws a line through the cloud of points, as shown in Figure 2. The line shows the best guess of a worker's wage based on years of schooling. The residual represents the deviation of the actual wage from the wage predicted by the line, acknowledging that the model will not fit the data perfectly.

linear regression
a statistical model in which the dependent variable is linearly related to one or more independent variables plus a random residual

The key parameter of interest is β_1, which tells us how much each year of schooling increases a person's wage. The other parameter, β_0, determines the intercept of the line. Taken literally, β_0 would be the average wage for a person with zero years of schooling. But because our sample does not include anyone with no education at all, it is best to avoid that literal interpretation. Our focus is on β_1.

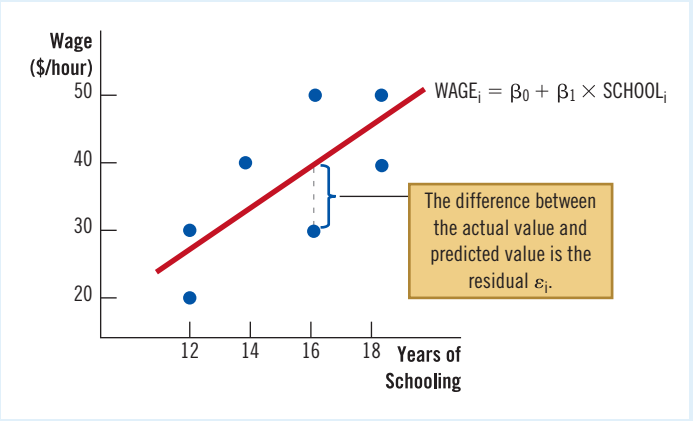
The question we now face is how to best estimate the parameters from the data we have. We could just try to draw a well-fitting line through the cloud of points by hand, but that approach is too inexact (and not easily generalized to more complex cases discussed in a moment). The standard method for finding the best-fitting line is called **ordinary least squares**, or OLS. We won't go into the details of OLS in this chapter, but the intuition is simple. OLS aims to determine the parameters

ordinary least squares
a statistical method for estimating parameter values by minimizing the sum of squared residuals

Figure 2

Estimating the Best-Fitting Line

A statistical model posits that wages are a linear function of education plus a residual representing other random influences on wages. The parameters of the model (β_0 and β_1) can be estimated by ordinary least squares (OLS), which yields the line that fits best as gauged by the sum of squared residuals.



(β_0 and β_1) representing the line that is closest to the data points. Closeness here is measured by the squared residuals. Squaring the residuals ensures that both positive and negative residuals are deemed to detract from the goal of closeness. OLS finds the parameters that minimize the sum of squared residuals.

When OLS is applied to the seven data points in Table 1, it yields the following result:

$$\text{WAGE}_i = -10.7 + 3.16 \times \text{SCHOOL}_i.$$

According to the estimated model, each year of schooling increases a worker's wage by $3.16 per hour. That is the answer to our question.

This example illustrates a general lesson: Economists often want to go beyond qualitative theoretical insights (education increases wages) to quantitative conclusions (how much each year of schooling increases wages). This leap requires data. The economists find the data relevant for the issue at hand, posit a statistical model that can plausibly explain the data, and estimate the model's parameters using a method such as ordinary least squares. Using the estimated model, they can reach quantitative conclusions.

38-3b Gauging Uncertainty

Economists use data to estimate key quantities (such as the wage benefit of a year of schooling in the previous section). Often, they want to know not only the best estimate but also how reliable that estimate is. That is, they want to know whether their estimate is precise or a rough guess.

Before returning to our model of wages and education, let's take a detour to consider a simpler example. Suppose you are interested in the height of the average person in New York City. One way to find that number is to get data on the height of every New Yorker. You then compute the average (also called the **mean**) by adding up all their heights and dividing by the number of people. Because your calculation includes the entire population, you arrive at the precisely correct answer. But there is a problem: New York has about 9 million people, so this approach is not practical.

Fortunately, there is an easier way if you are content with an answer that is not completely precise. Instead of trying to get data on the entire population, you can

choose a random sample of, say, 100 people. The meaning of "random" is that every person in the population has an equal chance of being chosen. Because of the randomness, the sample is likely to be representative of the population. You can estimate the height of the population by computing the average height of these 100 randomly chosen people.

Let's say that the average height of the 100 people in your sample is 66 inches. So far, so good. But you might start wondering how reliable this estimate is. Because the sample was chosen randomly, there is no reason to think that the sample average of 66 inches is too high or too low an estimate of the average height of the population. But it could be either. You could have gotten unlucky by randomly choosing a sample with a few more tall people or a few more short people than normal. This uncertainty comes from what statisticians call **sampling variation**—the variability that arises because different random samples lead to somewhat different estimates.

Statisticians have developed ways not only to estimate parameter values but also to gauge the uncertainty associated with parameter estimates resulting from sampling variation. The details of this process are beyond the scope of this chapter, but the basic idea can be gleaned from our example. Here is how it works.

First, after computing the mean height of the 100 people in your sample, you compute the standard deviation of heights. The **standard deviation** is a measure of variability across observations, and you may have learned about it in a math or statistics class. In our example, let's say you compute that the standard deviation of heights is 4 inches.

What is a standard deviation? Technically, it is the square root of the average squared deviation from the mean. That is a mouthful, but there is a more intuitive way to think about it. For a normal bell-shaped distribution, about 95 percent of observations fall within two standard deviations of the mean. In our example, the mean is 66 inches, and the standard deviation is 4 inches (making two standard deviations 8 inches). So, if you pick a New Yorker at random, the probability that this individual's height lies between 58 and 74 inches is 95 percent.

Next, using the standard deviation and the sample size, you can compute a measure of your estimate's reliability called the **standard error**. According to a formula developed by statisticians, the standard error of the sample mean as an estimate of the population mean is the standard deviation divided by the square root of the sample size. In this case, the standard error of your estimate is $4/\sqrt{100} = 4/10 = 0.4$. This number can be used to gauge the degree of sampling variation in your estimate. Just as the standard deviation measures the variability in the height of individual New Yorkers, the standard error measures the variability in the mean height of samples of New Yorkers.

standard error
a measure of the uncertainty associated with a parameter estimate that results from sampling variation

Here is a useful rule of thumb: The true value of a parameter lies within two standard errors of the estimated value about 95 percent of the time. In this example, the estimate is 66 inches, and the standard error is 0.4 inches. Two standard errors equal 0.8 inches. So you can be 95 percent confident that the true mean height of the population is between 65.2 inches and 66.8 inches.

Based on this rule of thumb, twice the standard error is sometimes called the **margin of error**. Journalists often use this term when reporting results from public opinion polls. For example, you might hear that, according to a poll of 400 people, 57 percent favor a particular candidate, with a margin of error of 5 percent. That means you can be 95 percent confident that the true support for that candidate lies between 52 and 62 percent.

Standard errors are useful not only in simple examples like estimating a population mean but also in other contexts. Depending on the circumstances, the formulas

to calculate standard errors can be complex. Fortunately, most statistical software used to produce parameter estimates automatically gives the standard errors associated with the estimates.

For the data on wages and education in Table 1, the estimated equation produced by Microsoft Excel, along with the standard errors in parentheses, is the following:

$$\text{WAGE}_i = -10.7 + 3.16 \times \text{SCHOOL}_i.$$
$$(20.7) \quad (1.35)$$

It turns out that the estimate of \$3.16 for each year of schooling is not very precise. Two standard errors are $2 \times 1.35 = 2.70$. Thus, we can be 95 percent confident that the true wage benefit of a year of schooling lies between \$0.46 and \$5.86. That is a large range. But we should not expect much precision when estimating a parameter using only seven data points. If we had 700 data points similar to these, the standard error would be 0.135, and the 95-percent confidence interval would lie between \$2.89 and \$3.43, which is a smaller range. Estimated parameters become more precise with larger samples.

38-3c Accounting for Confounding Variables

In many situations, the dependent variable is a function of more than one other variable. Wages, for example, depend not only on education but also on experience, cognitive ability, job characteristics, and so on. If data analysts are not careful, they can mistakenly confuse the effects of one variable with the effects of another. Fortunately, statistical methods are available to help avoid this problem of confounding variables.

As an example, let's continue with our statistical model of wages and education:

$$\text{WAGE}_i = \beta_0 + \beta_1 \times \text{SCHOOL}_i + \varepsilon_i.$$

Earlier, we assumed that the residual ε_i had a zero mean and was uncorrelated with the independent variable SCHOOL_i. The assumption of a zero mean is not crucial: If the mean is not zero, it will alter only the estimate of the constant term β_0. Because the constant term is not the key parameter of interest, we are not led astray.

The assumption that the residual is uncorrelated with the independent variable, however, is fraught with potential problems. The residual reflects all the forces that influence wages other than years of schooling. If those other forces are correlated with schooling, OLS will yield an incorrect estimate of the effect of schooling on wages. The problem is one of confounding variables.

For example, suppose that some people are simply smarter than others. Cognitive ability is another plausible determinant of a worker's wage. If ability is unmeasured, it will be reflected in the residual. That is not a problem as long as ability and schooling are not correlated. Estimating the model above would, on average, give the correct estimate of the benefit of schooling. In the language of statistics, the estimate of β_1 would be **unbiased**.

But suppose that people with greater ability get more years of schooling than people with less ability. In this case, the residual (which includes ability) would be positively correlated with the independent variable (schooling). The value of β_1 estimated by OLS would reflect not only the effect of schooling but also, to some degree, the effect of cognitive ability. In other words, OLS would confound the effect of the independent variable, schooling, with the effect of a related omitted variable, ability. As a result, the estimate of β_1 would be **biased upward**. That is, it would indicate that schooling has a more powerful effect on wages than it really does.

What is to be done? One approach is to find some way to measure the confounding variable. Suppose we return to our seven workers and find that, as children, they each took an IQ (intelligence quotient) test, a measure of cognitive ability. Table 2 shows the expanded data. As expected, IQ and schooling are positively correlated: Higher IQ workers tend to have more years of schooling. If IQ might have its own direct effect on wages, beyond its indirect effect through schooling, then our previous estimate of β_1 is suspect.

All is not lost, however. We can expand our statistical model as follows:

$$\text{WAGE}_i = \beta_0 + \beta_1 \times \text{SCHOOL}_i + \beta_2 \times \text{IQ}_i + \varepsilon_i.$$

In this new statistical model, a worker's wage depends on education and ability as measured by IQ. Because this statistical model has more than one independent variable, it is called a **multiple regression**.

multiple regression
a linear regression model with more than one independent variable

We can again apply ordinary least squares to estimate the model's parameters. OLS will now choose β_0, β_1, and β_2 to minimize the sum of squared residuals. OLS yields unbiased estimates of the parameters as long as the residual is uncorrelated with all the independent variables (schooling and IQ in this example). And the results will be reliable even though the independent variables are correlated with each other. In this case, a multiple regression estimated by OLS lets us sort out the relative importance of schooling and IQ in determining wages.

Estimating this model using the data in Table 2 yields the following results (with standard errors to gauge parameter uncertainty in parentheses):

$$\underset{}{\text{WAGE}_i} = \underset{}{-49.1} + \underset{(29.5)}{1.86} \times \text{SCHOOL}_i + \underset{(0.35)}{0.57} \times \text{IQ}_i.$$
$$\qquad\qquad\quad (1.41)$$

As we expected, once we control for IQ, the estimated effect of schooling decreases. The new estimate suggests that each year of schooling increases a worker's wage by \$1.86 per hour rather than \$3.16 per hour as estimated when we did not include IQ in the model.

To sum up: When an omitted variable (ability as measured by IQ in our example) directly influences the dependent variable (wages), and the omitted variable is correlated with the independent variable (schooling), OLS yields misleading results.

Table 2

Data on Wages, Education, and IQ

Worker	Wage (\$/hour)	Years of Schooling	IQ
Andy	20	12	90
Brooke	30	12	100
Chloe	30	16	90
Diego	40	14	105
Emma	40	18	105
Flynn	50	16	100
Gina	50	18	120

The OLS estimate confounds the effect of the independent variable with the effect of the omitted variable. One way to deal with this problem is to include the previously omitted variable in a multiple regression.

Multiple regression, however, is not the only way to handle the problem of confounding variables. The next section considers another approach.

38-3d Establishing Causal Effects

Data analysts are often interested in the causal effect of one variable on another. For instance, if workers get one more year of schooling, what change in their wages will that cause? If people double their intake of a particular food, what change in their body mass index will that cause? The estimation of causal effects using observational data is difficult due to both the potential omission of confounding variables and the possibility of reverse causality.

natural experiment
a chance event that causes variation in the data similar to that generated by a randomized controlled trial

Sometimes, these problems can be addressed by exploiting a **natural experiment**. A natural experiment is a chance event that generates variation in the data as if a randomized controlled trial had been conducted.

Consider an example. Imagine that one day, Phyllis Philanthropist gives a talk at a high school and makes a surprise announcement: Phyllis will pay four years of college tuition for all the students who graduate from that school and go on to college. This is great news for those students, more of whom will likely continue their education. Across town, there is another high school, much the same as the first but without a generous philanthropist. This situation offers a natural experiment. The first high school is the treatment group, and the second is the control group. By comparing the years of schooling and subsequent wages of the two groups, we can measure the causal effect of extra schooling.

Economists have developed a statistical approach to measure causal effects in data from natural experiments like this one and in similar situations. It is called the **instrumental variables** method, and you will study it if you take a course in econometrics. The key to this method is to find some random variable, called the **instrument**, that meets two conditions:

1. The instrument is correlated with the independent variable of interest.
2. The instrument does not affect the dependent variable other than through its effect on the independent variable.

In our example, the instrument is Phyllis's generosity in the first high school and its absence in the other. This random act increases the schooling of those in the treatment group (condition 1), but it does not affect their subsequent wages other than by increasing their schooling (condition 2). Under these conditions, we can use the instrumental variables method to identify the causal effect of schooling on wages.

For a numerical example, let's return to the workers in Table 1. Suppose Chloe, Emma, and Flynn attended the high school that received Phyllis's offer, while Andy, Brooke, Diego, and Gina attended the high school that did not. Simple calculation shows that the treatment group (Chloe, Emma, and Flynn) has average schooling of 16.7 years and an average wage of $40 per hour, while the control group (Andy, Brooke, Diego, and Gina) has average schooling of 14 years and an average wage of $35 per hour. Apart from exposure to the treatment (Phyllis's generosity), the two high schools are assumed to be the same. Because the treatment group has 2.7 more years of schooling and earns $5 per hour more than

the control group, we estimate that each year of schooling increases wages by \$5/2.7, or \$1.85 per hour.

The use of natural experiments always raises thorny questions about whether the experiment is as random and easily interpreted as it seems. In our example, the key question is whether the two high schools really are the same aside from Phyllis's offer. In reality, the students might not have been sorted into the two schools randomly. Maybe student placement was based on residence, and one side of town has a richer and better-educated population than the other. Maybe Phyllis made her offer to the one school because its students seemed especially hardworking or especially in need. Such differences could bias the results. Whenever researchers rely on natural experiments, they need to consider whether the process generating the data differs from a randomized controlled trial and whether these differences could contaminate their findings.

Despite these caveats, finding a natural experiment is often the best way to estimate the causal effect of one variable on another, as the following case study shows.

Case Study

How Military Service Affects Civilian Earnings

What is the effect of a period of military service on a person's subsequent earnings as a civilian? This question is important both for personal decision making and for public policy toward the military.

There are plausible arguments in both directions. One might argue that civilian earnings would be higher for those with some military service on their résumés because the military teaches discipline, teamwork, and valuable job skills. Alternatively, one might argue that civilian earnings would be lower for those with military service because the military takes time away from the private-sector experience and on-the-job training that a civilian job would provide. Either of these hypotheses could be true. Only the data can decide.

Consider how we can use data to address the issue. A starting point might be to compare workers with past military service and those without. The difference between the average earnings of the two groups might be taken to be the effect of military service.

That approach, however, has a problem. The difference in earnings might result from the personal characteristics of people who join the military rather than from a causal effect of military experience. Perhaps people who join the military are already more disciplined and good at teamwork than those who forgo the experience. In this case, those with military service would earn more even if the military service had no effect. Or perhaps those who join the military do so because they don't have the skills for well-paid civilian jobs. In this case, those with military service would earn less even if the military service had no effect. Because there are many omitted variables, military service cannot be interpreted as the cause of the difference in average earnings between these groups.

So how can we answer the question? A randomized controlled trial would do the trick. We could randomly divide the population into two groups: one that is forced into a period of military service and one that is prohibited from it. We then compare the subsequent civilian earnings of the two groups. The difference must reflect the causal effect of military service because the randomization ensures that the two groups are otherwise the same.

We cannot conduct exactly that experiment, but historically, the U.S. government has done something close. In the early 1970s, during the Vietnam War, young men

were drafted into the military based on a lottery. If a young man had a low lottery number, he was likely to be drafted. If he had a high lottery number, he would likely avoid military service. To be sure, the lottery number was not the only determinant of military service: The rich and well-connected could more easily avoid the draft, and those with high lottery numbers could nonetheless volunteer. But the lottery number itself was completely random.

In an important study, the economist Joshua Angrist noted that the draft lottery number is an ideal variable for applying the instrumental variables method. The lottery number satisfies the two conditions discussed earlier. It affected military service (condition 1). And it did not affect subsequent earnings other than through its effect on military service (condition 2).

What did Angrist learn from studying this natural experiment? Here is his bottom line: "In the early 1980s, long after their service in Vietnam had ended, the earnings of white veterans were approximately 15 percent less than the earnings of comparable nonveterans." Serving your country in the military can be a noble act, perhaps even more so once we recognize the long-term economic cost to many of those who serve.

In 2021, Angrist won the Nobel Prize for his "methodological contributions to the analysis of causal relationships." ●

QuickQuiz

4. Ordinary least squares is a statistical technique to
   a. find the best fitting parameters.
   b. avoid the problem of reverse causality.
   c. identify the smallest rectangle that includes all the data.
   d. turn observational data into experimental data.
5. Standard errors are used to
   a. categorize common mistakes.
   b. gauge the reliability of estimates.
   c. avoid confounding variables.
   d. provide more accurate forecasts.
6. Multiple regression is used
   a. to avoid the problem of reverse causality.
   b. to avoid the problem of large standard errors.
   c. when a statistical model has two or more independent variables.
   d. when a statistical model has two or more dependent variables.
7. The technique of instrumental variables is used
   a. to avoid the problem of reverse causality.
   b. to avoid the problem of large standard errors.
   c. when there is an excessive amount of data.
   d. when there is an insufficient amount of data.

Answers are at the end of the chapter.

38-4 Conclusion

This chapter has been a whirlwind tour of the large and often technical field of econometrics. It discussed the kinds of data that economists use, what they aim to achieve with their data analysis, and how various statistical methods help in drawing reliable inferences from the data. But to feel ready to apply the tools of econometrics yourself, you will need to take a full course on the topic. The brief introduction in this chapter provides a foundation for that future study.

Chapter in a Nutshell

- Economists use two kinds of data to study how the world works: experimental data obtained from randomized controlled trials and observational data obtained from surveys and administrative records. Interpreting observational data requires extra care due to the problems of confounding variables and reverse causality.
- There are three types of data. Cross-sectional data present information about multiple subjects (such as people, firms, or nations) at a given point in time. Time-series data present information about a single subject over time. Panel data present information about multiple subjects over time.
- Economists usually have one of four goals when using data: describing the economy, quantifying relationships among variables, testing hypotheses, or predicting the future.
- To quantify relationships, statistical methods are used to find parameter estimates that best fit the data. One such method is ordinary least squares.
- Statistical methods not only estimate parameters but also determine the uncertainty associated with those estimates that arises from sampling variation. An estimate's standard error is a measure of that uncertainty.
- Data analysts can be led astray if a confounding variable is correlated with the independent variable and omitted from the statistical model. One approach to dealing with this problem is to add the confounding variable to the model and use multiple regression to estimate the true effect of the independent variable of interest.
- To estimate the causal effect of one variable on another, data analysts need to be careful about confounding variables and reverse causality. One approach is to look for natural experiments.

Key Concepts

data, p. 819
econometrics, p. 820
randomized controlled trial, p. 820
experimental data, p. 820
observational data, p. 821
confounding variable, p. 821
reverse causality, p. 822
cross-sectional data, p. 822
time-series data, p. 822
panel data, p. 822
parameters, p. 823
linear regression, p. 827
ordinary least squares, p. 827
standard error, p. 829
multiple regression, p. 831
natural experiment, p. 832

Questions for Review

1. Explain the difference between experimental and observational data.
2. Why don't economists always use experimental data?
3. What two problems arise in the analysis of observational data?
4. Explain the difference between cross-sectional data and time-series data, and give an example of each.
5. How does ordinary least squares choose the parameter values of a statistical model?
6. What does the standard error of a parameter estimate measure?
7. Explain the problem of confounding variables, and describe two methods for solving the problem.

Problems and Applications

1. Choose the right words to complete this summary of a hypothetical research project.
   a. Ellie, an economist, wants to study how population growth affects national income. She collects data on 50 countries, measuring each country's population growth rate and national income per person. This is an example of [cross-sectional, time-series] data.
   b. She posits a statistical model in which national income depends on population growth. She plots the data and draws the best-fitting line through the points using [randomized controlled trials, ordinary least squares].
   c. There is a negative relationship: Countries with higher population growth tend to have lower incomes. She concludes that this finding is not

due to sampling variation because she has a small [sample size, standard error].

d. Ellie recognizes that her data is [experimental, observational]…

e. …because it was not generated by a [randomized controlled trial, multiple regression].

f. She worries that a country's average educational attainment might affect both its income and population growth, leading to the problem of [confounding variables, linear regression].

g. She finds data on educational attainment in each country and adds it to her statistical model using [panel data, multiple regression].

h. Ellie also worries that the level of income may affect the availability of birth control and thereby population growth, leading to the problem of [reverse causality, standard errors].

i. She learns that some countries benefited from a UN program disseminating birth control while others did not and that the choice of countries enrolled in the program was random. She recognizes that this policy provides a [natural experiment, linear regression].

j. She can now estimate the causal impact of population growth on income using the technique of [ordinary least squares, instrumental variables].

QuickQuiz Answers

1. **d** 2. **b** 3. **d** 4. **a** 5. **b** 6. **c** 7. **a**

Glossary

A

ability-to-pay principle the idea that taxes should be levied on a person according to how well that person can shoulder the burden

absolute advantage the ability to produce a good using fewer inputs than another producer

accounting profit total revenue minus total explicit cost

adverse selection the tendency for the mix of unobserved attributes to become undesirable from the standpoint of an uninformed party

agent a person who performs an act for another person, called the principal

aggregate-demand curve a curve that shows the quantity of goods and services that households, firms, the government, and customers abroad want to buy at each price level

aggregate-supply curve a curve that shows the quantity of goods and services that firms choose to produce and sell at each price level

appreciation an increase in the value of a currency as measured by the amount of foreign currency it can buy

Arrow's impossibility theorem a mathematical result showing that, under certain assumed conditions, there is no method for aggregating individual preferences into a valid set of social preferences

automatic stabilizers changes in fiscal policy that stimulate aggregate demand when the economy goes into a recession but that occur without policymakers having to take any deliberate action

average fixed cost fixed cost divided by the quantity of output

average revenue total revenue divided by the quantity sold

average tax rate total taxes divided by total income

average total cost total cost divided by the quantity of output

average variable cost variable cost divided by the quantity of output

B

balanced trade a situation in which exports equal imports

bank capital the resources a bank's owners have put into the institution

behavioral economics the subfield of economics that integrates the insights of psychology

benefits principle the idea that people should pay taxes based on the benefits they receive from government services

bond a certificate of indebtedness

budget constraint the limit on the consumption bundles that a consumer can afford

budget deficit a shortfall of tax revenue from government spending

budget surplus an excess of tax revenue over government spending

business cycle fluctuations in economic activity, such as employment and production

C

capital the equipment and structures used to produce goods and services

capital flight a large and sudden reduction in the demand for assets located in a country

capital requirement a government regulation specifying a minimum amount of bank capital

cartel a group of firms acting in unison

catch-up effect the property whereby countries that start off poor tend to grow more rapidly than countries that start off rich

central bank an institution designed to oversee the banking system and regulate the quantity of money in the economy

circular-flow diagram a visual model of the economy that shows how dollars flow through markets among households and firms

classical dichotomy the theoretical separation of nominal variables and real variables

closed economy an economy that does not interact with other economies in the world

club goods goods that are excludable but not rival in consumption

Coase theorem the proposition that if private parties can bargain without cost over the allocation of resources, they can solve the problem of externalities on their own

collective bargaining the process by which unions and firms agree on the terms of employment

collusion an agreement among firms in a market about quantities to produce or prices to charge

commodity money money that takes the form of a commodity with intrinsic value

common resources goods that are rival in consumption but not excludable

comparative advantage the ability to produce a good at a lower opportunity cost than another producer

compensating differential a difference in wages that arises to offset the nonmonetary characteristics of different jobs

competitive market a market with many buyers and sellers trading identical products so that each buyer and seller is a price taker

complements two goods for which an increase in the price of one leads to a decrease in the demand for the other

compounding the accumulation of a sum of money in, say, a bank account, where the interest earned remains in the account to earn additional interest in the future

Condorcet paradox the failure of majority rule to produce transitive preferences for society

constant returns to scale the property whereby long-run average total cost stays the same as the quantity of output changes

consumer price index (CPI) a measure of the overall cost of the goods and services bought by a typical consumer

consumer surplus the amount a buyer is willing to pay for a good minus the amount the buyer actually pays for it

consumption spending by households on goods and services, with the exception of purchases of new housing

core CPI a measure of the overall cost of consumer goods and services excluding food and energy

corrective tax a tax designed to induce private decision makers to take into account the social costs that arise from a negative externality

cost the value of everything a seller must give up to produce a good

cost–benefit analysis a study that compares the costs and benefits to society of providing a public good

cross-price elasticity of demand a measure of how much the quantity demanded of one good responds to a change in the price of another good, calculated as the percentage change in the quantity demanded of the first good divided by the percentage change in the price of the second good

crowding out a decrease in investment that results from government borrowing

crowding-out effect the offset in aggregate demand that results when expansionary fiscal policy raises the interest rate and thereby reduces investment spending

currency the paper bills and coins in the hands of the public

cyclical unemployment the deviation of unemployment from its natural rate

D

deadweight loss the fall in total surplus that results from a market distortion

demand curve a graph of the relationship between the price of a good and the quantity demanded

demand deposits balances in bank accounts that depositors can access on demand by writing a check

demand schedule a table that shows the relationship between the price of a good and the quantity demanded

depreciation a decrease in the value of a currency as measured by the amount of foreign currency it can buy

depression a severe recession

diminishing marginal product the property whereby the marginal product of an input declines as the quantity of the input increases

diminishing returns the property whereby the benefit from an extra unit of an input declines as the quantity of the input increases

discount rate the interest rate on the loans that the Fed makes to banks

discouraged workers individuals who would like to work but have given up looking for a job

discrimination the offering of different opportunities to similar individuals who differ only by race, ethnicity, gender, age, religion, sexual orientation, or other personal characteristics

diseconomies of scale the property whereby long-run average total cost rises as the quantity of output increases

diversification the reduction of risk achieved by replacing a single risk with a large number of smaller, imperfectly correlated risks

dominant strategy a strategy that is best for a player in a game regardless of the strategies chosen by the other players

E

economic profit total revenue minus total cost, including both explicit and implicit costs

economics the study of how society manages its scarce resources

economies of scale the property whereby long-run average total cost falls as the quantity of output increases

efficiency the property of society getting the most it can from its scarce resources

efficiency wages above-equilibrium wages paid by firms to increase worker productivity

efficient markets hypothesis the theory that asset prices reflect all publicly available information about the value of an asset

efficient scale the quantity of output that minimizes average total cost

elasticity a measure of the responsiveness of quantity demanded or quantity supplied to a change in one of its determinants

equality the property of distributing economic prosperity uniformly among the members of society

equilibrium a situation in which the market price has reached the level at which quantity supplied equals quantity demanded

equilibrium price the price that balances quantity supplied and quantity demanded

equilibrium quantity the quantity supplied and the quantity demanded at the equilibrium price

excludability the property of a good whereby a person can be prevented from using it

explicit costs input costs that require an outlay of money by the firm

exports goods produced domestically and sold abroad

externality the uncompensated impact of one person's actions on the well-being of a bystander

F

factors of production the inputs used to produce goods and services

federal funds rate the interest rate at which banks make overnight loans to one another

Federal Reserve (Fed) the central bank of the United States

fiat money money without intrinsic value that is used as money by government decree

finance the field that studies how people make decisions regarding the allocation of resources over time and the handling of risk

financial intermediaries financial institutions through which savers can indirectly provide funds to borrowers

financial markets financial institutions through which savers can directly provide funds to borrowers

financial system the group of institutions in the economy that help to match one person's saving with another person's investment

firm-specific risk risk that affects only a single company

fiscal policy the setting of the levels of government spending and taxation by government policymakers

Fisher effect the one-for-one adjustment of the nominal interest rate to the inflation rate

fixed costs costs that do not vary with the quantity of output produced

fractional-reserve banking a banking system in which banks hold only a fraction of deposits as reserves

free rider a person who receives the benefit of a good but avoids paying for it

frictional unemployment unemployment that results because it takes time for workers to search for the jobs that best suit their tastes and skills

fundamental analysis the study of a company's accounting statements and future prospects to determine its value

future value the amount of money in the future that an amount of money today will yield, given prevailing interest rates

G

game theory the study of how people behave in strategic situations

GDP deflator a measure of the price level calculated as the ratio of nominal GDP to real GDP times 100

Giffen good a good for which an increase in the price raises the quantity demanded

government purchases spending on goods and services by local, state, and federal governments

gross domestic product (GDP) the market value of all final goods and services produced within a country in a given period of time

H

horizontal equity the idea that taxpayers with similar abilities to pay taxes should pay the same amount

human capital the knowledge and skills that workers acquire through education, training, and experience

I

implicit costs input costs that do not require an outlay of money by the firm

imports goods produced abroad and sold domestically

incentive something that induces a person to act

income effect the change in consumption that results when a price change moves the consumer to a higher or lower indifference curve

income elasticity of demand a measure of how much the quantity demanded of a good responds to a change in consumers' income, calculated as the percentage change in quantity demanded divided by the percentage change in income

indexation the automatic correction by law or contract of a dollar amount for the effects of inflation

indifference curve a curve that shows consumption bundles that give the consumer the same level of satisfaction

inferior good a good for which, other things being equal, an increase in income leads to a decrease in demand

inflation an increase in the overall level of prices in the economy

inflation rate the percentage change in the price index from the preceding period

inflation tax the revenue the government raises by creating money

informational efficiency the description of asset prices that rationally reflect all available information

in-kind transfers transfers given in the form of goods and services rather than cash

interest on reserves the interest rate paid to banks on the reserves held in deposit at the Fed

internalizing the externality altering incentives so that people take into account the external effects of their actions

investment spending on business capital, residential capital, and inventories

J

job search the process by which workers find appropriate jobs given their tastes and skills

L

labor force the total number of workers, including both the employed and the unemployed

labor-force participation rate the percentage of the adult population that is in the labor force

law of demand the claim that, other things being equal, the quantity demanded of a good falls when the price of the good rises

law of supply the claim that, other things being equal, the quantity supplied of a good rises when the price of the good rises

law of supply and demand the claim that the price of any good adjusts to bring the quantity supplied and the quantity demanded for that good into balance

leverage the use of borrowed money to supplement existing funds for investment purposes

leverage ratio the ratio of assets to bank capital

liberal contractarianism the political philosophy according to which the government should choose policies deemed just, as evaluated by impartial observers behind a "veil of ignorance"

libertarianism the political philosophy according to which the government should punish crimes and enforce voluntary agreements but not redistribute income

life cycle the regular pattern of income variation over a person's life

liquidity the ease with which an asset can be converted into the economy's medium of exchange

lump-sum tax a tax that is the same amount for every person

M

macroeconomics the study of economy-wide phenomena, including inflation, unemployment, and economic growth

marginal change an incremental adjustment to a plan of action

marginal cost the increase in total cost that arises from an extra unit of production

marginal product the increase in output that arises from an additional unit of input

marginal product of labor the increase in the amount of output from an additional unit of labor

marginal rate of substitution the rate at which a consumer is willing to trade one good for another

marginal revenue the change in total revenue from an additional unit sold

marginal tax rate the increase in taxes from an additional dollar of income

market a group of buyers and sellers of a particular good or service

market economy an economy that allocates resources through the decentralized decisions of many firms and households as they interact in markets for goods and services

market failure a situation in which a market left on its own does not allocate resources efficiently

market for loanable funds the market in which those who want to save supply funds and those who want to borrow to invest demand funds

market power the ability of a single economic actor (or small group of actors) to have a substantial influence on market prices

market risk risk that affects all companies in the stock market

maximin criterion the claim that the government should aim to maximize the well-being of the worst-off person in society

median voter theorem a mathematical result showing that if voters are choosing a point along a line and they all want the point closest to their own optimum, then majority rule will pick the optimum of the median voter

medium of exchange an item that buyers give to sellers when they want to purchase goods and services

menu costs the costs of changing prices

microeconomics the study of how households and firms make decisions and how they interact in markets

model of aggregate demand and aggregate supply the model that most economists use to explain short-run fluctuations in economic activity around its long-run trend

monetary neutrality the proposition that changes in the money supply do not affect real variables

monetary policy the setting of the money supply by policymakers in the central bank

money the set of assets in an economy that people regularly use to buy goods and services

money multiplier the amount of money that results from each dollar of reserves

money supply the quantity of money available in the economy

monopolistic competition a market structure in which many firms sell products that are similar but not identical

monopoly a firm that is the sole seller of a product without close substitutes

monopsony a market that has only one buyer

moral hazard the tendency of a person who is imperfectly monitored to

engage in dishonest or otherwise undesirable behavior

multiplier effect the additional shifts in aggregate demand that result when expansionary fiscal policy increases income and thereby increases consumer spending

mutual fund an institution that sells shares to the public and uses the proceeds to buy a portfolio of stocks and bonds

N

Nash equilibrium a situation in which economic actors interacting with one another each choose their best strategy given the strategies that all the other actors have chosen

national saving (saving) the total income in the economy that remains after paying for consumption and government purchases

natural level of output the production of goods and services that an economy achieves in the long run when unemployment is at its normal rate

natural monopoly a type of monopoly that arises because a single firm can supply a good or service to an entire market at a lower cost than could two or more firms

natural-rate hypothesis the claim that unemployment eventually returns to its normal, or natural, rate, regardless of the rate of inflation

natural rate of unemployment the normal rate of unemployment around which the unemployment rate fluctuates

natural resources the inputs into the production of goods and services that are provided by nature, such as land, rivers, and mineral deposits

negative income tax a tax system that collects revenue from high-income households and gives subsidies to low-income households

net capital outflow the purchase of foreign assets by domestic residents minus the purchase of domestic assets by foreigners

net exports spending on domestically produced goods by foreigners (exports) minus spending on foreign goods by domestic residents (imports)

nominal exchange rate the rate at which a person can trade the currency of one country for the currency of another

nominal GDP the production of goods and services valued at current prices

nominal interest rate the interest rate as usually reported without a correction for the effects of inflation

nominal variables variables measured in monetary units

normal good a good for which, other things being equal, an increase in income leads to an increase in demand

normative statements claims that attempt to prescribe how the world should be

O

oligopoly a market structure in which only a few sellers offer similar or identical products

open economy an economy that interacts freely with other economies around the world

open-market operations the purchase and sale of U.S. government bonds by the Fed

opportunity cost whatever must be given up to obtain some item

P

perfect complements two goods with right-angle indifference curves

perfect substitutes two goods with straight-line indifference curves

permanent income a person's normal income

Phillips curve a curve that shows the short-run trade-off between inflation and unemployment

physical capital the stock of equipment and structures that are used to produce goods and services

political economy the study of government using the analytic methods of economics

positive statements claims that attempt to describe the world as it is

poverty line an absolute level of income set by the federal government for each family size below which a family is deemed to be in poverty

poverty rate the percentage of the population whose family income falls below an absolute level called the poverty line

present value the amount of money today needed to produce a future amount of money, given prevailing interest rates

price ceiling a legal maximum on the price at which a good can be sold

price discrimination the business practice of selling the same good at different prices to different customers

price elasticity of demand a measure of how much the quantity demanded of a good responds to a change in its price, calculated as the percentage change in quantity demanded divided by the percentage change in price

price elasticity of supply a measure of how much the quantity supplied of a good responds to a change in its price, calculated as the percentage change in quantity supplied divided by the percentage change in price

price floor a legal minimum on the price at which a good can be sold

principal a person for whom another person, called the agent, performs some act

prisoners' dilemma a particular "game" between two captured prisoners that illustrates why cooperation is difficult to maintain even when it is mutually beneficial

private goods goods that are both excludable and rival in consumption

private saving the income that households have left after paying for taxes and consumption

producer price index (PPI) a measure of the cost of a basket of goods and services sold by domestic firms

producer surplus the amount a seller is paid for a good minus the seller's cost of providing it

production function the relationship between the quantity of inputs used to make a good and the quantity of output of that good

production possibilities frontier a graph that shows the combinations of output that the economy can possibly produce with the available factors of production and production technology

productivity the quantity of goods and services produced from each unit of labor

profit total revenue minus total cost

progressive tax a tax for which taxpayers with high incomes pay a larger fraction of their income than do taxpayers with low incomes

property rights the ability of an individual to own and exercise control over scarce resources

proportional tax a tax for which taxpayers at all income levels pay the same fraction of income

public goods goods that are neither excludable nor rival in consumption

public saving the tax revenue that the government has left after paying for its spending

purchasing-power parity a theory of exchange rates that says a unit of any given currency should be able to buy the same quantity of goods in all countries

Q

quantity demanded the amount of a good that buyers are willing and able to purchase

quantity equation the equation $M \times V = P \times Y$, which relates the quantity of money, the velocity of money, and the dollar value of the economy's output of goods and services

quantity supplied the amount of a good that sellers are willing and able to sell

quantity theory of money a theory asserting that the quantity of money available determines the price level and that the growth rate in the quantity of money available determines the inflation rate

R

random walk the path of a variable whose changes are impossible to predict

rational expectations the theory that people optimally use all the information they have, including information about government policies, when forecasting the future

rational people people who systematically and purposefully do the best they can to achieve their objectives

real exchange rate the rate at which a person can trade the goods and services of one country for the goods and services of another

real GDP the production of goods and services valued at constant prices

real interest rate the interest rate corrected for the effects of inflation

real variables variables measured in physical units

recession a period of declining real incomes and rising unemployment

regressive tax a tax for which taxpayers with high incomes pay a smaller fraction of their income than do taxpayers with low incomes

reserve ratio the fraction of deposits that banks hold as reserves

reserve requirements regulations on the minimum amount of reserves that banks must hold against deposits

reserves deposits that banks have received but have not loaned out

risk aversion a dislike of uncertainty

rivalry in consumption the property of a good whereby one person's use diminishes other people's use

S

sacrifice ratio the number of percentage points of annual output lost in the process of reducing inflation by 1 percentage point

scarcity the limited nature of society's resources

screening an action taken by an uninformed party to induce an informed party to reveal information

shoeleather costs the resources wasted when inflation encourages people to reduce their money holdings

shortage a situation in which the quantity demanded is greater than the quantity supplied

signaling an action taken by an informed party to reveal private information to an uninformed party

social insurance government policy aimed at protecting people against the risk of adverse events

stagflation a period of falling output and rising prices

statistical discrimination discrimination that arises because an irrelevant but observable personal characteristic is correlated with a relevant but unobservable attribute

stock a claim to partial ownership in a firm

store of value an item that people can use to transfer purchasing power from the present to the future

strike the organized withdrawal of labor from a firm by a union

structural unemployment unemployment that results because the number of jobs available in some labor markets is insufficient to provide a job for everyone who wants one

substitutes two goods for which an increase in the price of one leads to an increase in the demand for the other

substitution effect the change in consumption that results when a price change moves the consumer along a given indifference curve to a point with a new marginal rate of substitution

sunk cost a cost that has already been committed and cannot be recovered

supply curve a graph of the relationship between the price of a good and the quantity supplied

supply schedule a table that shows the relationship between the price of a good and the quantity supplied

supply shock an event that directly alters firms' costs and prices, shifting the economy's aggregate-supply curve and thus the Phillips curve

surplus a situation in which the quantity supplied is greater than quantity demanded

T

tariff a tax on goods produced abroad and sold domestically

tax incidence the manner in which the burden of a tax is shared among participants in a market

technological knowledge society's understanding of the best ways to produce goods and services

theory of liquidity preference Keynes's theory that the interest rate adjusts to bring money supply and money demand into balance

total cost the market value of the inputs a firm uses in production

total revenue the amount paid by buyers and received by the sellers of a good, calculated as the price of the good times the quantity sold

trade balance the value of a nation's exports minus the value of its imports; also called net exports

trade deficit an excess of imports over exports

trade policy government policy that directly influences the quantity of goods and services that a country imports or exports

trade surplus an excess of exports over imports

Tragedy of the Commons a parable that illustrates why common resources are used more than is desirable from the standpoint of society as a whole

transaction costs the costs that parties incur during the process of agreeing to and following through on a bargain

U

unemployment insurance a government program that partially protects the incomes of workers who become unemployed

unemployment rate the percentage of the labor force that is unemployed

union a worker association that bargains with employers over wages and working conditions

unit of account the yardstick people use to post prices and record debts

utilitarianism the political philosophy according to which the government should choose policies to maximize the total utility of everyone in society

utility a measure of satisfaction

V

value of the marginal product the marginal product of an input times the price of the output

variable costs costs that vary with the quantity of output produced

velocity of money the rate at which money changes hands

vertical equity the idea that taxpayers with a greater ability to pay taxes should pay larger amounts

W

welfare government programs that supplement the incomes of the needy

welfare economics the study of how the allocation of resources affects economic well-being

willingness to pay the maximum amount that a buyer will pay for a good

world price the price of a good that prevails in the world market for that good

Index

Page numbers in *boldface* refer to pages where key terms are defined.

B

C

D

E

F

G

H

I

M

N

P

Q

R

S

U

V

W

X

Y

Z