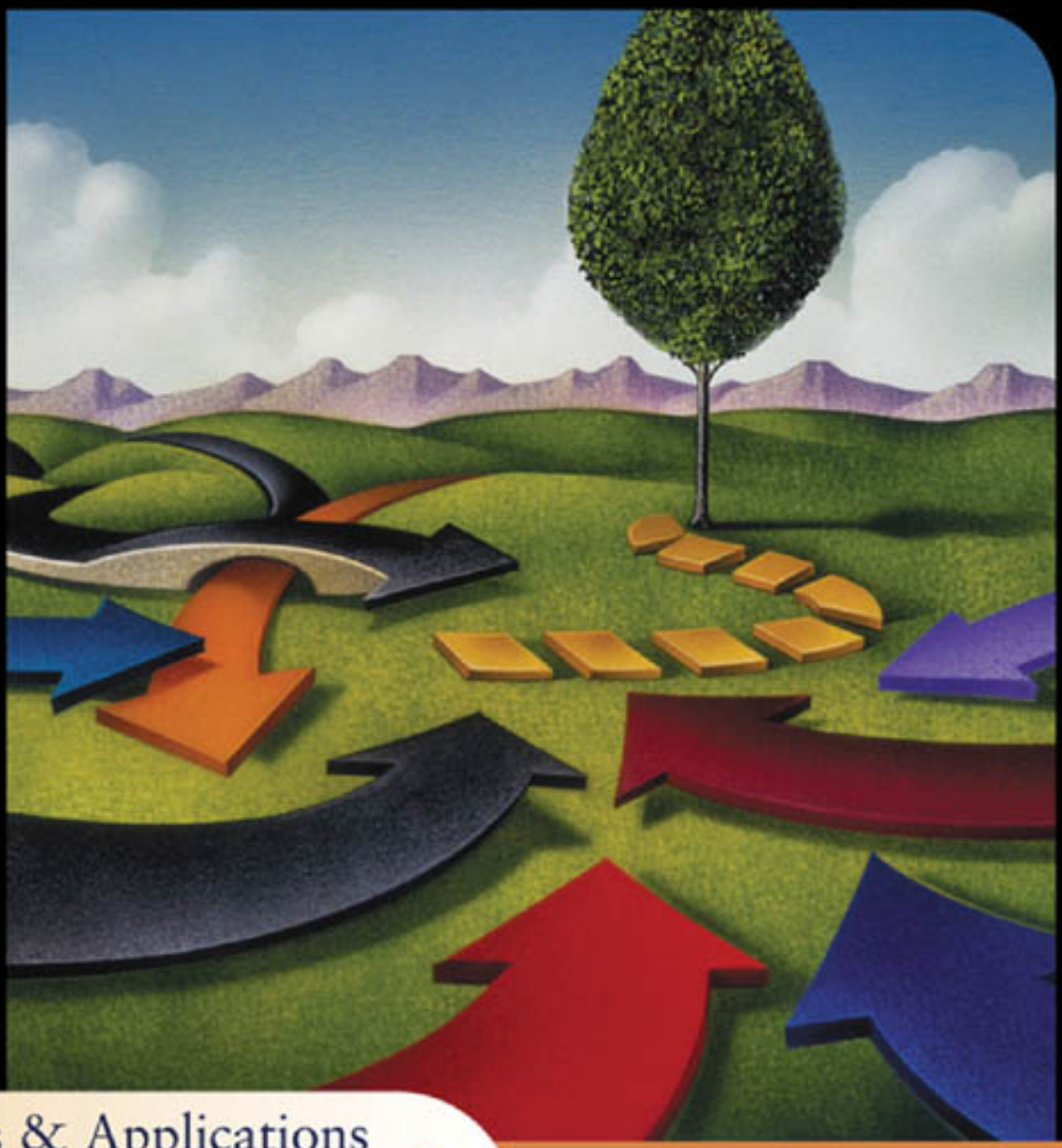
MICROECONOMICS



Principles & Applications 5E

Robert E. **HALL**

&

Marc **LIEBERMAN**

MICROECONOMICS

Principles & Applications 5E

Robert E. **HALL**

Department of Economics, Stanford University

Marc **LIEBERMAN**

Department of Economics, New York University

SOUTH-WESTERN
CENGAGE Learning™

Australia • Brazil • Japan • Korea • Mexico • Singapore • Spain • United Kingdom • United States

SOUTH-WESTERN
CENGAGE Learning™

Microeconomics: Principles & Applications, 5th Edition

Robert E. Hall

Marc Lieberman

Vice President of Editorial, Business: Jack W. Calhoun

Publisher: Joe Sabatino

Executive Editor: Michael Worls

Senior Developmental Editor: Susanna C. Smart

Senior Marketing Manager: John Carey

Senior Marketing Communications Manager: Sarah Greber

Content Project Manager: Corey Geissler

Media Editor: Deepak Kumar

Editorial Assistant: Lena Mortis

Manufacturing Coordinator: Sandra Milewski

Production Service: S4Carlisle Publishing Services

Senior Art Director: Michelle Kunkler

Internal Designer: Imbue Design

Cover Designer: Imbue Design

Cover Image: Copyright Theo Rudnak/images.com

Photography Permissions Manager: Deanna Ettinger

Text Permissions Manager: Mardell Glinski Schultz

Library of Congress Control Number: 2009933897

ISBN-10: 1-4390-3897-X
ISBN-13: 978-1-4390-3897-0

South-Western Cengage Learning
5191 Natorp Boulevard
Mason, OH 45040
USA

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

For your course and learning solutions, visit **www.cengage.com**
Purchase any of our products at your local college store or at our preferred online store **www.ichapters.com**

Printed in the United States of America
1 2 3 4 5 6 7 13 12 11 10 09

BRIEF CONTENTS

CONTENTS

Part V: Labor, Capital, and Financial Markets

Part VI: Efficiency, Government, and the Global Economy

Microeconomics: Principles and Applications is about economic principles and how economists use them to understand the world. It was conceived, written, and for the fifth edition, substantially revised to help your students focus on those basic principles and applications. We originally decided to write this book, because we thought that existing texts tended to fall into one of three categories. In the first category are the encyclopedias—the heavy tomes with a section or a paragraph on every topic or subtopic you might possibly want to present to your students. These books are often useful as reference tools. But because they cover so many topics—many of them superficially—the central themes and ideas can be lost in the shuffle. The second type of text we call the "scrapbook." In an effort to elevate student interest, these books insert multicolored boxes, news clippings, interviews, cartoons, and whatever else they can find to jolt the reader on each page. While these special features are often entertaining, there is a trade-off: These books sacrifice a logical, focused presentation of the material. Once again, the central themes and ideas are often lost. Finally, a third type of text, perhaps in response to the first two, tries to do less in every area—a *lot* less. But instead of just omitting extraneous or inessential details, these texts often throw out key ideas, models, and concepts. Students who use these books may think that economics is overly simplified and unrealistic. After the course, they may be less prepared to go on in the field, or to think about the economy on their own.

A Distinctive Approach

Our approach is very different. We believe that the best way to teach principles is to present economics as a coherent, unified subject. This does not happen automatically. On the contrary, principles students often miss the unity of what we call "the economic way of thinking." For example, they are likely to see the analysis of goods markets, labor markets, and financial markets as entirely different phenomena, rather than as a repeated application of the same methodology with a new twist here and there. So the principles course appears to be just "one thing after another," rather than the coherent presentation we aim for.

Careful Focus

Because we have avoided encyclopedic complexity, we have had to think hard about what topics are most important. As you will see:

We avoid nonessential material

When we believed a topic was not essential to a basic understanding of economics, we left it out. However, we have strived to include core material to *support* an instructor who wants to present special topics in class. So, for example, we do not have separate chapters on environmental economics, agricultural economics, urban economics, health care economics, or comparative systems. But instructors should find in the text a good foundation for building any of these areas—and many others—into their course. And we have included examples from each of these areas as *applications* of core theory where appropriate throughout the text.

We avoid distracting features

This text does not have interviews, news clippings, or boxed inserts with only distant connections to the core material. The features your students *will* find in our book are there to help them understand and apply economic theory itself, and to help them avoid common mistakes in applying the theory (the Dangerous Curves feature).

We explain difficult concepts patiently

By freeing ourselves from the obligation to introduce every possible topic in economics, we can explain the topics we *do* cover more thoroughly and patiently. We lead students, step-by-step, through each aspect of the theory, through each graph, and through each numerical example. In developing this book, we asked other experienced teachers to tell us which aspects of economic theory were hardest for their students to learn, and we have paid special attention to the trouble spots.

We use concrete examples

Students learn best when they see how economics can explain the world around them. Whenever possible, we develop the theory using real-world examples. You will find numerous references to real-world corporations and government policies throughout the text. When we employ hypothetical examples because they illustrate the theory more clearly, we try to make them realistic. In addition, almost every chapter ends with a thorough, extended application (the "Using the Theory" section) focusing on an interesting real-world issue.

Features That Reinforce

To help students see economics as a coherent whole, and to reinforce its usefulness, we have included some important features in this book.

The Three-Step Process

Most economists, when approaching a problem, begin by thinking about buyers and sellers, and the markets in which they come together to trade. They move on to characterize a market equilibrium, and then give their model a workout in a comparative statics exercise. To understand what economics is about, students need to understand this process and see it in action in different contexts. To help them do so, we have identified and stressed a "three-step process" that economists use in analyzing problems. The three key steps are:

1. **Characterize the Market.** Decide which market or markets best suit the problem being analyzed, and identify the decision makers (buyers and sellers) who interact there.
2. **Find the Equilibrium.** Describe the conditions necessary for equilibrium in the market, and a method for determining that equilibrium.
3. **Determine What Happens When Things Change.** Explore how events or government policies change the market equilibrium.

The steps themselves are introduced toward the end of Chapter 3. Thereafter, the content of most chapters is organized around this three-step process. We believe this helps students learn how to think like economists, and in a very natural way. And they come to see economics as a unified whole, rather than as a series of disconnected ideas.

Dangerous Curves

Anyone who teaches economics for a while learns that, semester after semester, students tend to make the same familiar errors. In class, in office hours, and on exams, students seem pulled, as if by gravity, toward certain logical pitfalls in thinking about, and using, economic theory. We've discovered in our own classrooms that merely explaining the theory properly isn't enough; the most common errors need to be *confronted,* and the student needs to be shown *specifically* why a particular logical path is incorrect. This was the genesis of our "Dangerous Curves" feature—boxes that anticipate the most common traps and warn students just when they are most likely to fall victim to them. We've been delighted to hear from instructors how effective this feature has been in overcoming the most common points of confusion for their students.

Using the Theory

This text is full of applications that are woven throughout the narrative. In addition, almost every chapter ends with an extended application ("Using the Theory") that pulls together several of the tools learned in that chapter. These are not news clippings or world events that relate only tangentially to the material. Rather, they are step-by-step presentations that help students see how the tools of economics can explain things about the world—things that would be difficult to explain without those tools.

Content Innovations

In addition to the special features just described, you will find some important differences from other texts in topical approach and arrangement. These, too, are designed to make the theory stand out more clearly, and to make learning easier. These are not pedagogical experiments, nor are they innovation for the sake of innovation. The differences you will find in this text are the product of years of classroom experience.

Scarcity, Choice, and Economic Systems (Chapter 2)

This early chapter, while covering standard material such as opportunity cost, also introduces some central concepts much earlier than other texts. Most importantly, it introduces the concept of *comparative advantage,* and the basic principle of *specialization and exchange.* We have placed them at the front of our book, because we believe they provide important building blocks for much that comes later. For example, comparative advantage and specialization *within* the firm help explain economies of scale (Chapter 6). International trade (Chapter 16) can be seen as a special application of these principles, extending them to trade between nations.

How Firms Make Decisions: Profit Maximization (Chapter 8)

Many texts introduce the theory of the firm using the perfectly competitive model first. While this has logical appeal to economists, we believe it is an unfortunate choice for students encountering this material for the first time. Leading with perfect competition forces students to simultaneously master the logic of profit maximization *and* the details of a rather counter-intuitive kind of market at the same time. Students quite naturally think of firms as facing *downward*-sloping demand curves—not horizontal ones. We have found that they have an easier time learning the theory of the firm with the more familiar, downward-sloping demand curve. Further, by treating the theory of the firm in a separate chapter, *before* perfect competition, we can separate concepts that apply in *all* market structures (the shapes of marginal cost and average cost curves, the MC and MR approach to profit maximization, the shut-down rule, etc.), from concepts that are unique to perfect competition (horizontal demand curve, marginal revenue the same as price, etc.). This avoids confusion later on.

Monopolistic Competition and Oligopoly (Chapter 11)

Two features of our treatment are worth noting. First, we emphasize advertising, a key feature of both of these types of

markets. Students are very interested in advertising and how firms make decisions about it. Second, we have omitted older theories of oligopoly that raised more questions than they answered, such as the kinked demand curve model. Our treatment of oligopoly is strictly game theoretic, but we have taken great care to keep it simple and clear. Here, as always, we provide the important tools to *support* instructors who want to take game theory further, without forcing every instructor to do so by including too much.

Capital and Financial Markets (Chapter 13)

This chapter focuses on the common theme of these subjects: the present value of future income. Moreover, it provides simple, principles-level analyses of the stock and bond markets—something that students are hungry for but that many principles textbooks neglect.

Description versus Assessment (Chapters 9–11 and 14–15)

In treating product market structures, most texts switch back and forth between the *description and analysis* of different markets on the one hand and their *efficiency properties* on the other. Our book deals with description and analysis first, and only then discusses efficiency, in two comprehensively chapters. The first of these (Chapter 14) covers the concept and measurement of economic efficiency, using Pareto improvements as well as consumer and producer surplus. The second (Chapter 15) deals with market failures and government's role in economic efficiency. This arrangement of the material permits instructors to focus on *description and prediction* when first teaching about market structures—a full plate, in our experience. Second, two chapters devoted to efficiency allows a more comprehensive treatment of the topic than we have seen elsewhere. Finally, our approach—in which students learn about efficiency *after* they have mastered the four market structures—allows them to study efficiency with the perspective needed to really understand it.

Comparative Advantage and the Gains from International Trade (Chapter 16)

We've found that international trade is best understood through clear numerical examples, and we've developed them carefully in this chapter. We also try to bridge the gap between the economics and politics of international trade with a systematic discussion of winners and losers.

Organizational Flexibility

We have arranged the contents of each chapter, and the table of contents as a whole, according to our recommended order of presentation. But we have also built in flexibility.

- Chapter 6 develops consumer theory with both marginal utility and (in an appendix) indifference curves, allowing you to present either method in class. (Instructors will find it even easier to make their choice in this edition—see following.)
- If you wish to highlight international trade or present comparative advantage earlier in the course, you could assign Chapter 16 immediately following Chapter 3.
- If you wish to introduce consumer and producer surplus earlier in the course, all of Chapter 14 can be assigned after Chapter 9. And if you feel strongly that economic efficiency should be interwoven bit-by-bit with the chapters on market structure, Chapter 14 can be easily broken into parts. The relevant sections can then be assigned separately with Chapters 3, 4, 9, and 10.

Finally, we have included only those chapters that we thought were both essential and teachable in a one-semester course. But not everyone will agree about what is essential. While we—as authors—cringe at the thought of a chapter being omitted in the interest of time, we have allowed for that possibility. Nothing in Chapter 12 (labor markets), Chapter 13 (capital and financial markets), Chapter 15 (government's role in economic efficiency), or Chapter 16 (international trade) is essential to any of the other chapters in the book. Skipping any of these should not cause continuity problems.

New to the Fifth Edition

The fifth edition is our most significant revision yet. This will not surprise anyone who was teaching an economics principles course during or after September 2008, when the financial crisis hit its peak. While teaching at the time, we had the daily task of integrating the flood of unprecedented events into the course. When the semester was over, the two of us thought long and hard about what worked, what didn't, and how the course should respond to the changes we had seen.

We wanted to be able to discuss recent events and draw out their long-lasting lessons and challenges. We knew this would require adding some new concepts and tools. But we were mindful that this is a *first* course in economics and did not want to migrate into areas that we could not fully explain at the principles level. In our discussions, we kept coming back to the same place: that by adding two new core concepts, we could open up a myriad of other doors to understanding recent economic events. Both of these concepts are introduced in Chapter 4 (Working with Supply and Demand).

Two New Concepts

The first new concept we've introduced in this new edition is *leverage*. While leverage is at the heart of the recent economic turmoil, it has not been part of the traditional principles pedagogy. We've introduced it in a simple, intuitive way in the body of Chapter 4. We then delve a bit deeper in the short appendix

to that chapter, which explains the concept of owners' equity (in a home), and presents a simple *leverage ratio* that students can work with. Teaching this concept not only creates an early, fresh connection between the classroom and current policy debates but also lays the foundation for later applications in the text. For example, students will see how leverage contributed to the recent housing boom and bust (in Chapter 4) and moral hazard in financial institutions (Chapter 15).

The second new core concept is how supply and demand can be used for *stock variables,* and not just flow variables. While this idea was present in prior editions, it came late in the text and was not fully established as a key concept. We've long wanted to introduce the stock-flow distinction earlier, and more carefully, so we could analyze the market for the housing *stock* with supply and demand. But we never thought this was essential . . . until now.

As you'll see in Chapter 4, treating housing as a stock variable opens another door to understanding the recent housing boom and bust. We also believe that teaching the stock-flow distinction early—with the rather intuitive case of housing—makes it easier to think about stock variables later in other contexts (such as financial markets, covered in Chapter 13).

Other Key Changes

Our overall approach, and the sequence of the material, will be mostly familiar to those who've used past editions. But we wanted to highlight some other pedagogical changes, in addition to the new concepts (discussed earlier) in Chapter 4.

In this edition, our biggest change (at the request of many instructors) is the new, simplified treatment of labor markets. The previous two chapters (one on labor markets and one on income inequality) are now combined into the single Chapter 12 (Labor Markets). The development of the labor demand curve is streamlined, so you can get to interesting applications (such as wage inequality) with less delay. (Those who liked the prior approach to labor demand will find it in the appendix to that chapter.)

Two other pedagogical changes we should note are the shift of the section on opportunity cost from Chapter 2 to Chapter 1 and an earlier introduction of international trade (within the discussion of comparative advantage in Chapter 2).

New Applications

There are dozens of new applications in this edition—some woven into the narrative, others as new or substantially revised "Using the Theory" sections, where the analysis is more extensive. The *entirely* new "Using the Theory" sections are:

- "The Oil Price Spike of 2007–2008" (Chapter 3)
- "The Housing Boom and Bust of 1997–2008" (Chapter 4)
- "Monopoly Pricing and Parallel Trade in Pharmaceuticals" (Chapter 10)
- "Moral Hazard in the Financial Crisis of 2008" (Chapter 15)

Within the chapters, there are new or substantially expanded sections on the role of elasticity in explaining commodity price fluctuations (Chapter 5), how the insights of behavioral economics have affected government policy (Chapter 6), interest rate spreads in financial markets (Chapter 13), what asset bubbles imply about efficient markets theory (Chapter 13), how information asymmetry affects the health insurance market (Chapter 15), and more.

Teaching and Learning Aids

To help you present the most interesting principles courses possible, we have created an extensive set of supplementary items. Many of them can be downloaded from the Hall/Lieberman Web site **www.cengage.com/economics/hall.** The list includes the following items.

For the Instructor

- The ***Instructor's Manual*** is revised by Natalija Novta, New York University, and Jeff Johnson, Sullivan University. The manual provides chapter outlines, teaching ideas, experiential exercises for many chapters, suggested answers to the end-of-chapter review questions, and solutions to all end-of-chapter problems.
- ***Instructor's Resource CD-ROM.*** This easy-to-use CD allows quick access to instructor ancillaries from your desktop. It also allows you to review, edit, and copy exactly the material you need. Or, you may choose to go to ***Instructor Resources*** on the *Product Support Web Site.*
- ***Instructor Resources*** on the *Product Support Web Site.* This site at **www.cengage.com/economics/hall** features the essential resources for instructors, password-protected, in downloadable format: the *Instructor's Manual* in Word, the test banks in Word, and PowerPoint lecture and exhibit slides.
- ***Microeconomics Test Bank.*** The micro test bank is revised by Toni Weiss of Tulane University. It contains more than 2,500 multiple-choice questions, arranged according to chapter headings and subheadings, making it easy to find the material needed to construct examinations.
- ***ExamView Computerized Testing Software.*** ExamView is an easy-to-use test creation package compatible with both Microsoft Windows and Macintosh client software, and contains all of the questions in all of the printed test banks. You can select questions by previewing them on the screen, selecting them by number, or selecting them randomly. Questions, instructions, and answers can be edited, and new questions can easily be added. You can also administer quizzes online over the Internet, through a local area network (LAN), or through a wide area network (WAN).
- ***PowerPoint Lecture and Exhibit Slides.*** Available on the Web site and the IRCD, the PowerPoint presentations are revised by Andreea Chiritescu, Eastern Illinois University

and consist of speaking points in chapter outline format, accompanied by numerous key graphs and tables from the main text, many with animation to show movement of demand and supply curves. A separate set of slides with exhibits only is also available.

- ***TextChoice.*** TextChoice is a custom format of Cengage Learning's online digital content. TextChoice provides the fastest, easiest way for you to create your own learning materials. You may select content from hundreds of best-selling titles, choose material from our numerous databases, and add your own material. Contact your South-Western/Thomson sales representative for more information at **www.cengagecustom.com.**

- ***WebTutor Toolbox.*** WebTutor ToolBox provides instructors with links to content from the book companion Web site. It also provides rich communication tools to instructors and students, including a course calendar, chat, and e-mail. For more information about the WebTutor products, please contact your local Cengage sales representative.

For the Student

- ***Hall/Lieberman EconCentral*** Multiple resources for learning and reinforcing principles concepts are now available in one place!

EconCentral is your one-stop shop for the learning tools and activities to help students succeed. Available for a minimal additional cost, EconCentral equips learners with a portal to a wealth of resources that help them both study and apply economic concepts. As they read and study the chapters, students can access video tutorials with *Ask the Instructor Videos.* They can review with *Flash Cards* and the *Graphing Workshop* as well as check their understanding of the chapter with *interactive quizzing.*

Ready to help students apply chapter concepts to the real world? EconCentral gives you ABC News videos, EconNews articles, Economic Debates, Links to Economic Data, and more. All the study and application resources in EconCentral are organized by chapter to help your students get the most from *Microeconomics: Principles and Applications,* fifth edition, and from your lectures.

Visit **www.cengage.com/economics/econcentral** and select **Hall/Lieberman 5e** to see the study options available!

- ***Global Economic Watch.*** A global economic crisis need not be a teaching crisis.

Students can now learn economic concepts through examples and applications using the most current information on the global economic situation. The Global Economic Resource Center Includes:

- A 32-page eBook that gives a general overview of the events that led up to the current situation, written by Mike Brandl from the University of Texas, Austin
- A Blog and Community Site updated daily by an economic journalist and designed to allow you and your colleagues to share thoughts, ideas and resources
- Thousands of articles from leading journals, news services, magazines, and newspapers from around the world revised 4 times a day and searchable by topic and key term
- Student and instructor resources such as PowerPoint® decks, podcasts, and videos
- Assessment materials allowing you to ensure student accountability

This resource can be bundled at no charge with this textbook. Visit **www.cengage.com/thewatch** for more information.

- ***Tomlinson Economics Videos.*** **"Like Office Hours 24/7"** Award winning teacher, actor, and professional communicator, Steven Tomlinson (Ph.D. Economics, Stanford) walks students through all of the topics covered in principles of economics in an online video format. Segments are organized to follow the organization of the Hall/Lieberman text and most videos include class notes that students can download and quizzes for students to test their understanding which can be sent to the professor if required. Find out more at **www.cengage.com/economics/tomlinson.**

- ***Aplia*** Founded in 2000 by economist and Stanford Professor Paul Romer, Aplia is dedicated to improving learning by increasing student effort and engagement. The most successful online product in economics by far, Aplia has been used by more than 1,000,000 students at more than 850 institutions. Visit **www.aplia.com/cengage** for more details. For help, questions, or a live demonstration, please contact Aplia at **support@aplia.com.**

- ***The Active Learning Guide*** provides numerous exercises and self-tests for problem-solving practice. It is a valuable tool for helping students strengthen their knowledge of economics, and includes a sample multiple-choice final exam, with answers and explanations. It is now available both in print and online.

- ***The Hall/Lieberman Web site*** (**www.cengage.com/economics/hall**). The text Web site contains a wealth of useful teaching and learning resources. Important features available at the Web site include interactive quizzes with feedback on answers—completed quizzes can be e-mailed directly to the instructor; a sample chapter from the *Active Learning Guide;* and links to other economic resources.

- ***Economics for Life*** Economics comes alive**!** Let Bruce Madariaga's clear, concise, nontechnical style transform your economic study from an academic exercise into an exciting journey. By using real-world applications, stories, misconceptions, mysteries, amazing facts, and statistics, ***Economics for Life*** gives you the foundation to understand how people make economic decisions every day. This softbound book (5½ × 8½) may be bundled at no extra charge with this text. Contact your sales representative for additional information.

Acknowledgments

Our greatest debt is to the many reviewers who carefully read the book and provided numerous suggestions for improvements. While we could not incorporate all their ideas, we did carefully evaluate each one of them. We are especially grateful to the participants in our survey who helped us with the revision for this fifth edition. To all of these people, we are most grateful:

Sindy Abadie	Southwest Tennessee Community College
Eric Abrams	Hawaii Pacific University
Ljubisa Adamovich	Florida State University
Brian A'Hearn	Franklin and Marshall College
Ali Akarca	University of Illinois, Chicago
Rashid Al-Hmoud	Texas Tech University
David Aschauer	Bates College
Richard Ballman	Augustana College
James T. Bang	Virginia Military Institute
Chris Barnett	Gannon University
Parantap Basu	Fordham University
Tibor Besedes	Rutgers University
Gautam Bhattacharya	University of Kansas
Margot B. Biery Blackburne	Tarrant County College Edward Sam Houston State University
Sylvain Boko	Wake Forest University
Barry Bomboy	J. Sargeant Reynolds Community College
John L. Brassel	Southwest Tennessee Community College
Mark Buenafe	Arizona State University
Steven Call	Metropolitan State College
Kevin Carey	American University
Siddharth Chandra	University of Pittsburgh
Steven Cobb	Xavier University
Christina Coles	Johnson & Wales University
Maria Salome E. Davis	Indian River State College
Dennis Debrecht	Carroll College
Selahattin Dibooglu	University of St. Louis, Missouri
Arthur M. Diamond Jr.,	University of Nebraska, Omaha
James E. Dietz	California State University, Fullerton
Khosrow Doroodian	Ohio University
John Duffy	University of Pittsburgh
Debra S. Dwyer	SUNY Stony Brook
Stephen Erfle	Dickinson College
Barry Falk	Iowa State University
James Falter	Mount Marty College
Sasan Fayazmanesh	California State University, Fresno
Lehman B. Fletcher	Iowa State University
Richard Fowles	University of Utah
Mark Frascatore	Clarkson College
Mark Funk	University of Arkansas at Little Rock
James R. Gale	Michigan Technological University
Sarmila Ghosh	University of Scranton
Satyajit Ghosh	University of Scranton
Michelle Gietz	Southwest Tennessee Community College
Scott Gilbert	Southern Illinois University, Carbondale
Susan Glanz	St. John's University
Michael Gootzeit	University of Memphis
John Gregor	Washington and Jefferson University
Jeff Gropp	DePauw University
Arunee C. Grow	Mesa Community College
Ali Gungoraydinoglu	The University of Mississippi
Rik Hafer	Southern Illinois University
Robert Herman	Nassau Community College
Michael Heslop	Northern Virginia Community College
Paul Hettler	California University of Pennsylvania
Roger Hewett	Drake University
Andrew Hildreth	University of California, Berkeley
Nathan Himelstein	Essex County College
Stella Hofrenning	Augsburg College
Shahruz Hohtadi	Suffolk University
Daniel Horton	Cleveland State
Jack W. Hou	California State University–Long Beach
Thomas Husted	American University
Jeffrey Johnson	Sullivan University
James Jozefowicz	Indiana University of Pennsylvania
Jack Julian	Indiana University of Pennsylvania
Farrokh Kahnamoui	Western Washington University
Leland Kempe	California State University, Fresno
Jacqueline Khorassani	Marietta College
Philip King	San Francisco State University
Frederic R. Kolb	University of Wisconsin, Eau Claire
Kate Krause	University of New Mexico
Brent Kreider	Iowa State University
Viju Kulkarni	San Diego State University
Nazma Latif-Zaman	Providence College
Teresa Laughlin	Palomar College
Bruce Madariaga	Montgomery College
Judith Mann	University of California, San Diego
Thomas McCaleb	Florida State University

Mark McCleod Virginia Tech University
Michael McGuire University of the Incarnate Word
Steve McQueen Barstow Community College
William R. Melick Kenyon College
Arsen Melkumian West Virginia University
Frank Mixon University of Southern Mississippi
Shahruz Mohtadi Suffolk University
Gary Mongiovi St. John's University
Joseph R. Morris Broward Community College-South Campus
Paul G. Munyon Grinnell College
Rebecca Neumann University of Wisconsin, Milwaukee
Chris Niggle University of Redlands
Emmanuel Nnadozie Truman State University
Nick Noble Miami University, Ohio
Farrokh Nourzad Marquette University
Lee Ohanian University of California, Los Angeles
Jim Palmieri Simpson College
Zaohong Pan Western Connecticut State University
Yvon Pho American University
Gregg Pratt Mesa Community College
Teresa Riley Youngstown State University
William Rosen Cornell University
Alannah Rosenberg Saddleback College
Thomas Pogue University of Iowa
Scott Redenius Bryn Mawr College
Jeff Rubin Rutgers University
Rose Rubin University of Memphis
Thomas Sadler Pace University
Jonathan Sandy University of San Diego
Ramazan Sari Texas Tech University
Ghosh Sarmila University of Scranton
Edward Scahill University of Scranton
Robert F. Schlack Carthage College
Pamela M. Schmitt U.S. Naval Academy
Mary Schranz University of Wisconsin, Madison
Gerald Scott Florida Atlantic University
Peter M Shaw Tidewater Community College
Alden Shiers California Polytechnic State University
Kevin Siqueira Clarkson University
William Doyle Smith University of Texas, El Paso Kevin
Sontheimer University of Pittsburgh
Mark Steckbeck Campbell University
Richard Steinberg Indiana University, Purdue University, Indianapolis
Martha Stuffler Irvine Valley College
Mohammad Syed Miles College
Manjuri Talukdar Northern Illinois University
Kiril Tochkov Binghamton University
John Vahaly University of Louisville
Mikayel Vardanyan Oregon State University
Thomas Watkins Eastern Kentucky University
Robert Whaples Wake Forest University
Glen Whitman California State University, Northridge
Michael F. Williams University of St. Thomas
Melissa Wiseman Houston Baptist University Dirk
Yandell University of San Diego
Petr Zemcik Southern Illinois University, Carbondale

We also wish to acknowledge the talented and dedicated group of instructors who helped put together a supplementary package that is second to none. Geoffrey A. Jehle of Vassar College co-wrote the *Active Learning Guide* for several editions and helped make it user-friendly and *active*. Natalija Novta, New York University, and Jeff Johnson, Sullivan University, revised the *Instructor's Manual,* and the test bank was carefully revised by Toni Weiss, Tulane University. Finally, special thanks go to Dennis Hanseman, who was our development editor for the first edition; his insights and ideas are still present in this fifth edition, and his continued assistance has proved invaluable.

The beautiful book you are holding would not exist except for the hard work of a talented team of professionals. Book production was overseen by Corey Geissler, Content Project Manager at Cengage South-Western and undertaken by Cindy Sweeney, Project Editor at S4Carlisle Publishing Services. Corey and Cindy showed remarkable patience, as well as an unflagging concern for quality throughout the process. We couldn't have asked for better production partners. Several NYU students helped to locate and fix the few remaining errors: Carla Bernal, Eric Branting, Junli Chen, Joseph Colucci, Jason Flamendorf, Anna Gaysynsky, and Kyle Kozman, The overall look of the book and cover was planned by Michelle Kunkler and executed by Lisa Albonetti. Deanna Ettinger managed the photo program, and Sandee Milewski made all the pieces come together in her role as Manufacturing Coordinator. We are especially grateful for the hard work of the dedicated and professional South-Western editorial, marketing, and sales teams. Mike Worls, Senior Acquisitions Editor, has once again shepherded this text through publication with remarkable skill and devotion. John Carey, Senior Marketing Manager, has done a first-rate job getting the message out to instructors and sales reps. Susan Smart, who has been Senior Development Editor on several editions, once again delved into every chapter and contributed to their improvement. She showed even more than her usual patience, flexibility, and skill in managing both content and authors—and didn't flinch when it became clear that this revision would

require much more work than previous editions. Deepak Kumar, Media Editor, has put together a wonderful package of media tools, and the Cengage South-Western sales representatives have been extremely persuasive advocates for the book. We sincerely appreciate all their efforts!

Finally, we want to acknowledge the amazing team at Aplia, who modified existing Aplia problems, and wrote new ones, to create the closest possible fit with our textbook. In particular, Paul Romer (CEO and Founder), Kristen Ford (Managing Editor), Chris Makler (Senior Economist), and Kasie Jean (Senior Content Developer) showed remarkable skill, knowledge, and patience in working on content.

A Request

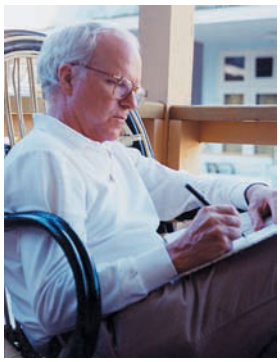
Although we have worked hard on the five editions of this book, we know there is always room for further improvement. For that, our fellow users are indispensable. We invite your comments and suggestions wholeheartedly. We especially welcome your suggestions for additional "Using the Theory" sections and Dangerous Curves. You may send your comments to either of us, care of South-Western.

Robert E. Hall

Marc Lieberman

About the Authors

Robert E. Hall



Robert E. Hall is a prominent applied economist. He is the Robert and Carole McNeil Professor of Economics at Stanford University and Senior Fellow at Stanford's Hoover Institution where he conducts research on inflation, unemployment, taxation, monetary policy, and the economics of high technology. He received his Ph.D. from MIT and has taught there as well as at the University of California, Berkeley. Hall is director of the research program on Economic Fluctuations of the National Bureau of Economic Research, and chairman of the Bureau's Committee on Business Cycle Dating, which maintains the semiofficial chronology of the U.S. business cycle. He has published numerous monographs and articles in scholarly journals, in addition to co-authoring this well-known intermediate text. Hall has advised the Treasury Department and the Federal Reserve Board on national economic policy and has testified on numerous occasions before congressional committees. Hall is President-elect of the American Economic Association and will serve as President in 2010. He presented the Ely Lecture to the Association in 2001 and served as Vice President in 2005.

Marc Lieberman

Marc Lieberman is Clinical Professor of Economics at New York University. He received his Ph.D. from Princeton University. Lieberman has presented his extremely popular Principles of Economics course at Harvard, Vassar, the University of California at Santa Cruz, and the University of Hawaii, as well as at NYU. He has twice won NYU's Golden Dozen teaching award, and also the Economics Society Award for Excellence in Teaching. He is coeditor and contributor to *The Road to Capitalism: Economic Transformation in Eastern Europe and the Former Soviet Union*. Lieberman has consulted for the Bank of America and the Educational Testing Service. In his spare time, he is a professional screenwriter, and teaches screenwriting at NYU's School of Continuing and Professional Studies.

CHAPTER 1

What Is Economics?

Economics. The word conjures up all sorts of images: manic stock traders on Wall Street, an economic summit meeting in a European capital, a somber television news anchor announcing good or bad news about the economy. . . . You probably hear about economics several times each day. What exactly *is* economics?

First, economics is a *social science*. It seeks to explain something about *society*, just like other social sciences, such as psychology, sociology, and political science. But economics is different from these other social sciences because of *what* economists study and *how* they study it. Economists ask different questions, and they answer them using tools that other social scientists find rather exotic.

A good definition of economics, which stresses its differences from other social sciences, is the following:

> ***Economics** is the study of choice under conditions of scarcity.*

Economics The study of choice under conditions of scarcity.

This definition may appear strange to you. Where are the familiar words we ordinarily associate with economics: "money," "stocks and bonds," "prices," "budgets," . . . ? As you will soon see, economics deals with all of these things and more. But first, let's take a closer look at two important ideas in this definition: scarcity and choice.

Scarcity and Individual Choice

Think for a moment about your own life. Is there anything you don't have that you'd *like* to have? Anything you'd like *more* of? If your answer is "no," congratulations! You are well advanced on the path of Zen self-denial. The rest of us, however, feel the pinch of limits to our material standard of living. This simple truth is at the very core of economics. It can be restated this way: We all face the problem of **scarcity**.

Scarcity A situation in which the amount of something available is insufficient to satisfy the desire for it.

At first glance, it may seem that you suffer from an infinite variety of scarcities. There are so many things you might like to have right now—a larger room or apartment, a new car, more clothes . . . the list is endless. But a little reflection suggests that your limited ability to satisfy these desires is based on two other, more basic limitations: scarce *time* and scarce *spending power*.

> *As individuals, we face a scarcity of time and spending power. Given more of either, we could each have more of the goods and services that we desire.*

The scarcity of spending power is no doubt familiar to you. We've all wished for higher incomes so that we could afford to buy more of the things we want. But the scarcity of time is equally important. So many of the activities we enjoy—seeing movies, taking vacations, making phone calls—require time as well as money. Just as we have limited spending power, we also have a limited number of hours in each day to satisfy our desires.

Because of the scarcities of time and spending power, each of us is forced to make *choices*. We must allocate our scarce *time* to different activities: work, play, education, sleep, shopping, and more. We must allocate our scarce *spending power* among different goods and services: housing, food, furniture, travel, and many others. And each time we choose to buy something or do something, we also choose *not* to buy or do something else.

Economists study the choices we make as individuals, as well as their consequences. When some of the consequences are harmful, economists study what—if anything—the government can or should do about them.

For example, in the United States, as incomes have risen, more and more people have chosen to purchase automobiles. The result is increasing traffic jams in our major cities. The problem is even worse in rapidly developing countries. In China and India, for example, recent income growth and migration from rural to urban areas has led to an explosion of driving. Economists have come up with some creative ideas to reduce traffic congestion, while preserving individual choices about driving. A few cities have used these ideas, with some success, and more are considering them.

The Concept of Opportunity Cost

What does it cost you to go to the movies? If you answered 9 or 10 dollars because that is the price of a movie ticket, then you are leaving out a lot. Most of us are used to thinking of "cost" as the money we must pay for something. Certainly, the money we pay for goods or services is a *part* of its cost. But economics takes a broader view of costs. The true cost of any choice we make—buying a car, producing a computer, or even reading a book—is everything we must *give up* when we take that action. This cost is called the *opportunity cost* of the action, because we give up the opportunity to have other desirable things.

Opportunity cost What is given up when taking an action or making a choice.

*The **opportunity cost** of any choice is what we must forego when we make that choice.*

Opportunity cost is the most accurate and complete concept of cost—the one we should use when making our own decisions or analyzing the decisions of others.

Suppose, for example, it's 8 P.M. on a weeknight and you're spending a couple of hours reading this chapter. As authors, that thought makes us very happy. We know there are many other things you could be doing: going to a movie, having dinner with friends, playing ping pong, earning some extra money, watching TV. . . . But, assuming you're still reading—and you haven't just run out the door because we've given you better ideas—let's relate this to opportunity cost.

What *is* the opportunity cost of reading this chapter? Is it *all* of those other possibilities we've listed? Not really, because in the time it takes to read this chapter, you'd probably be able to do only *one* of those other activities. You'd no doubt

choose whichever one you regarded as best. So, by reading, you sacrifice only the *best* choice among the alternatives that you could be doing instead.

When the alternatives to a choice are mutually exclusive, only the next best choice—the one that would actually be chosen—is used to determine the opportunity cost of the choice.

For many choices, a large part of the opportunity cost is the money sacrificed. If you spend $15 on a new DVD, you have to part with $15, which is money you could have spent on something else (whatever the best choice among the alternatives turned out to be). But for other choices, money may be only a small part, or no part, of what is sacrificed. If you walk your dog a few blocks, it will cost you time but not money.

Still, economists often like to attach a monetary value even to the parts of opportunity cost that *don't* involve money. The opportunity cost of a choice can then be expressed as a dollar value, albeit a roughly estimated one. That, in turn, enables us to compare the cost of a choice with its benefits, which we also often express in dollars.

An Example: The Opportunity Cost of College

Let's consider an important choice you've made for this year: to attend college. What is the opportunity cost of this choice? A good starting point is to look at the actual monetary costs—the annual out-of-pocket expenses borne by you or your family for a year of college. Table 1 shows the College Board's estimates of these expenses for the average student (ignoring scholarships). For example, the third column of the table shows that the average in-state resident at a four-year state college pays $6,585 in tuition and fees, $1,077 for books and supplies, $7,748 for room and board, and $2,916 for transportation and other expenses, for a total of $18,326 per year.

So, is that the average opportunity cost of a year of college at a public institution? Not really. Even if $18,326 is what you or your family actually pays out for college, this is not the dollar measure of the opportunity cost.

TABLE 1

Average Cost of a Year of College, 2008–2009

Type of Institution	Two-Year Public	Four-Year Public	Four-Year Private
Tuition and fees	$2,402	$6,585	$25,143
Books and supplies	$1,036	$1,077	$1,054
Room and board	$7,341	$7,748	$8,989
Transportation and other expenses	$3,275	$2,916	$2,204
Total out-of-pocket costs	$14,054	$18,326	$37,390

Source: *Trends in College Pricing*, 2008, The College Board, New York, NY.

Notes: Averages are enrollment-weighted by institution, to reflect the average experience among students across the United States. Average tuition and fees at public institutions are for in-state residents only. Room and board charges are for students living on campus at four-year institutions, and off-campus (but not with parents) at two-year institutions. Four-year private includes nonprofit only.

First, the $18,326 your family pays in this example includes some expenses that are *not* part of the opportunity cost of college. For example, room and board is something you'd need no matter *what* your choice. For example, if you didn't go to college, you might have lived in an apartment and paid rent. But suppose, instead, that if you didn't go to college you would have chosen to live at home in your old room. Even then, you could not escape a cost for room and board. Your family *could* have rented out the room to someone else, or used it for some other valuable purpose. Either way, something would be sacrificed for room and board, whether you go to college or not.

Let's suppose, for simplicity, that if you weren't in college, you or your family would be paying the same $7,748 for room and board as your college charges. Then, the room and board expense should be excluded from the opportunity cost of going to college. And the same applies to transportation and other expenses, at least the part that you would have spent anyway even if you weren't in college. We'll assume these other expenses, too, are the same whether or not you go to college.

Now we're left with payments for tuition and fees, and for books and supplies. For an in-state resident going to a state college, this averages $6,585 + $1,077 = $7,662 per year. Since these dollars are paid only when you attend college, they represent something sacrificed for that choice and are part of its opportunity cost. Costs like these—for which dollars are actually paid out—are called **explicit costs**, and they are *part* of the opportunity cost.

Explicit cost The dollars sacrificed—and actually paid out—for a choice.

Implicit cost The value of something sacrificed when no direct payment is made.

But college also has **implicit costs**—sacrifices for which no money changes hands. The biggest sacrifice in this category is *time*. But what is that time worth? That depends on what you *would* be doing if you weren't in school. For many students, the alternative would be working full-time at a job. If you are one of these students, attending college requires the sacrifice of the income you *could* have earned at a job—a sacrifice we call *foregone income*.

How much income is foregone when you go to college for a year? In 2008, the average yearly income of an 18- to 24-year-old high school graduate who worked full-time was about $24,000. If we assume that only nine months of work must be sacrificed to attend college and that you could still work full-time in the summer, then foregone income is about 3/4 of $24,000, or $18,000.

Summing the explicit and implicit costs gives us a rough estimate of the opportunity cost of a year in college. For a public institution, we have $7,662 in explicit costs and $18,000 in implicit costs, giving us a total of $25,662 per year. Notice that this is significantly greater than the total charges estimated by the College Board we calculated earlier. When you consider paying this opportunity cost for four years, its magnitude might surprise you. Without financial aid in the form of tuition grants or other fee reductions, the average in-state resident will sacrifice about $103,000 to get a bachelor's degree at a state college and about $177,000 at a private one.

Our analysis of the opportunity cost of college is an example of a general, and important, principle:

The opportunity cost of a choice includes both explicit costs and implicit costs.

A Brief Digression: Is College the Right Choice?

Before you start questioning your choice to be in college, there are a few things to remember. First, for many students, scholarships reduce the costs of college to less than those in our example. Second, in addition to its high cost, college has substantial *benefits*, including financial ones. In fact, over a 40-year work life, the average college graduate will make about $2.5 million, which is about a million dollars *more* than the average high school graduate.

True, much of that income is earned in the future, and a dollar gained years from now is worth less than a dollar spent today. Also, *some* of the higher earnings of college graduates result from the personal characteristics of people who are likely to attend college, rather than from the education or the degree itself. But even when we make reasonable adjustments for these facts, attending college appears to be one of the best *financial* investments you can make.[1]

Finally, remember that we've left out of our discussion many important aspects of this choice that would be harder to estimate in dollar terms but could be very important to you. Do you *enjoy* being at college? If so, your enjoyment is an added benefit, even though it may be difficult to value that enjoyment in dollars. (Of course, if you *hate* college and are only doing it for the financial rewards or to satisfy your parents, that's an implicit cost—which is part of your opportunity cost—that we haven't included.)

Time Is Money

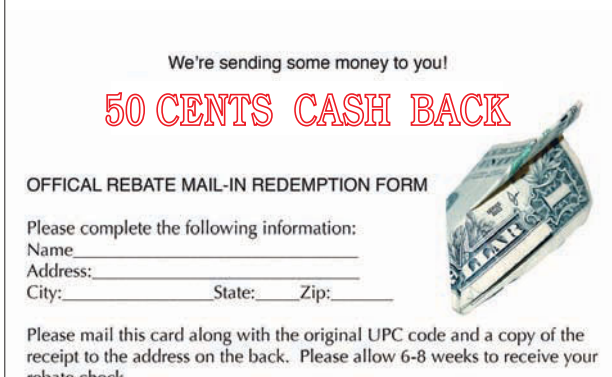
Our analysis of the opportunity cost of college points out a general principle, one understood by economists and noneconomists alike. It can be summed up in the expression, "Time is money."

For some people, this maxim applies directly: when they spend time on something, they *actually* give up money—money they *could* have earned during that time. Consider Jessica, a freelance writer with a backlog of projects on which she can earn $25 per hour. For each hour Jessica spends *not* working, she sacrifices $25.

What if Jessica decides to see a movie? What is the opportunity cost, in dollar terms? Suppose the ticket costs $10 and the entire activity takes three hours—including time spent getting there and back. The opportunity cost is the sum of the explicit cost ($10 for the ticket) and the implicit cost ($75 for three hours of foregone income), making the total opportunity cost $85.

The idea that a movie "costs" $85 might seem absurd to you. But if you think about it, $85 is a much better estimate than $10 of what the movie actually costs Jessica—$85 is what she sacrifices to see the movie.

What's wrong with this picture?



dangerous curves

If you think the opportunity cost of your time is zero . . . What if you can't work extra hours for additional pay, so you cannot *actually* turn time into money? Does this mean that the opportunity cost of your time is zero?

If you think the answer is yes, the authors of this textbook would like to hire you for help with some household chores, for 25 cents an hour. Does this sound like a good deal to you? It would, if the opportunity cost of your time really had no value. If it doesn't sound like a good deal, then the time you'd be giving up must have some positive value to you. If pressed, you could state that value in money terms—and it would no doubt exceed 25 cents per hour.

[1] If you are studying microeconomics, you'll learn more about the value of college as an investment and the general technique economists use to compare future earnings with current costs in a later chapter.

Our examples about the cost of college and the cost of a movie point out an important lesson about opportunity cost:

> *The explicit (direct money) cost of a choice may only be a part—and sometimes a small part—of the opportunity cost of a choice.*

Scarcity and Social Choice

Now let's think about scarcity and choice from *society*'s point of view. What are the goals of our society? We want a high standard of living for our citizens, clean air, safe streets, good schools, and more. What is holding us back from accomplishing all of these goals in a way that would satisfy everyone? You already know the answer: scarcity. In society's case, the problem is a scarcity of **resources**—the things we use to make goods and services that help us achieve our goals.

Resources The labor, capital, land (including natural resources), and entrepreneurship that are used to produce goods and services.

The Four Resources

Resources are the most basic elements used to make goods and services. We can classify resources into four categories:

Labor The time human beings spend producing goods and services.

- **Labor**—the time human beings spend producing goods and services.

Capital A long-lasting tool that is used to produce other goods.

Physical capital The part of the capital stock consisting of physical goods, such as machinery, equipment, and factories.

Human capital The skills and training of the labor force.

- **Capital**—any long-lasting tool, that is itself produced, and helps us make other goods and services.

  More specifically, **physical capital** consists of things like machinery and equipment, factory buildings, computers, and even hand tools like hammers and screwdrivers. These are all long-lasting *physical* tools that we produce to help us make other goods and services.

  Another type of capital is **human capital**—the skills and knowledge possessed by workers. These satisfy our definition of capital: They are *produced* (through education and training), they help us produce *other* things, and they last for many years, typically through an individual's working life.

  Note the word *long-lasting* in the definition. If something is used up quickly in the production process—like the flour a baker uses to make bread—it is generally *not* considered capital. A good rule of thumb is that capital should last at least a year, although most types of capital last considerably longer.

  The **capital stock** is the total amount of capital at a nation's disposal at any point in time. It consists of all the capital—physical and human—created in previous periods that is still productively useful.

Capital stock The total amount of capital in a nation that is productively useful at a particular point in time.

Land The physical space on which production takes place, as well as the natural resources that come with it.

- **Land**—the physical space on which production takes place, as well as useful materials—*natural resources*—found under it or on it, such as crude oil, iron, coal, or fertile soil.

Entrepreneurship The ability and willingness to combine the *other* resources—labor, capital, and land—into a productive enterprise.

- **Entrepreneurship**—the ability (and the willingness to *use* it) to combine the *other* resources into a productive enterprise. An entrepreneur may be an *innovator* who comes up with an original idea for a business or a *risk taker* who provides her own funds or time to nurture a project with uncertain rewards.

Anything *produced* in the economy comes, ultimately, from some combinations of the four resources.

Think about the last lecture you attended at your college. Some resources were used *directly*: Your instructor's labor and human capital (his or her knowledge of economics); physical capital (the classroom building, a blackboard or projector); and land (the property on which your classroom building sits). Somebody played the role of entrepreneur, bringing these resources together to create your college in the first place. (If you attend a public institution, the entrepreneurial role was played by your state government.)

Many other inputs—besides those special inputs we call resources—were also used to produce the lecture. But these other inputs were themselves produced from resources, as illustrated in Figure 1. For example, the electricity used to power the lights in your classroom is an input, not a resource. Electricity is produced using crude oil, coal or natural gas (land and natural resources); coal miners or oil-riggers (labor); and electricity-generating turbines and power cables (capital).

Input Anything (including a resource) used to produce a good or service.

dangerous curves

Resources versus inputs The term *resources* is often confused with another, more general term—**inputs.** An input is *anything* used to make a good or service. Inputs include not only resources but also many other things made from them (cement, rolled steel, electricity), which are, in turn, used to make goods and services. *Resources*, by contrast, are the *special* inputs that fall into one of four categories: labor, land, capital, and entrepreneurship. They are the ultimate source of everything that is produced.

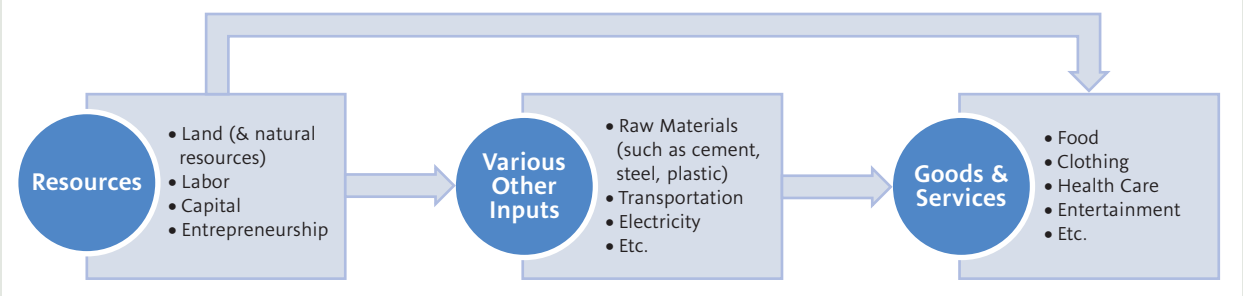
Opportunity Cost and Society's Tradeoffs

For an individual, opportunity cost arises from the scarcity of time or money. But for society as a whole, opportunity cost arises from the scarcity of *resources*. Our desire for goods is limitless, but we have limited resources to produce them. Therefore,

> *virtually all production carries an opportunity cost: To produce more of one thing, society must shift resources away from producing something else.*

For example, we'd all like better health for our citizens. What would be needed to achieve this goal? Perhaps more frequent medical checkups for more people and greater access to top-flight medicine when necessary. These, in turn, would require more and better-trained doctors, more hospital buildings and laboratories, and more high-tech medical equipment. In order for us to produce these goods and services, we would have to pull resources—land, labor, capital,

FIGURE 1 Resources and Production



All goods and services come ultimately from the four resources. Resources are used directly by firms that produce goods and services. They are also used indirectly, to make the other inputs firms use to produce goods and services.

and entrepreneurship—out of producing other things that we also enjoy. The opportunity cost of improved health care, then, consists of those other goods and services we would have to do without.

The World of Economics

The field of economics is surprisingly broad. It ranges from the mundane (why does a pound of steak cost more than a pound of chicken?) to the personal (how do couples decide how many children to have?) to the profound (could we ever have another Great Depression in the United States, with tens of millions plunged into sudden poverty?). With a field this broad, it is useful to have some way of classifying the different types of problems economists study and the different methods they use to analyze them.

Microeconomics and Macroeconomics

Microeconomics The study of the behavior of individual households, firms, and governments; the choices they make; and their interaction in specific markets.

Macroeconomics The study of the behavior of the overall economy.

Positive economics The study of how the economy works.

The field of economics is divided into two major parts: microeconomics and macroeconomics. **Microeconomics** comes from the Greek word *mikros*, meaning "small." It takes a close-up view of the economy, as if looking through a microscope. Microeconomics is concerned with the behavior of *individual* actors on the economic scene—households, business firms, and governments. It looks at the choices they make and how they interact with each other when they come together to trade *specific* goods and services. What will happen to the cost of movie tickets over the next five years? How many management-trainee jobs will open up for college graduates? These are microeconomic questions because they analyze individual *parts* of an economy rather than the *whole*.

Macroeconomics—from the Greek word *makros*, meaning "large"—takes an *overall* view of the economy. Instead of focusing on the production of carrots or computers, macroeconomics lumps all goods and services together and looks at the economy's *total output*. Instead of focusing on employment of management trainees or manufacturing workers, it considers *total employment* in the economy. Macroeconomics focuses on the big picture and ignores the fine details.

dangerous curves

Seemingly Positive Statements Be alert to statements that may *seem* purely positive, but contain hidden value judgments. Here's an example: "If we want to reduce greenhouse gas emissions, our society will have to use less gasoline." This may *sound* positive, because it seems to refer only to a fact about the world. But it's also at least partly normative. Why? Cutting back on gasoline is just *one* policy among many that could reduce emissions. To say that we *must* choose this method makes a value judgment about its superiority to other methods. A purely positive statement on this topic would be, "Using less gasoline—with no other change in living habits—would reduce greenhouse gas emissions."

Similarly, be alert to statements that use vague terms that hide value judgments. An example: "All else equal, the less gasoline we use, the better our quality of life." Whether you agree or disagree, this is *not* a purely positive statement. People can disagree over the meaning of the phrase "quality of life," and what would make it better. This disagreement could not be resolved just by looking at the facts.

Positive and Normative Economics

The micro versus macro distinction is based on the level of detail we want to consider. Another useful distinction has to do with our *purpose* in analyzing a problem. **Positive economics** explains how the economy works, plain and simple. If someone says, "The decline in home prices during 2008 and 2009 was a major cause of the recent recession," he or she is making a positive economic statement. A statement need not be accurate or even sensible to be classified as positive. For example, "Government policy has no effect on our standard of living" is a statement that virtually every economist would regard as false. But it is still a positive economic statement. Whether true or

not, it's about how the economy works and its accuracy can be tested by looking at the facts—and just the facts.

Normative economics *prescribes solutions* to economic problems. It goes beyond just "the facts" and tells us what we should *do* about them. Normative economics requires us to make judgments about different outcomes and therefore depends on our values.

Normative economics The practice of recommending policies to solve economic problems.

If an economist says, "We should cut total government spending," he or she is engaging in normative economic analysis. Cutting government spending would benefit some citizens and harm others, so the statement rests on a value judgment. A normative statement—like the one about government spending earlier—cannot be proved or disproved by the facts alone.

Positive and normative economics are intimately related in practice. For one thing, we cannot properly argue about what we should or should not do unless we know certain facts about the world. Every normative analysis is therefore based on an underlying positive analysis. But while a positive analysis can, at least in principle, be conducted without value judgments, a normative analysis is always based, at least in part, on the values of the person conducting it.

Why Economists Disagree about Policy

Suppose the country is suffering from a serious recession—a significant, nationwide decrease in production and employment. Two economists are interviewed on a cable news show.

Economist A says, "We should increase government spending on roads, bridges, and other infrastructure. This would directly create jobs and help end the recession." Economist B says, "No, we should cut taxes instead. This will put more money in the hands of households and businesses, leading them to spend more and create jobs that way." Why do they disagree?

It might be based on *positive* economics—different views about how the economy works. Economist A might think that government spending will create more jobs, dollar for dollar, than will tax cuts. Economist B might believe the reverse. Positive differences like these can arise because our knowledge of how the economy works—while always improving—remains imperfect.

But the disagreement might stem from a difference in values—specifically, what each economist believes about government's proper role in the economy. Those toward the left of the political spectrum tend to believe that government should play a larger economic role. They tend to view increases in government spending more favorably. Those toward the right tend to believe that government's role should be smaller. They would prefer tax cuts that result in more private, rather than government, spending. This difference in values can explain why two economists—even if they have the same *positive* views about the outcome of a policy—might disagree about its wisdom.

> *Policy differences among economists arise from (1) positive disagreements (about what the outcome of different policies will be), or (2) differences in values (how those outcomes are evaluated).*

Policy disputes among economists are common. But on *some* policy issues, most economists agree. For example, in microeconomics there is wide agreement that certain types of goods and services should be provided by private business firms and that certain others are best provided by government. In macroeconomics, almost all economists agree that some of the government policies during the Great Depression

of the 1930s were mistakes. Indeed, as the U.S. and global economies sank deeper into recession in 2008–2009, economists were virtually united in warning against repeating these mistakes. You will learn more about these and other areas of agreement in the chapters to come.

Why Study Economics?

If you've read this far into the chapter, chances are you've already decided to allocate some of your scarce time to studying economics. We think you've made a wise choice. But it's worth taking a moment to consider what you might gain from this choice.

Why study economics?

To Understand the World Better

Applying the tools of economics can help you understand global and catastrophic events such as wars, famines, epidemics, and depressions. But it can also help you understand much of what happens to you locally and personally—the salary you will earn after you graduate, or the rent you'll pay on your apartment. Economics has the power to help us understand these phenomena because they result, in large part, from the choices we make under conditions of scarcity.

Economics has its limitations, of course. But it is hard to find any aspect of life about which economics does not have *something* important to say. Economics cannot explain why so many Americans like to watch television, but it *can* explain how TV networks decide which programs to offer. Economics cannot protect you from a robbery, but it *can* explain why some people choose to become thieves and why no society has chosen to eradicate crime completely. Economics will not improve your love life, resolve unconscious conflicts from your childhood, or help you overcome a fear of flying, but it *can* tell us how many skilled therapists, ministers, and counselors are available to help us solve these problems.

To Achieve Social Change

If you are interested in making the world a better place, economics is indispensable. There is no shortage of serious social problems worthy of our attention—unemployment, hunger, poverty, disease, child abuse, drug addiction, violent crime. Economics can help us understand the origins of these problems, explain why previous efforts to solve them haven't succeeded, and help us to design new, more effective solutions.

To Help Prepare for Other Careers

Economics has long been a popular college major for individuals intending to work in business. But it has also been popular among those planning careers in politics, international relations, law, medicine, engineering, psychology, and other professions. This is for good reason: Practitioners in each of these fields often find themselves confronting economic issues. For example, lawyers increasingly face judicial rulings based on the principles of economic efficiency. And doctors will need to understand how new technologies or changes in the structure of health insurance will affect their practices.

To Become an Economist

Only a tiny minority of this book's readers will decide to become economists. This is welcome news to the authors, and after you have studied labor markets in your

microeconomics course you will understand why. But if you do decide to become an economist—obtaining a master's degree or a Ph.D.—you will find many possibilities for employment. The economists with whom you have most likely had personal contact are those who teach and conduct research at colleges and universities. But as many economists work outside of colleges and universities as work inside them. Economists are hired by banks to assess the risk of investing abroad; by manufacturing companies to help them determine new methods of producing, marketing, and pricing their products; by government agencies to help design policies to fight crime, disease, poverty, and pollution; by international organizations to help create and reform aid programs for less developed countries; by the media to help the public interpret global, national, and local events; and by nonprofit organizations to provide advice on controlling costs and raising funds more effectively.

The Methods of Economics

One of the first things you will notice as you begin to study economics is the heavy reliance on *models*.

You have no doubt encountered many models in your life. As a child, you played with model trains, model planes, or model people (dolls). You may have also seen architects' cardboard models of buildings. These are physical models, three-dimensional replicas that you can pick up and hold. Economic models, on the other hand, are built not with cardboard, plastic, or metal but with words, diagrams, and mathematical statements.

What, exactly, is a model?

*A **model** is an abstract representation of reality.*

Model An abstract representation of reality.

The two key words in this definition are *abstract* and *representation*. A model is not supposed to be exactly like reality. Rather, it *represents* the real world by *abstracting* or *taking from* the real world that which will help us understand it. By definition, a model leaves out features of the real world.

The Art of Building Economic Models

When you build a model, how do you know which real-world details to include and which to leave out? There is no simple answer to this question. The right amount of detail depends on your purpose in building the model in the first place. There is, however, one guiding principle:

A model should be as simple as possible to accomplish its purpose.

This means that a model should contain only the *necessary* details.

To understand this a little better, think about a map. A map is a model that represents a part of the earth's surface. But it leaves out many details of the real world. First, a map leaves out the third dimension—height—of the real world. Second, maps always ignore small details, such as trees and houses and potholes. But when you buy a map, how much detail do you want it to have?

Suppose you're in Boston and you want to drive from Logan Airport to the downtown convention center. You will need a detailed street map, as on the left side

FIGURE 2 Maps as Models

© SUSAN VAN ETTEN

These maps are models. But each would be used for a different purpose.

of Figure 2. The highway map on the right doesn't show any streets, so it wouldn't do at all.

But if you instead wanted to find the best driving route from Boston to Cincinnati, you would want a highway map. The map with individual streets would have too much detail and be harder to use.

The same principle applies in building economic models. The level of detail that would be just right for one purpose will usually be too much or too little for another. When you feel yourself objecting to an economic model because something has been left out, keep in mind the purpose for which the model is built. In introductory economics, the purpose is entirely educational—to help you understand some simple, but powerful, principles about how the economy operates. Keeping the models simple makes it easier to see these principles at work and remember them later.

Assumptions and Conclusions

Every economic model makes two types of assumptions: *simplifying* assumptions and *critical* assumptions.

Simplifying assumption Any assumption that makes a model simpler without affecting any of its important conclusions.

A **simplifying assumption** is just what it sounds like—a way of making a model simpler without affecting any of its important conclusions. A road map, for example, makes the simplifying assumption, "There are no trees." Having trees on a map would only get in the way. Similarly, in an economic model, we might assume that there are only two goods that households can choose from or that there are only two nations in the world. We make such assumptions *not* because we think they are true, but because they make a model easier to follow and do not change any of the important insights we can get from it.

Critical assumption Any assumption that affects the conclusions of a model in an important way.

A **critical assumption**, by contrast, is an assumption that affects the conclusions of a model in important ways. When you use a road map, you make the critical assumption, "All of these roads are open." If that assumption is wrong, your conclusion—the best route to take—might be wrong as well.

In an economic model, there are always one or more critical assumptions. You don't have to look very hard to find them because economists like to make them explicit right from the outset. For example, when we study the behavior of business firms, our model will assume that firms try to earn the highest possible profit for

their owners. By stating this critical assumption up front, we can see immediately where the model's conclusions spring from.

Math, Jargon, and Other Concerns . . .

Economists often express their ideas using mathematical concepts and a special vocabulary. Why? Because these tools enable economists to express themselves more precisely than with ordinary language. For example, someone who has never studied economics might say, "When gas is expensive, people don't buy big, gas-guzzling cars." That statement might not bother you right now. But once you've finished your first economics course, you'll be saying it something like this: "When the price of gas rises, the demand curve for big, gas-guzzling cars shifts leftward."

Does the second statement sound strange to you? It should. First, it uses a special term—a *demand curve*—that you've yet to learn. Second, it uses a mathematical concept—a *shifting curve*—with which you might not be familiar. But while the first statement might mean a number of different things, the second statement—as you will see in Chapter 3—can mean only *one* thing. By being precise, we can steer clear of unnecessary confusion.

If you are worried about the special vocabulary of economics, you can relax. After all, you may never have heard the term "opportunity cost" before, but now you know what it means. New terms will be defined and carefully explained as you encounter them. Indeed, this textbook does not assume you have any special knowledge of economics. It is truly meant for a "first course" in the field.

But what about the math? Here, too, you can relax. While professional economists often use sophisticated mathematics to solve problems, only a little math is needed to understand basic economic *principles*. And virtually all of this math comes from high school algebra and geometry.

Still, if you have forgotten some of your high school math, a little brushing up might be in order. This is why we have included an appendix at the end of this chapter. It covers some of the most basic concepts—such as interpreting graphs, the equation for a straight line, and the concept of a slope—that you will need in this course. You may want to glance at this appendix now, just so you'll know what's there. Then, from time to time, you can go back to it when you need it.

How to Study Economics

As you read this book or listen to your instructor, you may find yourself following along and thinking that everything makes perfect sense. Economics may even seem easy. Indeed, it *is* rather easy to *follow* economics, since it's based so heavily on simple logic. But *following* and *learning* are two different things. You will eventually discover (preferably *before* your first exam) that economics must be studied actively, not passively.

If you are reading these words lying back on a comfortable couch, a phone in one hand and a remote control in the other, you are going about it in the wrong way. Active studying means reading with a pencil in your hand and a blank sheet of paper in front of you. It means closing the book periodically and *reproducing* what you have learned. It means listing the steps in each logical argument, retracing the flow of cause and effect in each model, and drawing the graphs. It does require some work, but the payoff is a good understanding of economics and a better understanding of your own life and the world around you.

SUMMARY

One of the most fundamental concepts in economics is *opportunity cost*. The opportunity cost of any choice is what we give up when we make that choice.

Economics is the study of choice under conditions of scarcity. As individuals—and as a society—we have unlimited desires for goods and services. Unfortunately, our ability to satisfy those desires is limited, so we must usually sacrifice something for any choice we make.

The correct measure of the cost of a choice is not just the money price we pay, but the *opportunity cost*: what we must give up when we make a choice.

At the individual level, opportunity cost arises from the scarcity of time or money. For society as a whole, it arises from the scarcity of *resources*: *land, labor, capital,* and *entrepreneurship*. To produce and enjoy more of one thing, society must shift resources away from producing something else. Therefore, we must choose which desires to satisfy and how to satisfy them. Economics provides the tools that explain those choices.

The field of economics is divided into two major areas. *Microeconomics* studies the behavior of individual households, firms, and governments as they interact in specific markets. *Macroeconomics*, by contrast, concerns itself with the behavior of the entire economy. It considers variables such as total output, total employment, and the overall price level.

Economics makes heavy use of *models*—abstract representations of reality—to help us understand how the economy operates. All models are simplifications, but a good model will have just enough detail for the purpose at hand. The *simplifying* assumptions in a model just make it easier to use. The *critical* assumptions are the ones that affect the model's conclusions.

PROBLEM SET

Answers to even-numbered Questions and Problems can be found on the text website at www.cengage.com/economics/hall.

1. Discuss whether each statement is a purely positive statement, or also contains normative elements and/or value judgments:
   a. An increase in the personal income tax will slow the growth rate of the economy.
   b. The goal of any country's economic policy should be to increase the well-being of its poorest, most vulnerable citizens.
   c. The best way to reduce the national poverty rate is to increase the federal minimum wage.
   d. The 1990s were a disastrous decade for the U.S. economy. Income inequality increased to its highest level since before World War II.

2. For each of the following, state whether economists would consider it a *resource*, and if they would, identify which of the four types of resources the item is.
   a. A computer used by an FBI agent to track the whereabouts of suspected criminals.
   b. The office building in which the FBI agent works.
   c. The time that an FBI agent spends on a case.
   d. A farmer's tractor.
   e. The farmer's knowledge of how to operate the tractor.
   f. Crude oil.
   g. A package of frozen vegetables.
   h. A food scientist's knowledge of how to commercially freeze vegetables.
   i. The ability to bring together resources to start a frozen food company.
   j. Plastic bags used by a frozen food company to hold its product.

3. Suppose you are using the second map in Figure 2, which shows main highways only. You've reached a conclusion about the fastest way to drive from the Boston city center to an area south of the city. State whether each of the following assumptions of the map would be a *simplifying* or *critical* assumption for your conclusion, and explain briefly. (Don't worry about whether the assumption is true or not.)
   a. The thicker, numbered lines are major highways without traffic lights.
   b. The earth is two-dimensional.
   c. When two highways cross, you can get from one to the other without going through city traffic.
   d. Distances on the map are proportional to distances in the real world.

4. Suppose that you are considering what to do with an upcoming weekend. Here are your options, from least to most preferred: (1) study for upcoming midterms; (2) fly to Colorado for a quick ski trip; (3) go into seclusion in your dorm room and try to improve your score on a computer game. What is the opportunity

cost of a decision to play the computer game all weekend?

5. Use the information in Table 1 as well as the assumption about foregone income made in the chapter to calculate the average opportunity cost of a year in college for a student at a four-year private institution under each of the following assumptions:
   a. The student receives free room and board at home at no opportunity cost to the parents.
   b. The student receives an academic scholarship covering all tuition and fees (in the form of a grant, not a loan or a work study aid).
   c. The student works half time while at school at no additional emotional cost.
6. Use the information in Table 1 (as well as the assumption about foregone income made in the chapter) to compare the opportunity cost of attending a year of college for a student at a two-year public college, under each of the following assumptions:
   a. The student receives free room and board at home at no opportunity cost to the parents.
   b. The student receives an academic scholarship covering all tuition and fees (in the form of a grant, not a loan or a work study aid).
   c. The student works half time while at school (assume that the leisure or study time sacrificed has no opportunity cost).
7. Consider Kylie, who has been awarded academic scholarships covering all tuition and fees at three different colleges. College #1 is a two-year public college. College #2 is a four-year public college, and College #3 is a four-year private college. Explain why, if the decision is based solely on opportunity cost, Kylie will turn down her largest scholarship offers. (Use Table 1 in the chapter.)

APPENDIX

Graphs and Other Useful Tools

Tables and Graphs

A brief glance at this text will tell you that graphs are important in economics. Graphs provide a convenient way to display information and enable us to immediately *see* relationships between variables.

Suppose that you've just been hired at the advertising department of Len & Harry's—an up-and-coming manufacturer of high-end ice cream products, located in Texas. You've been asked to compile a report on how advertising affects the company's sales. It turns out that the company's spending on advertising has changed repeatedly in the past, so you have lots of data on monthly advertising expenditure and monthly sales revenue, both measured in thousands of dollars.

Table A.1 shows a useful way of arranging this data. The company's advertising expenditure in different months are listed in the left-hand column, while the right-hand column lists total sales revenue ("sales" for short) during the same months. Notice that the data in this table is organized so that spending on advertising increases as we move down the first column. Often, just looking at a table like this can reveal useful patterns. In this example, it's clear that higher spending on advertising is associated with higher monthly sales. These two variables—advertising and sales—have a *positive relationship*. A rise in one is associated with a rise in the other. If higher advertising had been associated with *lower* sales, the two variables would have a *negative* or *inverse relationship:* A rise in one would be associated with a fall in the other.

We can be even more specific about the positive relationship between advertising and sales: Logic tells us that the association is very likely *causal*. We'd expect that sales revenue *depends on* advertising outlays, so we call sales our *dependent variable* and advertising our *independent variable*. Changes in an independent variable cause changes in a dependent variable, but not the other way around.

To explore the relationship further, let's graph it. As a rule, the *independent* variable is measured on the *horizontal* axis and the *dependent* variable on the *vertical* axis. In economics, unfortunately, we do not always stick to this rule, but for now we will. In Figure A.1, monthly advertising expenditure—our independent variable—is measured on the horizontal axis. If we start at the *origin*—the corner where the two axes intersect—and move rightward along the horizontal axis, monthly advertising spending increases from \$0 to \$1,000 to \$2,000 and so on. The vertical axis measures monthly sales—the dependent variable. Along this axis, as we move upward from the origin, sales rise.

The graph in Figure A.1 shows six labeled points, each representing a different pair of numbers from our

TABLE A.1

Advertising and Sales at Len & Harry's

Advertising Expenditures ($1,000 per Month)	Sales ($1,000 per Month)
2	24
3	27
6	36
7	39
11	51
12	54

FIGURE A.1 A Graph of Advertising and Sales

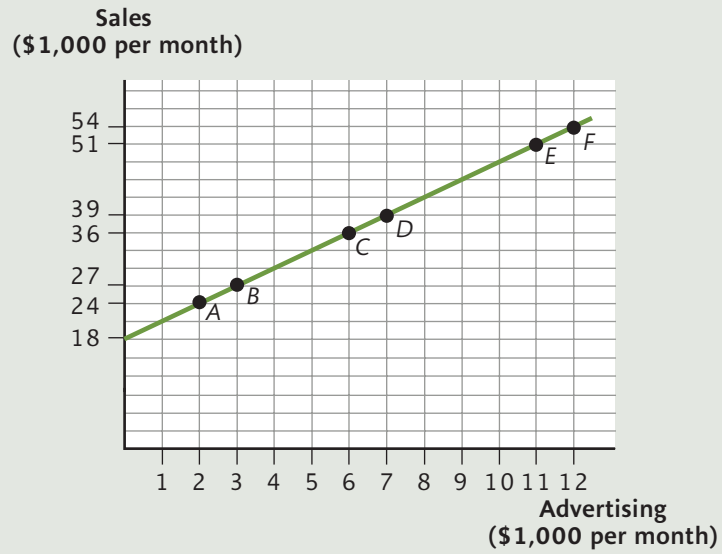


table. For example, point *A*—which represents the numbers in the first row of the table—shows us that when the firm spends $2,000 on advertising, sales are $24,000 per month. Point *B* represents the *second* row of the table, and so on. Notice that all of these points lie along a *straight line*.

Straight-Line Graphs

You'll encounter straight-line graphs often in economics, so it's important to understand one special property they possess: The "rate of change" of one variable compared with the other is always the same. For example, look at what happens as we move from point *A* to point *B*: Advertising rises by $1,000 (from $2,000 to $3,000), while sales rise by $3,000 (from $24,000 to $27,000). If you study the graph closely, you'll see that anywhere along this line, whenever advertising increases by $1,000, sales increase by $3,000. Or, if we define a "unit" as "one thousand dollars," we can say that every time advertising increases by one unit, sales rise by three units. So the "rate of change" is three units of sales for every one unit of advertising.

The rate of change of the *vertically* measured variable for a one-unit change in the *horizontally* measured variable is also called the *slope* of the line. The slope of the line in Figure A.1 is three, and it remains three no matter where along the line we measure it. For example, make sure you can see that from point *C* to point *D*, advertising rises by one unit and sales rise by three units.

What if we had wanted to determine the slope of this line by comparing points *D* and *E*, which has advertising rising by four units instead of just one? In that case, we'd have to calculate the rise in one variable *per unit* rise in the other. To do this, we divide the change in the vertically measured variable by the change in the horizontally measured variable.

$$\text{Slope of a straight line} = \frac{\text{Change in vertical variable}}{\text{Change in horizontal variable}}.$$

We can make this formula even simpler by using two shortcuts. First, we can call the variable on the vertical axis "*Y*" and the variable on the horizontal axis "*X*." In our case, *Y* is sales, while *X* is spending on advertising. Second, we use the Greek letter Δ ("delta") to denote the words "change in." Then, our formula becomes:

$$\text{Slope of straight line} = \frac{\Delta Y}{\Delta X}.$$

Let's apply this formula to get the slope as we move from point *D* to point *E*, so that advertising (*X*) rises

from 7 units to 11 units. This is an increase of 4, so $\Delta X = 4$. For this move, sales rise from 39 to 51, an increase of 12, so $\Delta Y = 12$. Applying our formula,

$$\text{Slope} = \frac{\Delta Y}{\Delta X} = \frac{12}{4} = 3.$$

This is the same value for the slope that we found earlier. Not surprising, since it's a straight line and a straight line has the same slope everywhere. The particular pair of points we choose for our calculation doesn't matter.

Curved Lines

Although many of the relationships you'll encounter in economics have straight-line graphs, many others do not. Figure A.2 shows *another* possible relationship between advertising and sales that we might have found from a different set of data. As you can see, the line is curved. But as advertising rises, the curve gets flatter and flatter. Here, as before, each time we spend another $1,000 on advertising, sales rise. But now, the rise in sales seems to get smaller and smaller. This means that the *slope* of the curve is *itself changing* as we move along this curve. In fact, the slope is getting smaller.

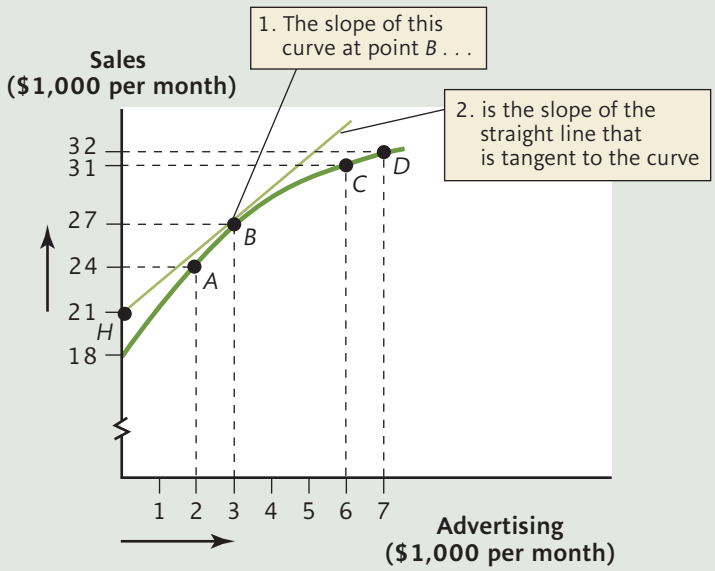
How can we measure the slope of a curve? First, note that since the slope is different at every point along the curve, we aren't really measuring the slope of "the curve" but the slope of the curve *at a specific point along it*. How can we do this? By drawing a *tangent line*—a straight line that touches the curve at just one point and that has the same slope as the curve at that point. For example, in the figure, a tangent line has been drawn for point *B*. To measure the slope of this tangent line, we can compare any two points on it, say, *H* and *B*, and calculate the slope as we would for any straight line. Moving from point *H* to point *B*, we are moving from 0 to 3 on the horizontal axis ($\Delta X = 3$) and from 21 to 27 on the vertical axis ($\Delta Y = 6$). Thus, the slope of the tangent line—which is the same as the slope of the curved line at point *B*—is

$$\frac{\Delta Y}{\Delta X} = \frac{6}{3} = 2.$$

This says that, at point *B*, the rate of change is two units of sales for every one unit of advertising. Or, going back to dollars, the rate of change is $2,000 in sales for every $1,000 spent on advertising.

The curve in Figure A.2 slopes everywhere upward, reflecting a positive relationship between the variables. But a curved line can also slope downward to

FIGURE A.2 Measuring the Slope of a Curve



illustrate a negative relationship between variables, or slope first one direction and then the other. You'll see plenty of examples of each type of curve in later chapters, and you'll learn how to interpret each one as it's presented.

Linear Equations

Let's go back to the straight-line relationship between advertising and sales, as shown in Table A.1. What if you need to know how much in sales the firm could expect if it spent $5,000 on advertising next month? What if it spent $8,000, or $9,000? It would be nice to be able to answer questions like this without having to pull out tables and graphs to do it. As it turns out, anytime the relationship you are studying has a straight-line graph, it is easy to figure out an equation for the entire relationship—a *linear equation*. You then can use the equation to answer any such question that might be put to you.

All straight lines have the same general form. If Y stands for the variable on the vertical axis and X for the variable on the horizontal axis, every straight line has an equation of the form

$$Y = a + bX,$$

where a stands for some number and b for another number. The number a is called the vertical *intercept*, because it marks the point where the graph of this equation hits (intercepts) the vertical axis; this occurs when X takes the value zero. (If you plug $X = 0$ into the equation, you will see that, indeed, $Y = a$.) The number b is the slope of the line, telling us how much Y will change every time X changes by one unit. To confirm this, note that when $X = 0$, the equation tells us that $Y = a$. When $X = 1$, it tells us that $Y = a + b$. So as X increases from 0 to 1, Y goes from a to $a + b$. The number b is therefore the change in Y corresponding to a one-unit change in X—exactly what the slope of the graph should tell us.

If b is a positive number, a one-unit increase in X causes Y to *increase* by b units, so the graph of our line would slope upward, as illustrated by the line in the upper left panel of Figure A.3. If b is a negative number, then a one-unit increase in X will cause Y to *decrease* by b units, so the graph would slope downward, as the line does in the lower left panel. Of course, b could equal zero. If it does, a one-unit increase in X causes no change in Y, so the graph of the line is flat, like the line in the middle left panel.

The value of a has no effect on the slope of the graph. Instead, different values of a determine the graph's position. When a is a positive number, the graph will intercept the vertical Y-axis above the origin, as the line does in the upper right panel of Figure A.3. When a is negative, however, the graph will intercept the Y-axis *below* the origin, like the line in the lower right panel. When a is zero, the graph intercepts the Y-axis right at the origin, as the line does in the middle right panel.

Let's see if we can figure out the equation for the relationship depicted in Figure A.1. There, X denotes advertising and Y denotes sales. Earlier, we calculated that the slope of this line, b, is 3. But what is a, the vertical intercept? In Figure A.1, you can see that when advertising outlays are zero, sales are $18,000. That tells us that $a = 18$. Putting these two observations together, we find that the equation for the line in Figure A.1 is

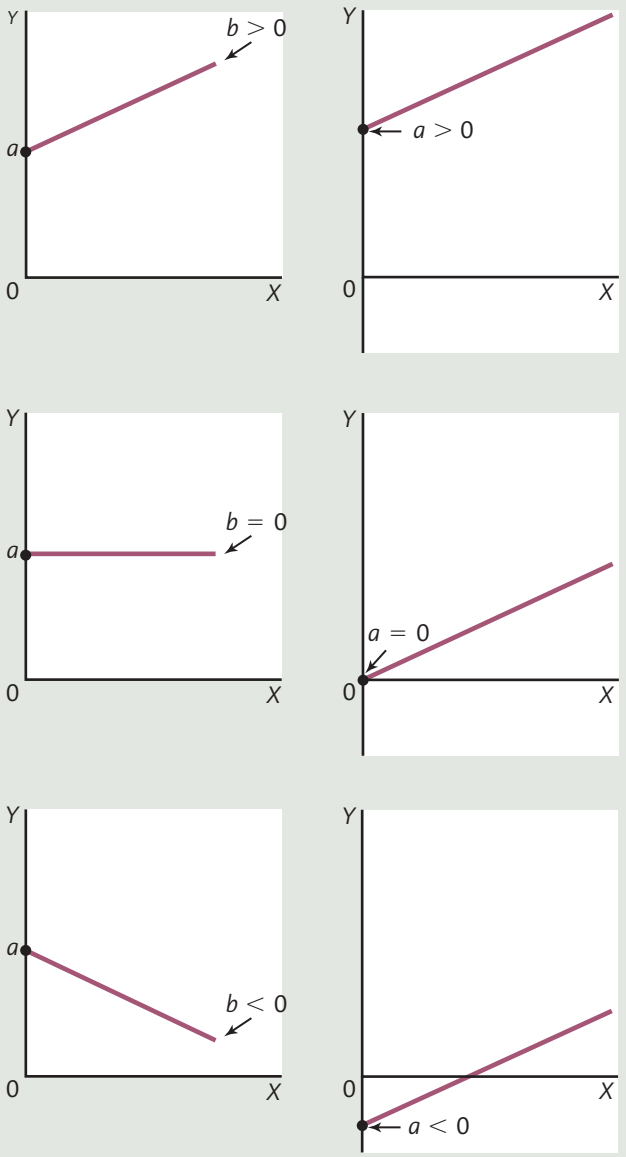
$$Y = 18 + 3X.$$

Now if you need to know how much in sales to expect from a particular expenditure on advertising (both in thousands of dollars), you would be able to come up with an answer: You'd simply multiply the amount spent on advertising by 3, add 18, and that would be your sales in thousands of dollars. To confirm this, plug in for X in this equation any amount of advertising in dollars from the left-hand column of Table A.1. You'll see that you get the corresponding amount of sales in the right-hand column.

How Straight Lines and Curves Shift

So far, we've focused on relationships where some variable Y depends on a single other variable, X. But in many of our theories, we recognize that some variable of interest to us is actually affected by more than just one other variable. When Y is affected by both X and some third variable, changes in that third variable will usually cause a *shift* in the graph of the relationship between X and Y. This is because whenever we draw the graph between X and Y, we are holding fixed every other variable that might possibly affect Y.

FIGURE A.3 **Straight Lines with Different Slopes and Vertical Intercepts**

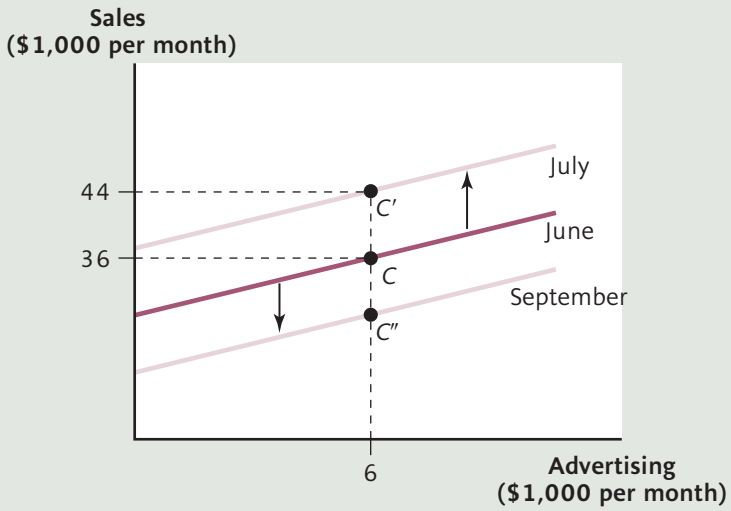


A graph between two variables X and Y is only a picture of their relationship when all other variables affecting Y are held constant.

But suppose one of these other variables *does* change? What happens then?

Think back to the relationship between advertising and sales. Earlier, we supposed sales depend only on advertising. But suppose we make an important discovery: Ice cream sales are *also* affected by how hot the weather is. What's more, all of the data in Table A.1 on which we previously based our analysis turns out to have been from the month of June in different years, when the average temperature in Texas is 80 degrees. What's going to happen in July, when the average temperature rises to 100 degrees?

FIGURE A.4 **Shift in the Graphs of Advertising and Sales**



In Figure A.4 we've redrawn the graph from Figure A.1, this time labeling the line "June." Often, a good way to determine how a graph will shift is to perform a simple experiment like this: Put your pencil tip anywhere on the graph labeled June—let's say at point *C*. Now ask the following question: If I hold advertising constant at $6,000, do I expect to sell more or less ice cream as temperature rises in July? If you expect to sell more, then the amount of sales corresponding to $6,000 of advertising will be *above* point *C*, at a point such as *C′* (pronounced "C prime"), representing sales of $44,000. From this, we can tell that the graph will *shift upward* as temperature rises. In September, however, when temperatures fall, the amount of sales corresponding to $6,000 in advertising would be less than it is at point *C*. It would be shown by a point such as *C″*, (pronounced "C double-prime"). In that case, the graph would shift downward.

The same procedure works well whether the original graph slopes upward or downward and whether it is a straight line or a curved one. Figure A.5 sketches two examples. In panel (a), an increase in some third variable, *Z*, increases the value of *Y* for each value of *X*, so the graph of the relationship between *X* and *Y* shifts upward as *Z* increases. We often phrase it this way: "An increase in *Z* causes an increase in *Y, at any value of X*." In panel (b), we assume that an increase in *Z decreases* the value of *Y*, at any value of *X*, so the graph of the relationship between *X* and *Y* shifts *downward* as *Z* increases.

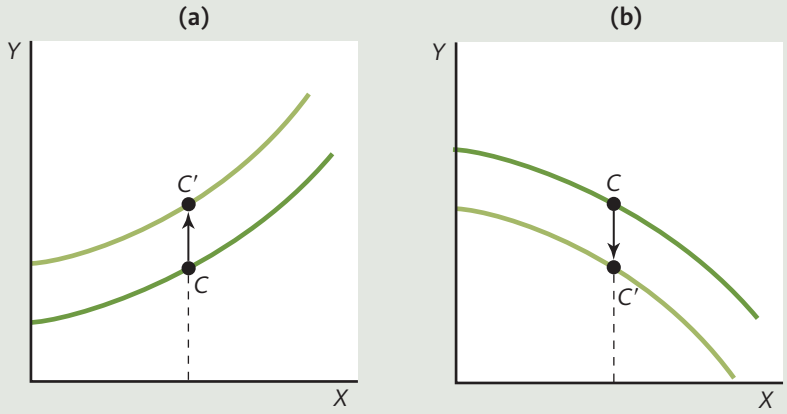
You'll notice that in Figures A.4 and A.5, the original line is darker, while the new line after the shift is drawn in a lighter shade. We'll often use this convention—a lighter shade for the new line after a shift in this book.

Shifts versus Movements Along a Line

If you look back at Figure A.1, you'll see that when advertising increases (say, from $2,000 to $3,000), we *move along* our line, from point *A* to point *B*. But you've just learned that when average temperature changes, the entire line *shifts*. This may seem strange to you. After all, in both cases, an independent variable changes (either advertising or temperature). Why should we move *along* the line in one case and *shift* it in the other?

The reason for the difference is that in one case (advertising), the independent variable is *in our graph*, measured along one of the axes. When an independent variable in the graph changes, we simply move along the line. In the other case (temperature), the independent variable does *not* appear in our graph. Instead, it's been in the background, being held constant.

FIGURE A.5 **Shifts of Curved Lines**



An increase in *Z* causes an increase in *Y* at any value of *X*

An increase in *Z* causes a decrease in *Y* at any value of *X*

Here's a very simple—but crucial—rule:

> *Suppose Y is the dependent variable, which is measured on one of the axes in a graph. If the independent variable* ***measured on the other axis changes,*** *we* ***move along*** *the line. But if* ***any other*** *independent variable changes, the* ***entire line shifts.***

Be sure you understand the phrase "any other independent variable." It refers to any variable that actually *affects* *Y* but is *not* measured on either axis in the graph.

This rule applies to straight lines as well as curved lines. And it applies even in more complicated situations, such as when *two different* lines are drawn in the same graph, and a shift of one causes a movement along the other. (You'll encounter this situation in Chapter 3.) But for now, make sure you can see how we've been applying this rule in our example, where the three variables are total sales, advertising, and temperature.

Solving Equations

When we first derived the equation for the relationship between advertising and sales, we wanted to know what level of sales to expect from different amounts of advertising. But what if we're asked a slightly different question? Suppose, this time, you are told that the sales committee has set an ambitious goal of $42,000 for next month's sales. The treasurer needs to know how much to budget for advertising, and you have to come up with the answer.

Since we know how advertising and sales are related, we ought to be able to answer this question. One way is just to look at the graph in Figure A.1. There, we could first locate sales of $42,000 on the vertical axis. Then, if we read over to the line and then down, we find the amount of advertising that would be necessary to generate that level of sales. Yet even with that carefully drawn diagram, it is not always easy to see just exactly how much advertising would be required. If we need to be precise, we'd better use the equation for the graph instead.

According to the equation, sales (Y) and advertising (X) are related as follows:

$$Y = 18 + 3X.$$

In the problem before us, we know the value we want for sales, and we need to solve for the corresponding amount of advertising. Substituting the sales target of $42, for Y, we need to find that value of X for which

$$42 = 18 + 3X.$$

In this case, X is the unknown value for which we want to solve.

Whenever we solve an equation for one unknown, say, X, we need to *isolate* X on one side of the equals sign and everything else on the other side of the equals sign. We do this by performing identical operations on both sides of the equals sign. Here, we can first subtract 18 from both sides, getting

$$24 = 3X.$$

We can then divide both sides by 3 and get

$$8 = X.$$

This is our answer. If we want to achieve sales of \$42,000, we'll need to spend \$8,000 on advertising.

Of course, not all relationships are linear, so this technique will not work in every situation. But no matter what the underlying relationship, the idea remains the same:

To solve for X in any equation, rearrange the equation, following the rules of algebra, so that X appears on one side of the equals sign and everything else in the equation appears on the other side.

CHAPTER 2

Scarcity, Choice, and Economic Systems

In the last chapter, you learned that economists use models—abstract representations of reality—to understand and explain how the economy works. In this chapter, you'll encounter your first simple economic model, designed to illustrate and help us analyze opportunity cost for society's choices.

Society's Production Choices

Let's consider a specific choice that faces every society: how much of its resources to allocate toward national defense versus how much to use for civilian production. To make this choice more concrete, we'll make a simplifying assumption: In the economy we're studying, there is one kind of military good (tanks) and one kind of civilian good (wheat).

Table 1 lists some possible combinations of yearly tank production and yearly wheat production this society could manage, given its available resources and the currently available production technology. For example, the first row of the table (choice A) tells us what would happen if all available resources were devoted to wheat production and no resources at all to producing tanks. The resulting quantity of wheat—1 million bushels per year—is the most this society could possibly produce. In the second row (choice B), society moves enough resources into tank production to make 1,000 tanks per year. This leaves fewer resources for wheat production, which now declines to 950,000 bushels per year.

TABLE 1

Production of Tanks and Wheat

Choice	Tank Production (number per year)	Wheat Production (bushels per year)
A	0	1,000,000
B	1,000	950,000
C	2,000	850,000
D	3,000	700,000
E	4,000	400,000
F	5,000	0

As we continue down the table, moving to choices C, D, and E, tank production increases by increments of 1,000. The last column shows us the maximum quantity of wheat that can be produced for each given quantity of tanks. Finally, look at the last row (choice F). It shows us that when society throws all of its resources into tank production (with none for wheat), tank production is 5,000 and wheat production is zero.

The table gives us a quantitative measure of opportunity cost for this society. For example, suppose this society currently produces 1,000 tanks per year, along with 950,000 bushels of wheat (choice B). What would be the opportunity cost of producing another 1,000 tanks? Moving down to choice C, we see that producing another 1,000 tanks (for a total of 2,000) would require wheat production to drop from 950,000 to 850,000 bushels, a decrease of 100,000 bushels. Thus, the opportunity cost of 1,000 more tanks is 100,000 bushels of wheat. The opportunity cost of having more of one good is measured in the units of the other good that must be sacrificed.

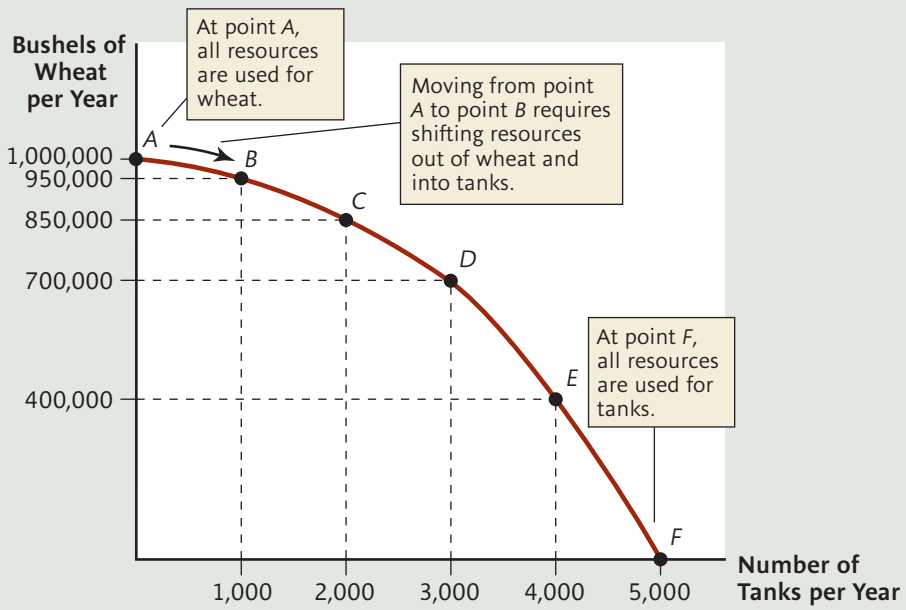
The Production Possibilities Frontier

We can see opportunity cost even more clearly in Figure 1, where the data in Table 1 has been plotted on a graph. In the figure, tank production is measured along the horizontal axis, and wheat production along the vertical axis. Each of the six points labeled *A* through *F* corresponds to one of society's choices in the table. For example, point *B* represents 1,000 tanks and 950,000 bushels of wheat.

When we connect these points with a smooth line, we get a curve called society's **production possibilities frontier (PPF).** Specifically, this PPF tells us the maximum quantity of wheat that can be produced for each quantity of tanks produced. Alternatively, it tells us the maximum number of tanks that can be produced for

Production possibilities frontier (PPF) A curve showing all combinations of two goods that can be produced with the resources and technology currently available.

FIGURE 1 The Production Possibilities Frontier



each different quantity of wheat. Positions outside the frontier are unattainable with the technology and resources at the economy's disposal. Society's choices are limited to points on or inside the PPF.

Now recall our earlier example of moving from choice B to choice C in the table. When tank production increased from 1,000 to 2,000, wheat production decreased from 950,000 to 850,000. In the graph, this change would be represented by a movement along the PPF from point *B* to point *C*. We're moving rightward (1,000 more tanks) and also downward (100,000 fewer bushels of wheat). Thus, the opportunity cost of 1,000 more tanks can be viewed as the vertical drop along the PPF as we move from point *B* to point *C*.

Increasing Opportunity Cost

Suppose we have arrived at point *C* and society then decides to produce still more tanks. Once again, resources must be shifted into tank production to make an additional 1,000 of them, moving from point *C* to point *D*. This time, however, there is an even *greater opportunity cost*: Production of wheat falls from 850,000 to 700,000 bushels, a sacrifice of 150,000 bushels. The opportunity cost of 1,000 more tanks has risen. Graphically, the vertical drop along the curve is greater for the same move rightward.

You can see that as we continue to increase tank production by increments of 1,000—moving from point *C* to point *D* to point *E* to point *F*—the opportunity cost of producing an additional 1,000 tanks keeps rising, until the last 1,000 tanks costs us 400,000 bushels of wheat. (You can also see this in the table, by running down the numbers in the right column. Each time tank production rises by 1,000, wheat production falls by more and more.)

The behavior of opportunity cost described here—the more tanks we produce, the greater the opportunity cost of producing still more—applies to a wide range of choices facing society. It can be generalized as the *law of increasing opportunity cost*.

> *According to the law of increasing opportunity cost, the more of something we produce, the greater the opportunity cost of producing even more of it.*

The law of increasing opportunity cost causes the PPF to have a concave (upside-down bowl) shape, becoming steeper as we move rightward and downward. That's because the slope of the PPF—the change in the quantity of wheat divided by the change in the quantity of tanks—can be interpreted as the change in wheat *per additional tank*. For example, moving from point *C* to point *D*, we give up 150,000 bushels of wheat to get 1,000 more tanks, or *150 bushels of wheat per tank*. Thus, the slope of the PPF between points *C* and *D* is approximately −150. (We say approximately because the PPF is curved, so its slope changes slightly as we move along the interval from *C* to *D*.) If we remove the minus sign from this slope and consider just its absolute value, it tells us the opportunity cost of *one more tank*.

Now—as we've seen—this opportunity cost increases as we move rightward. Therefore, the absolute value of the PPF's slope must rise as well. The PPF gets steeper and steeper, giving us the concave shape we see in Figure 1.[1]

[1] You might be wondering if the law of increasing opportunity cost applies in both directions. That is, does the opportunity cost of producing more wheat increase as we produce more of it? The answer is yes, as you'll be asked to find in an end-of-chapter problem.

The Reason for Increasing Opportunity Cost

Why does opportunity cost increase as we move along a PPF? Because most resources—by their very nature—are better suited to some purposes than to others. If the economy were operating at point *A*, for example, we'd be using *all* of our resources for wheat, even those that are much better suited to make tanks. People who would be better at factory work than farming would nevertheless be pressed into working on farms. And we'd be growing wheat on all the land available, even land that would be fine for a tank factory but awful for growing crops.

Now, as we begin to move rightward along the PPF, say from *A* to *B*, we would shift resources out of wheat production and into tank production. But we would *first* shift those resources *best suited* to tank production—and least suited for wheat. When these resources are shifted, an additional thousand tanks causes only a small drop in wheat production. This is why, at first, the PPF is very flat: a small vertical drop for the rightward movement.

As we continue moving rightward, however, we are forced to shift resources away from wheat—resources that are less and less suited to tanks and more and more suited to wheat. As a result, the PPF becomes steeper.

The principle of increasing opportunity cost applies to most of society's production choices, not just that between wheat and tanks. If we look at society's choice between food and oil, we would find that some land is better suited to growing food and other land is better suited to drilling for oil. As we continue to produce more oil, we would find ourselves drilling on land that is less and less suited to producing oil, but better and better for producing food. The opportunity cost of producing additional oil will therefore increase. The same principle applies if we want to produce more health care, more education, more automobiles, or more computers: The more of something we produce, the greater the opportunity cost of producing still more.

The Search for a Free Lunch

This chapter has argued that every decision to produce *more* of something requires us to pay an opportunity cost by producing less of something else. Nobel Prize–winning economist Milton Friedman summarized this idea in his famous remark, "There is no such thing as a free lunch." Friedman was saying that, even if a meal is provided free of charge to someone, society still uses up resources to provide it. Therefore, a "free lunch" is not *really* free: Society pays an opportunity cost by not producing other things with those resources. *Some* members of society will have less of something else.

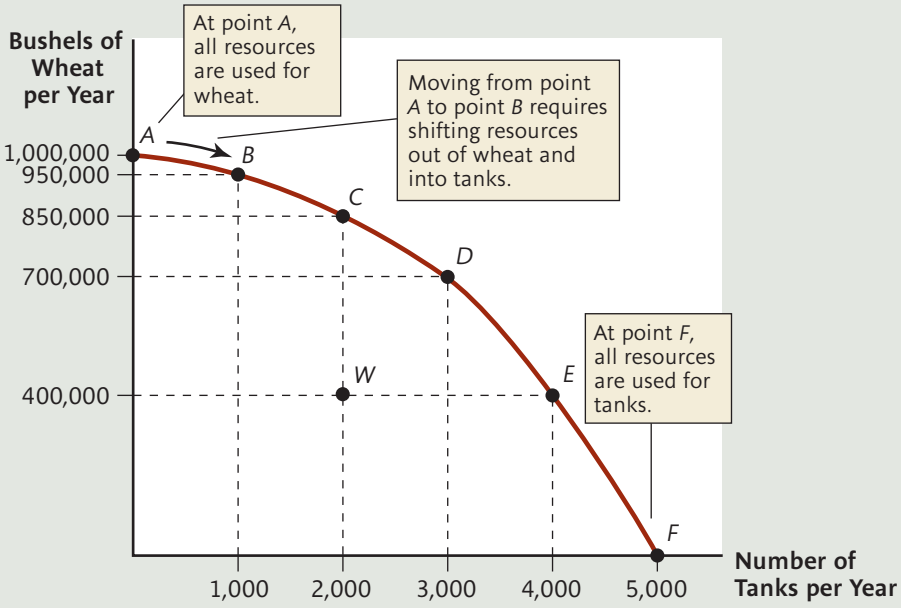
The same logic applies to other supposedly "free" goods and services. From society's point of view, there is no such thing as free Internet service, free broadcast television, or free medical care, even if those who enjoy these things don't pay for them as individuals. Providing any of these things requires a sacrifice of *other* things, as illustrated by a movement along society's PPF.

But there are some situations that seem, at first glance, to violate Freidman's dictum. Let's explore them.

Operating Inside the PPF

What if an economy is not living up to its productive potential, but is instead operating *inside* its PPF? For example, in Figure 2, suppose we are currently operating at

FIGURE 2 The Production Possibilities Frontier



point *W*, where we are producing 2,000 tanks and 400,000 bushels of wheat. Then we could move from point *W* to point *E* and produce 2,000 more tanks, with no sacrifice of wheat. Or, starting at point *W*, we could move to point *C* (more wheat with no sacrifice of tanks), or to a point like *D* (more of *both* wheat and tanks).

But why would an economy ever operate inside its PPF?

Productive Inefficiency

One possibility is that resources are not being used in the most productive way. Suppose, for example, that many people who could be outstanding wheat farmers are instead making tanks, and many who would be great at tank production are instead stuck on farms. Then switching people from one job to the other could enable us to have more of *both* tanks *and* wheat. That is, because of the mismatch of workers and jobs, we would be *inside* the PPF at a point like *W*. Creating better job matches would then move us to a point *on* the PPF (such as point *E*).

Economists use the phrase *productive inefficiency* to describe this type of situation that puts us inside our PPF.

Productively inefficient A situation in which more of at least one good can be produced without sacrificing the production of any other good.

A firm, an industry, or an entire economy is ***productively inefficient*** *if it could produce more of at least one good without pulling resources from the production of any other good.*

The phrase *productive efficiency* means the absence of any productive *in*efficiency.

Although no firm, industry, or economy is ever 100 percent productively efficient, cases of gross inefficiency are not as common as you might think. Business firms have strong incentives to identify and eliminate productive inefficiency, since

any waste of resources increases their costs and decreases their profit. When one firm discovers a way to eliminate waste, others quickly follow.

For example, empty seats on an airline flight represent productive inefficiency. Since the plane is making the trip anyway, filling the empty seat would enable the airline to serve more people with the flight (produce more transportation services) without using any additional resources (other than the trivial resources of in-flight snacks). Therefore, more people could fly without sacrificing any other good or service. When American Airlines developed a computer model in the late 1980s to fill its empty seats by altering schedules and fares, the other airlines followed its example very rapidly. And when—in the late 1990s—Priceline.com enabled airlines to auction off empty seats on the Internet, several airlines jumped at the chance and others quickly followed. As a result, a case of productive inefficiency in the airline industry—and therefore in the economy—was eliminated.

Starbucks provides another example. In 2000, the company analyzed how it makes drinks and eliminated several productively inefficient practices that it hadn't previously noticed. For example, it ended the practice of requiring signatures for small credit card purchases. It also began using larger scoops so that iced drinks could be made with one dip into the ice machine instead of two. These and other changes freed up labor time and enabled the company to make more drinks and serve more customers without using any additional resources.

Economists, logistics experts, and engineers are continually identifying and designing policies to eliminate cases of productive inefficiency. But many instances still remain. Why?

Usually, it's because the inefficiency creates benefits for individuals or groups who will resist changes in the status quo. For example, the government currently requires every taxpayer to file a federal tax return. About 40 percent of these returns are so simple that they merely provide the Internal Revenue Service (IRS) with information it already has, and contain calculations that the IRS duplicates anyway, to check for mistakes. Yet each taxpayer in this 40 percent group must spend hours doing his or her own return or else pay someone to do it. Why not have the IRS send these people filled-out returns, requiring only a signature if they approve?

One economist has estimated that this simple change would save a total of 250 million hours per year (for those who currently fill out their own returns), and $2 billion per year (for those who pay accountants). With resources freed up by this change, we could produce and enjoy more of all the things that we value. But if you reread this paragraph, you can probably guess who might lobby the government to oppose this change, if and when it is seriously considered.

When political obstacles prevent us from eliminating inefficiency, we are back to where we started: producing more of one thing requires taking resources away from something else we value, rather than getting "free" resources from greater efficiency. Productive inefficiency does create a theoretical possibility for a free lunch. But in practice, it does not offer as many hearty meals as you might think.

Recessions

Another reason an economy might operate inside its PPF is a *recession*—a slowdown in overall economic activity. During recessions, many resources are idle. For one thing, there is widespread *unemployment*—people *want* to work but are unable to find jobs. In addition, factories shut down, so we are not using all of our available capital. An end to the recession would move the economy from a point *inside* its PPF to a point *on* its PPF—using idle resources to produce more goods and services without sacrificing anything.

dangerous curves

False Benefits from Employment Often, you'll hear an evaluation of some economic activity that includes "employment" as one of the benefits. For example, an article in the online magazine *Slate*, after discussing the costs of e-mail spam, pointed out that spam also has "a corresponding economic payoff. Anti-spam efforts keep well-paid software engineers employed."[2]

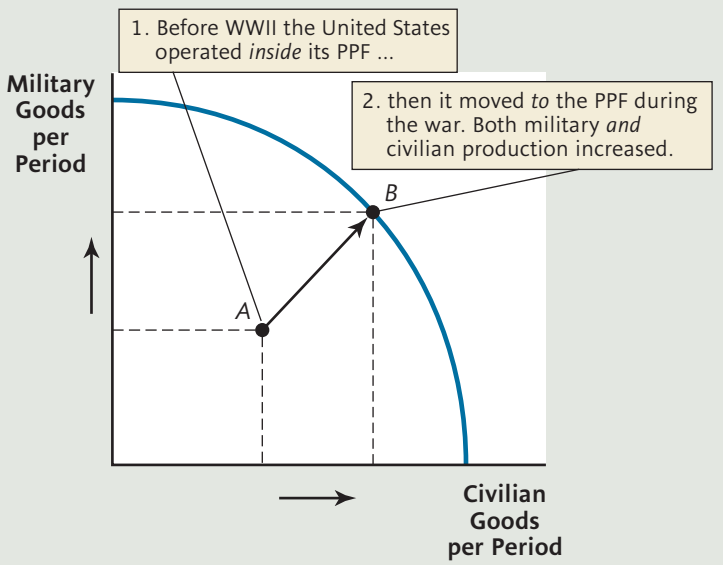
This kind of thinking is usually incorrect. True, when the economy is in recession, an increase in any kind of employment can be regarded as a benefit—especially to those who get the jobs. But this is a special, temporary situation. Once a recession ends, the software engineers—if not for the spam—would be employed elsewhere. At that point, employment in the spam-fighting industry—far from being a benefit—is actually part of the *opportunity cost* of spam: we sacrifice the goods and services these spam-fighting engineers would otherwise produce.

This simple observation can help us understand an otherwise confusing episode in U.S. economic history. During the early 1940s, after the United States entered World War II and began using massive amounts of resources to produce military goods and services, the standard of living in the United States did *not* decline as we might have expected but actually improved slightly. Why?

When the United States entered the war in 1941, it was still suffering from the Great Depression—the most serious and long-lasting economic downturn in modern history, which began in 1929 and hit most of the developed world. As we joined the allied war effort, the economy also recovered from the depression. (Indeed, for reasons you will learn when you study macroeconomics, U.S. participation in the war may have accelerated the depression's end.) So, in Figure 3, this moved our economy from a point like *A*, *inside* the PPF, to a point like *B*, *on* the frontier. Military production like tanks increased, but so did the production of civilian goods such as wheat. Although there were shortages of some consumer goods, the overall result was a rise in total production and an increase in the material well-being of the average U.S. citizen.

No government would ever *choose* war as a purely economic policy to end a downturn because other, economically superior policies could accomplish the same goal. But do these other methods of promoting recovery give us a free lunch—more of some things without any sacrifice? Not really. When you study macroeconomics you'll learn that policies to cure or avoid recessions have risks and costs of their own. Indeed, when the new Obama administration proposed some of these policies

FIGURE 3 Production and Unemployment



[2] Jeff Merron, "Workus Interruptus," *Slate*, Posted March 16, 2006, 12:06PM ET.

in early 2009, in the midst of a deep recession, a heated debate broke out about whether the costs and risks were worth it.

Economic Growth

What if the PPF itself were to change? Couldn't we then produce more of everything? This is exactly what happens when an economy's productive capacity grows.

One way that productivity capacity grows is by an increase in available resources. Historically, the resource that has contributed most to rising living standards is capital. More physical capital (factory buildings, tractors, and medical equipment) or more human capital (skilled doctors, engineers, and construction workers) can enable us to produce more of *any* goods and services that use these tools.

The other major source of economic growth is *technological change*—the discovery of new ways to produce more from a given quantity of resources. The development of the Internet, for example, enabled people to find information in a few minutes that used to require hours of searching through printed documents. As a result, a variety of professionals—teachers, writers, government officials, attorneys, and physicians—can produce more of their services with the same amount of labor hours.

Figure 4 shows three examples of how economic growth can change the PPF. Panel (a) illustrates a change that initially affects only wheat production—say, the acquisition of more tractors (usable in wheat farming but not in tank-making) or the discovery of a new, higher-yielding technique for growing wheat. If we used *all* of our resources to produce wheat, we could now produce more of it than before. For that reason, the vertical intercept of the PPF rises from point A to a point like A', where the economy could produce a maximum of 1,200,000 bushels per year. But the horizontal intercept of the PPF remains at point F, because the changes we're considering apply only to wheat. If we were to use all of our resources in tank production, we'd be able to produce the same number of tanks as before. The final effect is to stretch the PPF upward along the vertical axis.

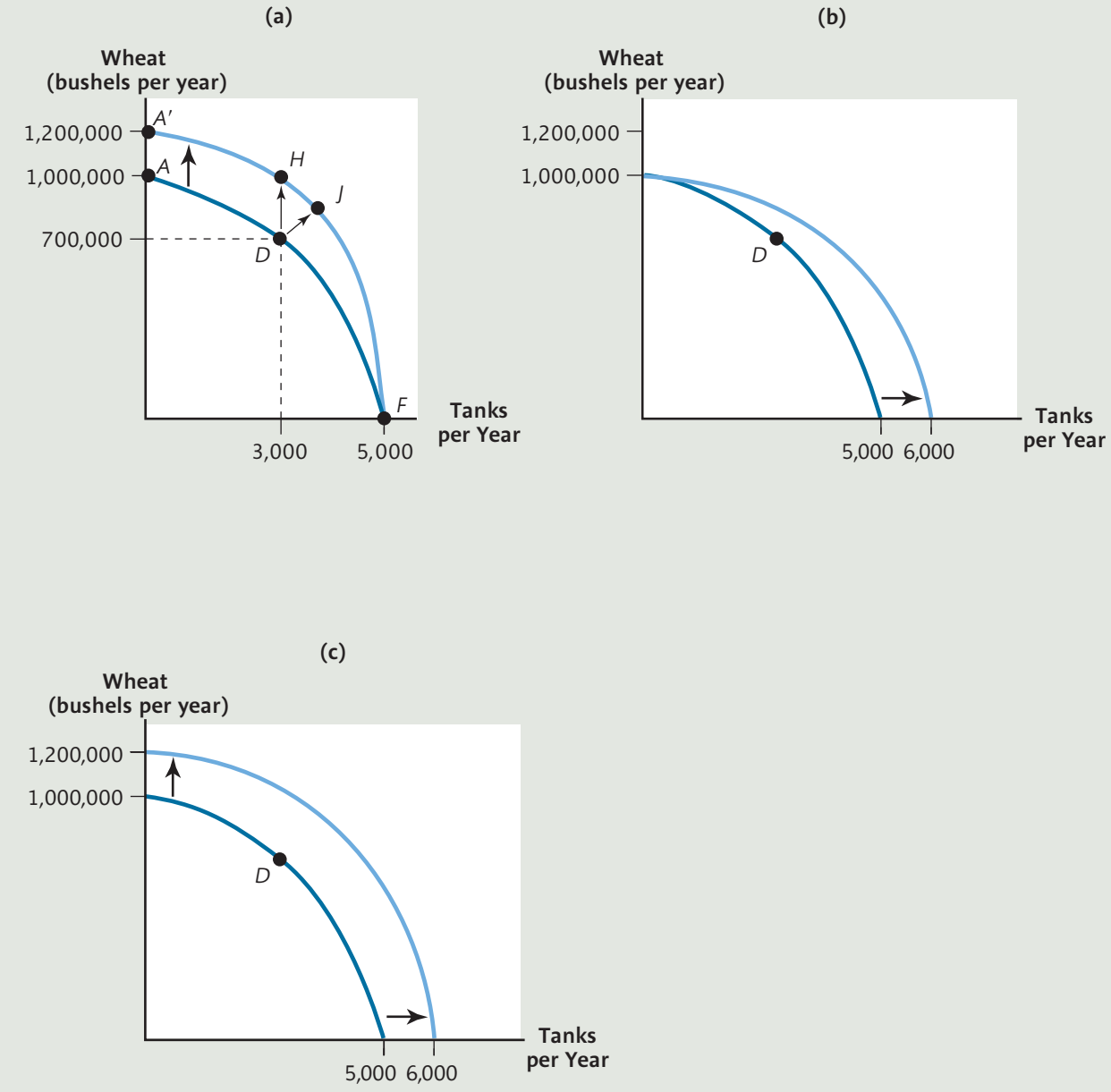
Suppose we were originally operating at point D on the old PPF. Then, with our new PPF, we could choose to produce more wheat and the same number of tanks (point H). Or we could produce more of *both* goods (point J). We could even choose to produce more tanks and the same amount of wheat as before. (See if you can identify this point on the new PPF.)

But wait . . . how can having more tractors or a new type of seed—changes that directly affect only the wheat industry—enable us to produce more tanks? The answer is: after the change in the PPF, society can choose to shift some resources out of wheat farming and have the same amount of wheat as before at point D on the original PPF. The shifted resources can be used to increase tank production.

Panel (b) illustrates the opposite type of change in the PPF—from a technological change in producing tanks, or an increase in resources usable only in the tank industry. This time, the *horizontal* intercept of the PPF increases, while the vertical intercept remains unchanged. (Can you explain why?) As before, we could choose to produce more tanks, more wheat, or more of both. (See if you can identify points on the new PPF in panel (b) to illustrate all three cases.)

Finally, panel (c) illustrates the case where technological change occurs in both the wheat and the tank industries, or there is an increase in resources (such as workers or capital) that could be used in either. Now both the horizontal and the vertical intercepts of the PPF increase. But as before, society can choose to locate anywhere along the new PPF, producing more tanks, more wheat, or more of both.

FIGURE 4 Economic Growth and the PPF



All three panels show economic growth from an increase in resources or a technological change. In panel (a), the additional resources or technological advance directly affect only wheat production. However, society can choose to have more wheat and more tanks if it desires, such as at point J. *In panel (b), the additional resources or technological advance directly affect only tank production. But once again, society can choose to have more of both goods. In panel (c), the additional resources or technological advance directly affect production of both goods.*

Panels (a) and (b) can be generalized to an important principle about economic growth:

> *A technological change or an increase in resources, even when the direct impact is to increase production of just one type of good, allows us to choose greater production of all types of goods.*

This conclusion certainly *seems* like a free lunch. But is it?

Yes . . . and no. True, comparing the new PPF to the old, it looks like we can have more of something—in fact, more of everything—without any sacrifice. But Figure 4 tells only part of the story. It leaves out the sacrifice that creates the change in the PPF in the first place.

Consumption vs. Growth

Suppose we want more capital. First, note that capital plays two roles in the economy. On the one hand, capital is a *resource* that we use to produce goods and services. On the other hand, capital is *itself* a good and is produced . . . using resources! A tractor, for example, is produced using land, labor, and *other* capital (a tractor factory and all of the manufacturing equipment inside the factory).

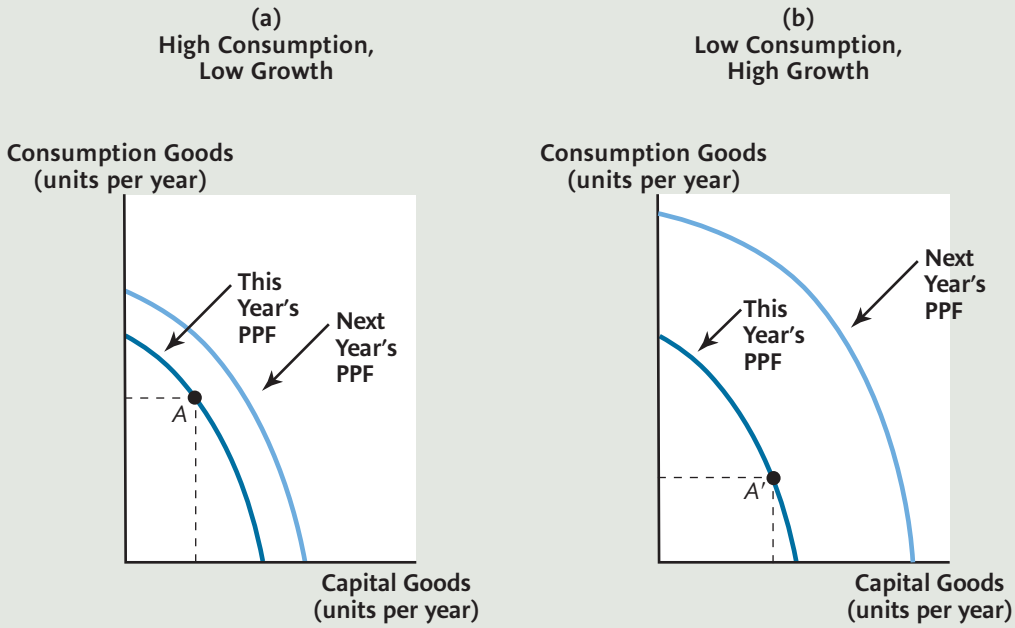
Each year, we must choose how much of our available resources to devote to producing capital, as opposed to other things. On the plus side, the more capital we produce this year, the more capital we'll have in the future to produce other things. (Remember: capital, once produced, is a long-lasting tool.) But there's a tradeoff: Any resources used to produce capital this year are *not* being used to produce *consumer goods*—food, health care, and other things we can enjoy right now.

Figure 5 illustrates this tradeoff. In each panel, the total quantity of capital goods is measured on the horizontal axis, and consumption goods on the vertical axis. In each panel, the darker curve is *this year's* PPF. In the left panel, point *A* shows one choice we could make this year: relatively high production of consumer goods and low production of capital goods. We'd have a relatively high standard of living this year (lots of consumer goods), but we'd be adding little to our total stock of capital during the year. As a result, *next year's* PPF—the lighter curve—will be shifted outward somewhat (because we'll have a bit more capital next year than we had this year), but not by much.

The right panel shows a different choice for the same economy. If we are situated at point *A′* on this year's PPF, we sacrifice more consumption goods now and produce more capital goods than at point *A* in the left panel. Living standards are lower this year. But next year, when we have considerably more capital, the PPF will have shifted outward even more. We can then choose a point on next year's PPF with greater production of consumer goods than we could have had if we had chosen point *A*. So, choosing point *A′* rather than *A* can lead to a greater rise in living standards next year, but requires greater sacrifice of consumer goods this year.

A similar tradeoff exists when technological change drives growth. New technologies don't just "happen"—resources must be used *now* for research and development. These resources could have been used to produce to other things that we'd enjoy today. For example, doctors who work at developing new drugs in pharmaceutical companies could instead be providing health care to patients right now. We could show this using the same PPFs as in Figure 5. But on the horizontal axis, instead of "capital goods," we'd have some measure of "research and development activities."

FIGURE 5 How Current Production Affects Economic Growth



In panel (a), production is tilted toward current consumption goods, with relatively few resources devoted to production of capital goods. As a result, in the future, there will not be much of an increase in capital, so the PPF will not shift out much in the future. In panel (b), production is tilted more toward capital goods, with a greater sacrifice of current consumption. As a result, there will be a greater increase in capital, so the PPF will shift out more in the future.

And we'd come to the same conclusion about technological change that we came to earlier about having more capital:

> *In order to produce more goods and services in the future, we must shift resources toward R&D and capital production, and away from producing things we'd enjoy right now.*

We must conclude that although economic growth—at first glance—*appears* to be a free lunch, someone ends up paying the check. In this case, the bill is paid by the part of society who will have to make do with less in the present.

Economic Systems

As you read these words—perhaps sitting at home or in the library—you are experiencing a very private moment. It is just you and this book; the rest of the world might as well not exist. Or so it seems. . . .

Actually, even in this supposedly private moment, you are connected to others in ways you may not have thought about. In order for you to be reading this book, the authors had to write it. Someone had to edit it, to help make sure that all necessary material was covered and explained as clearly as possible. Someone else had to prepare the graphics. Others had to run the printing presses and the binding

machines, and still others had to pack the book, ship it, unpack it, put it on a store shelf, and then sell it to you.

And there's more. People had to manufacture all kinds of goods: paper and ink, the boxes used for shipping, the computers used to keep track of inventory, and so on. It is no exaggeration to say that thousands of people were involved in putting this book in your hands.

Take a walk in your town or city, and you will see even more evidence of our economic interdependence: People are collecting garbage, helping schoolchildren cross the street, transporting furniture across town, constructing buildings, repairing roads, painting houses. Everyone is producing goods and services for *other people*.

Why is it that so much of what we consume is produced by other people? Why are we all so heavily dependent on each other for our material well-being? Why don't we all—like Robinson Crusoe on his island—produce our own food, clothing, housing, and anything else we desire? And how did it come about that *you*—who did not produce any of these things yourself—are able to consume them?

These are all questions about our *economic system*—the way our economy is organized. Ordinarily, we take our economic system for granted, like the water that runs out of our faucets. But now it's time to begin looking at the plumbing—to learn how our economy serves so many millions of people, enabling them to survive and prosper.

Specialization and Exchange

If we were forced to, many of us could become economically *self-sufficient*. We could stake out a plot of land, grow our own food, make our own clothing, and build our own homes. But in no society is there such extreme self-sufficiency. On the contrary, every economic system has been characterized by two features: (1) **specialization,** in which each of us concentrates on a limited number of productive activities, and (2) **exchange,** in which most of what we desire is obtained by trading with others rather than producing for ourselves.

Specialization A method of production in which each person concentrates on a limited number of activities.

Exchange The act of trading with others to obtain what we desire.

Specialization and exchange enable us to enjoy greater production and higher living standards than would otherwise be possible. As a result, all economies exhibit high degrees of specialization and exchange.

Where do the gains from specialization and exchange come from?

Development of Expertise. Each of us can learn only so much in a lifetime. By limiting ourselves to a narrow set of tasks—fixing plumbing, managing workers, writing music, or designing Web pages—we can hone our skills. We can become experts at one or two things, instead of remaining amateurs at a lot of things. An economy of experts will produce more than an economy of amateurs.

Minimizing Downtime. When people specialize, and thus spend their time doing one type of task, there is less unproductive "downtime" from switching activities. On a smaller scale, we see this in most households every night, when it's time to wash the dishes. "I'll wash, you dry" enables both members of a couple to keep doing their part nonstop, and the dishes get done faster.

Comparative Advantage. We would gain from developing expertise and minimizing downtime even if everyone, initially, had identical capabilities. But a third gain arises because of our *differences:* we are not all equally suited to different tasks. This

TABLE 2

Labor Requirements for Fish and Berries

	Labor Required For:	
	1 Fish	1 Cup of Berries
Maryanne	1 hour	1 hour
Gilligan	3 hours	1½ hours

idea applies not just to labor, but to the other resources as well. Not all plots of land are equally suited for all crops, nor are all types of capital equipment equally suited for all types of production. When we allocate our resources according to their *suitability* for different types of production, we get further gains from specialization. The principle behind this gain—*comparative advantage*—is such a central idea in economics that it gets its own section.

Comparative Advantage

Imagine a shipwreck in which there are only two survivors—let's call them Maryanne and Gilligan—who wash up on opposite shores of a deserted island. Initially they are unaware of each other, so each is forced to become completely self-sufficient. And there are only two kinds of food on the island: fish and berries.

Table 2 shows how much time it takes for each castaway to pick a cup of berries or catch one fish. For simplicity, we'll assume that the time requirement remains constant no matter how much time is devoted to these activities.

On one side of the island, Maryanne finds that it takes her 1 hour to catch a fish and 1 hour to pick one cup of berries, as shown in the first row of the table. On the other side of the island, Gilligan—who is less adept at both tasks—requires 3 hours to catch a fish and 1½ hours to pick a cup of berries, as listed in the second row of the table. Since both castaways would want some variety in their diets, we can assume that each would spend part of the week catching fish and part picking berries.

Suppose that, one day, Maryanne and Gilligan discover each other. After rejoicing at the prospect of human companionship, they decide to develop a system of production that will work to their mutual benefit. Let's rule out any of the gains from specialization that we discussed earlier (minimizing downtime or developing expertise). Will it still pay for these two to specialize? The answer is yes, as you will see after a small detour.

Absolute Advantage: A Detour

When Gilligan and Maryanne sit down to figure out who should do what, they might fall victim to a common mistake: basing their decision on *absolute advantage*.

Absolute advantage The ability to produce a good or service, using fewer resources than other producers use.

An individual has an ***absolute advantage*** *in the production of some good when he or she can produce it using fewer resources than another individual can.*

On the island, labor is the only resource. The castaways might (mistakenly) reason as follows: Maryanne can catch a fish more quickly than Gilligan (see Table 2), so she has an *absolute advantage* in fishing. Therefore, Maryanne should be the one to catch fish.

But wait! Maryanne can also pick berries more quickly than Gilligan, so she has an absolute advantage in that as well. If absolute advantage is the criterion for

assigning work, then Maryanne should do *both* tasks. This, however, would leave Gilligan doing nothing, which is certainly *not* in the pair's best interests.

What can we conclude from this example? That absolute advantage is an unreliable guide for allocating tasks to different workers.

Comparative Advantage

The correct principle to guide the division of labor on the island is comparative advantage:

> *A person has a* ***comparative advantage*** *in producing some good if he or she can produce it with a smaller opportunity cost than some other person can.*

Comparative advantage The ability to produce a good or service at a lower opportunity cost than other producers.

Notice the important difference between absolute advantage and comparative advantage: You have an *absolute* advantage in producing a good if you can produce it using fewer *resources* than someone else can. But you have a *comparative* advantage if you can produce it with a smaller *opportunity cost*. As you'll see, these are not necessarily the same thing.

Who Has a Comparative Advantage in What?

Let's first see who has a comparative advantage in fishing. To do this, we need to calculate the opportunity cost—for each castaway—of catching one more fish.

For Maryanne, catching one more fish takes one more hour. That requires one hour less picking berries and that, in turn, means sacrificing 1 cup of berries. We can summarize it this way:

Maryanne: 1 more fish ⇒ 1 more hour fishing
⇒ 1 less hour picking berries ⇒ 1 less cup berries

Therefore, for Maryanne: *the opportunity cost of one more fish is 1 cup of berries*.

Doing the same analysis for Gilligan, who needs three hours to catch a fish, we find:

Gilligan: 1 more fish ⇒ 3 more hours fishing
⇒ 3 fewer hours picking berries ⇒ 2 fewer cups berries

So for Gilligan, the *opportunity cost of one more fish is 2 cups of berries*. Because Maryanne has the lower opportunity cost for one more fish (1 cup of berries rather than 2 cups), *Maryanne has the comparative advantage in fishing*.

Now let's see who has the comparative advantage in picking berries. We can summarize the steps as follows:

Maryanne: 1 more cup berries ⇒ 1 more hour picking berries
⇒ 1 less hour fishing ⇒ 1 less fish

Gilligan: 1 more cup berries ⇒ 1½ more hours picking berries
⇒ 1½ fewer hours fishing ⇒ ½ less fish

You can see that the opportunity cost of one more cup of berries is 1 fish for Maryanne, and ½ fish for Gilligan.[3] When it comes to berries, Gilligan has the lower

Even castaways do better when they specialize and exchange with each other, instead of trying to be self-sufficient.

[3] Of course, no one would ever catch half a fish unless they were using a machete. The number just tells us the rate of tradeoff of one good for the other.

TABLE 3

A Beneficial Change in Production

	Change in Fish Production	Change in Berry Production
Maryanne	+1	−1
Gilligan	−1	+2
Total Island	+0	+1

opportunity cost. So Gilligan—who has an *absolute* advantage in nothing—has a *comparative* advantage in berry-picking.

Gains from Comparative Advantage

What happens when each castaway produces more of their comparative advantage good? The results are shown in Table 3. In the first row, we have Maryanne catching one more fish each day (+1) at an opportunity cost of one cup of berries (−1). In the second row, we have Gilligan producing one fewer fish (−1), which enables him to produce two more cups of berries (+2).

Now look at the last row. It shows what has happened to production of both goods on the island as a result of this little shift between the two. While fish production remains unchanged, berry production rises by one cup. If the castaways specialize and trade with each other, they can both come out ahead: consuming the same quantity of fish as before, but more berries.

As you can see in Table 3, when each castaway moves toward producing more of the good in which he or she has a *comparative advantage*, total production rises. Now, let's think about this. If they gain by making this small shift toward their comparative advantage goods, why not make the change again? And again after that? In fact, why not keep repeating it until the opportunities for increasing total island production are exhausted, which occurs when one or both of them is devoting all of their time to producing just their comparative advantage good, and none of the other? In the end, specializing according to comparative advantage and exchanging with each other gives the castaways a higher standard of living than they each could achieve on their own.

Beyond the Island

What is true for our shipwrecked island dwellers is also true for the entire economy:

> *Total production of every good or service will be greatest when individuals specialize according to their comparative advantage. This is another reason why specialization and exchange lead to higher living standards than does self-sufficiency.*

When we turn from our fictional island to the real world, is production, in fact, consistent with the principle of comparative advantage? Indeed, it is. A journalist may be able to paint her house more quickly than a house painter, giving

her an *absolute* advantage in painting her home. Will she paint her own home? Except in unusual circumstances, no, because the journalist has a *comparative* advantage in writing news articles. Indeed, most journalists—like most college professors, attorneys, architects, and other professionals—hire house painters, leaving more time to practice the professions in which they enjoy a comparative advantage.

Even fictional superheroes seem to behave consistently with comparative advantage. Superman can no doubt cook a meal, fix a car, chop wood, and do virtually *anything* faster than anyone else on the earth. Using our new vocabulary, we'd say that Superman has an absolute advantage in everything. But he has a clear *comparative* advantage in catching criminals and saving the universe from destruction, which is exactly what he spends his time doing.

International Comparative Advantage

You've seen that comparative advantage is one reason people *within* a country can benefit from specializing and trading with each other. The same is true for trading among *nations*. We say that

> *A nation has a comparative advantage in producing a good if it can produce it at a* lower opportunity cost *than some other nation.*

To illustrate this, let's consider a hypothetical world that has only two countries: the United States and China. Both are producing only two goods: soybeans and T-shirts. And—to keep our model simple—we'll assume that these goods are being produced with just one resource: labor.

Table 4 shows the amount of labor, in hours, required to produce one bushel of soybeans or one T-shirt in each country. We'll assume that hours per unit remain *constant,* no matter how much of a good is produced. For example, the entry "5 hours" tells us that it takes 5 hours of labor to produce one bushel of soybeans in China. This will be true no matter how many bushels China produces.

In the table, the United States has an *absolute advantage* in producing both goods. That is, it takes fewer resources (less labor time) to produce either soybeans or T-shirts in the United States than in China. But—as you are about to see—China has a *comparative* advantage in one of these goods.

Determining a Nation's Comparative Advantage

Just as we did for our mythical island, we can determine comparative advantage for a country by looking at opportunity cost.

TABLE 4

Labor Requirements for Soybeans and T-Shirts

	Labor Required For:	
	1 Bushel of Soybeans	**1 T-Shirt**
United States	½ hour	¼ hour
China	5 hours	1 hour

Let's first look at the opportunity cost of producing, say, 10 more bushels of soybeans. (Using 10 bushels rather than just one bushel enables us to use round numbers in this example.) For the United States, these 10 bushels would require 5 hours (10 bushels, each requiring ½ hour). That means 5 hours less spent producing T-shirts. Since each T-shirt requires ¼ hour, 5 fewer hours means sacrificing 20 T-shirts. We can summarize it this way:

United States: 10 more bshls soybeans ⇒ 5 more hours producing soybeans
⇒ 5 fewer hours producing T-shirts
⇒ 20 fewer T-shirts

Therefore, for the U.S. the opportunity cost of 10 more bushels of soybeans is 20 T-shirts.

Doing the same analysis for China, where it takes 5 hours to produce each bushel of soybeans, we find:

China: 10 more bshls soybeans ⇒ 50 more hours producing soybeans
⇒ 50 fewer hours producing T-shirts
⇒ 50 fewer T-shirts

So for China, the opportunity cost of 10 bushels of soybeans is 50 T-shirts. Since the U.S. has the lower opportunity cost for soybeans (20 T-shirts rather than 50), the *U.S. has the comparative advantage in soybeans*.

Now let's see who has the comparative advantage in T-shirts. We can choose any number of T-shirts to start off our analysis, so let's (arbitrarily) consider the opportunity cost of producing 40 more T-shirts:

United States: 40 more T-shirts ⇒ 10 more hours producing T-shirts
⇒ 10 fewer hours producing soybeans
⇒ 20 fewer bshls soybeans

China: 40 more T-shirts ⇒ 40 more hours producing T-shirts
⇒ 40 fewer hours producing soybeans
⇒ 8 fewer bshls soybeans

To sum up: the opportunity cost of 40 more T-shirts is 20 bushels of soybeans for the U.S., but only 8 bushels for China. Since China has the lower opportunity cost, China—which has an *absolute* advantage in neither good—has a *comparative* advantage in T-shirts.

Global Gains from Comparative Advantage

What happens when each nation produces more of its comparative advantage good? The results, for our example, are shown in Table 5. In the first row, we have the U.S.

TABLE 5

A Beneficial Change in World Production

	Soybeans (Bushels)	T-Shirts
United States	+10	−20
China	−8	+40
Total World Production	**+2**	**+20**

producing 10 more bushels of soybeans (+10) at an opportunity cost of 20 T-shirts (−20). In the second row, China produces 8 fewer bushels of soybeans (−8), and 40 more T-shirts (+40).

As you see in our example, world production of *both* goods increases: 2 more bushels of soybeans and 20 more T-shirts. With greater world production, each country can enjoy a higher standard of living. How will the potential gains in living standards be shared among the two countries? That depends on how many T-shirts trade for a bushel of soybeans in the international market where these goods are traded. You'll learn more about this—and other aspects of international trade—in a later chapter. But you can already see how this example illustrates a general point:

> *Total production of every good or service is greatest when* nations *shift production toward their comparative advantage goods, and trade with each other.*

Resource Allocation

More than ten thousand years ago, the Neolithic age began and human society switched from hunting and gathering to farming and simple manufacturing. At the same time, human wants grew beyond mere food and shelter to the infinite variety of things that can be *made*. Ever since, all societies have been confronted with three important questions:

1. ***Which* goods and services should be produced with society's resources?**
   Should we produce more consumer goods for enjoyment now or more capital goods to increase future production? Should we produce more health care, and if so, what should we produce less of? In other words, where on its production possibilities frontier should the economy operate?
2. ***How* should they be produced?**
   Most goods and services can be produced in a variety of different ways, with each method using more of some resources and less of others. For example, there are many ways to dig a ditch. We could use *no capital at all* and have dozens of workers digging with their bare hands. We could use *a small amount of capital* by giving each worker a shovel and thereby use less labor, since each worker would now be more productive. Or we could use *even more capital*—a power trencher—and dig the ditch with just one or two workers. In every economic system, there must always be some mechanism that determines *how* goods and services will be produced from the many options available.
3. ***Who* should get them?**
   This is where economics interacts most strongly with politics. There are so many ways to divide ourselves into groups: men and women, rich and poor, skilled and unskilled, workers and owners, families and single people, young and old . . . the list is almost endless. How should the products of our economy be distributed among these different groups and among individuals within each group? Determining *who* gets the economy's output is always the most controversial aspect of resource allocation. Over the last half-century, our society has become more sensitized to the way goods and services are distributed, and we increasingly ask whether that distribution is fair.

The Three Methods of Resource Allocation

Throughout history, every society has relied primarily on one of three mechanisms for allocating resources. In a **traditional economy,** resources are allocated according to the long-lived practices of the past. Tradition was the dominant method of

Traditional economy An economy in which resources are allocated according to long-lived practices from the past.

resource allocation for most of human history and remains strong in many tribal societies and small villages in parts of Africa, South America, Asia, and the Pacific. Typically, traditional methods of production are handed down by the village elders, and traditional principles of fairness govern the distribution of goods and services.

Economies in which resources are allocated mostly by tradition tend to be stable and predictable. But these economies have one serious drawback: They don't grow. With everyone locked into the traditional patterns of production, there is little room for innovation and technological change. Traditional economies are therefore likely to be stagnant economies.

Command or centrally planned economy An economic system in which resources are allocated according to explicit instructions from a central authority.

In a **command economy,** resources are allocated mostly by explicit instructions from some higher authority.

Because the government must plan these instructions in advance, command economies are also called **centrally planned economies.** But command economies are disappearing fast. Until about 25 years ago, examples would have included the former Soviet Union, Poland, Romania, Bulgaria, Albania, China, and many others. Beginning in the late 1980s, all of these nations began abandoning central planning. The only examples left today are Cuba and North Korea, and even these economies—though still dominated by central planning—occasionally take steps away from it.

Market economy An economic system in which resources are allocated through individual decision making.

The third method of allocating resources—and the one with which you are no doubt most familiar—is "the market." In a **market economy,** instead of following long-held traditions or commands from above, people are largely free to do what they want with the resources at their disposal. In the end, resources are allocated as a result of individual decision making. *Which* goods and services are produced? The ones that producers *choose* to produce. *How* are they produced? However producers *choose* to produce them. *Who* gets these goods and services? Anyone who *chooses* to buy them.

Of course, in a market economy, freedom of choice is constrained by the resources one controls. And in this respect, we do not all start in the same place in the economic race. Some of us have inherited great intelligence, talent, or beauty; and some, such as the children of successful professionals, are born into a world of helpful personal contacts. Others, unfortunately, will inherit none of these advantages. In a market system, those who control more resources will have more choices available to them than those who control fewer resources. Nevertheless, given these different starting points, individual choice plays the major role in allocating resources in a market economy.

But wait . . . isn't there a problem here? People acting according to their own desires, without command or tradition to control them? This sounds like a recipe for chaos! How, in such a free-for-all, could resources possibly be *allocated*?

The answer is contained in two words: *markets* and *prices*.

The Nature of Markets

The market economy gets its name from something that nearly always happens when people are free to do what they want with the resources they possess. Inevitably, people decide to specialize in the production of one or a few things—often organizing themselves into business firms—and then sellers and buyers *come together to trade*. A **market** is a collection of buyers and sellers who have the potential to trade with one another.

Market A group of buyers and sellers with the potential to trade with each other.

In some cases, the market is *global;* that is, the market consists of buyers and sellers who are spread across the globe. The market for oil is an example of a global market, since buyers in any country can buy from sellers in any country. In other cases, the market is local. Markets for restaurant meals, haircuts, and taxi service are examples of local markets.

Markets play a major role in allocating resources by forcing individual decision makers to consider very carefully their decisions about buying and selling. They do so because of an important feature of every market: the *price* at which a good is bought and sold.

The Importance of Prices

A **price** is *the amount of money a buyer must pay to a seller for a good or service*. Price is not always the same as *cost*. In economics, as you've learned in this chapter, cost means *opportunity cost*—the *total* sacrifice needed to buy the good. While the price of a good is a *part* of its opportunity cost, it is not the only cost. For example, the price does not include the value of the time sacrificed to buy something. Buying a new jacket will require you to spend time traveling to and from the store, trying on different styles and sizes, and waiting in line at the cash register.

Price The amount of money that must be paid to a seller to obtain a good or service.

Still, in most cases, the price of a good is a significant part of its opportunity cost. For large purchases such as a home or automobile, the price will be *most* of the opportunity cost. And this is why prices are so important to the overall working of the economy: They confront individual decision makers with the costs of their choices.

Consider the example of purchasing a car. Because you must pay the price, you know that buying a new car will require you to cut back on purchases of other things. In this way, the opportunity cost to *society* of making another car is converted to an opportunity cost *for you*. If you value a new car more highly than the other things you must sacrifice for it, you will buy it. If not, you won't buy it.

Why is it so important that people face the opportunity costs of their actions? The following thought experiment can answer this question.

A Thought Experiment: Free Cars

Imagine that the government passes a new law: When anyone buys a new car, the government will reimburse that person for it immediately. The consequences would be easy to predict.

First, on the day the law was passed, everyone would rush out to buy new cars. Why not, if cars are free? The entire stock of existing automobiles would be gone within days—maybe even hours. Many people who didn't value cars much at all, and who seldom used them, would find themselves owning several—one for each day of the week or to match the different colors in their wardrobe. Others who weren't able to act in time—including some who desperately needed a new car for their work or to run their households—would be unable to find one at all.

Over time, automobile companies would drastically increase production to meet the surge in demand for cars. So much of our available labor, capital, land, and entrepreneurial talent would be diverted to making cars that we'd have to sacrifice huge quantities of all other goods and services. Thus, we'd end up *paying* for those additional cars in the end, by having less education, less medical care, perhaps even less food—all to support the widespread, frivolous use of cars. Almost everyone would conclude that society had been made worse off with the new "free-car" policy. By eliminating a price for automobiles, and severing the connection between society's opportunity cost of producing cars and individuals' decisions to have them, we would have created quite a mess for ourselves.

When resources are allocated by the market, and people must pay for their purchases, they are forced to consider the opportunity cost to society of their individual actions. In this way, markets are able to create a sustainable allocation of resources.

Resource Allocation in the United States

The United States has always been considered the leading example of a market economy. Each day, millions of distinct items are produced and sold in markets. Our grocery stores are always stocked with broccoli and tomato soup, and the drugstore always has Kleenex and aspirin—all due to the choices of individual producers and consumers. The goods that are traded, the way they are traded, and the price at which they trade are determined by the traders themselves.

But even in the United States, there are numerous cases of resource allocation *outside* the market. For example, many economic decisions are made within families, which do *not* operate like little market economies. Instead, many decisions are based on tradition. For example, even when children get an allowance or have other earnings, they don't have to pay for goods consumed within the home. Other decisions are based on command ("No TV until you finish your homework!").

In the broader economy, there are many examples of resource allocation by command. Various levels of government collect, in total, about one-third of our incomes as taxes. We are *told* how much tax we must pay, and those who don't comply suffer serious penalties, including imprisonment. Government—rather than individual decision makers—spends the tax revenue. In this way, the government plays a major role in allocating resources—especially in determining which goods are produced and who gets them.

There are also other ways, aside from strict commands, that the government limits our market freedoms. Regulations designed to protect the environment, maintain safe workplaces, and ensure the safety of our food supply are just a few examples of government-imposed constraints on our individual choice.

Why, then, do we say that the U.S. is "a market economy"? Because for each example we can find where resources are allocated by tradition or command, or where government restrictions seriously limit some market freedom, we can find hundreds of examples where individuals make choices according to their own desires. The things we buy, the jobs at which we work, the homes in which we live—in almost all cases, these result from market choices. The market, though not pure, is certainly the dominant method of resource allocation in the United States.

The complete label for the economic system in the U.S. (and—to varying degrees—in most other countries as well) is *market capitalism*. While the market describes how resources are *allocated,* capitalism refers to one way that resources are *owned*. Under **capitalism,** most resources are owned by private citizens, who are mostly free to sell or rent them to others as they wish. The alternative mode of ownership is **socialism,** a system in which most resources are owned by the state, as in the former Soviet Union.

Capitalism A type of economic system in which most resources are owned privately.

Socialism A type of economic system in which most resources are owned by the state.

Just as the U.S. is a leading example of resource allocation by the market, it is also a leading example of capitalism. True, there are examples of state ownership of resources (national parks, government buildings, state highways systems, and more). But most of our nation's land, labor, and capital are privately owned and managed, and can be sold or rented in markets as the owners wish.

Understanding the Market

The market is simultaneously the most simple and the most complex way to allocate resources. For individual buyers and sellers, the market is simple. There are no traditions or commands to be memorized and obeyed. Instead, we enter the markets we *wish* to trade in, and we respond to prices there as we *wish* to, unconcerned about the overall process of resource allocation.

But from the economist's point of view, the market is quite complex. Resources are allocated indirectly, as a *byproduct* of individual decision making, rather than through easily identified traditions or commands. As a result, it often takes some skillful economic detective work to determine just how individuals are behaving and how resources are being allocated as a consequence.

How can we make sense of all of this apparent chaos and complexity? That is what economics is all about. You will begin your detective work in Chapter 3, where you will learn about the most widely used model in the field of economics: the model of supply and demand.

Using the Theory

© PHOTODISC/GETTY IMAGES

ARE WE SAVING LIVES EFFICIENTLY?

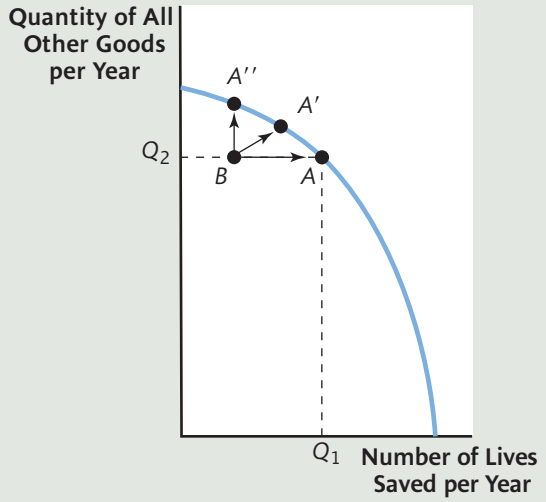
Earlier in this chapter, you learned that instances of gross productive inefficiency are not as easy to find in our economy as one might imagine. But many economists argue that our allocation of resources to lifesaving efforts is a glaring exception. In this section, we'll use some of the tools and concepts you've learned in this chapter to ask whether we are saving lives efficiently.

We can view "saving lives" as the output—a service—produced by the "lifesaving industry." This industry consists of private firms (such as medical practices and hospitals), as well as government agencies (such as the Department of Health and Human Services or the Environmental Protection Agency). In a productively efficient economy, we must pay an opportunity cost whenever we choose to save additional lives. That's because saving more lives—by building another emergency surgery center, running an advertising campaign to encourage healthier living, or requiring the use of costly but safe materials instead of cheaper but toxic ones—requires additional land, labor, capital, and entrepreneurship. And these resources could be used to produce other goods and services that we value.

Figure 6 illustrates this opportunity cost with a production possibilities frontier. The number of lives saved per year is measured along the horizontal axis, and the quantity of all other goods (lumped together into a single category) is measured on the vertical axis. A productively efficient economy would be *on* the frontier, producing the maximum quantity of all other goods for any given number of lives saved. Equivalently, productive efficiency would mean saving the maximum possible number of lives for any given quantity of other goods. Point *A* on the PPF is one such productively efficient point, where we would save Q_1 lives per year, and produce the quantity Q_2 of all other goods. Once we are on the frontier, we can only save more lives by pulling resources away from producing other goods, and paying an opportunity cost in other goods foregone.

FIGURE 6 Efficiency and Inefficiency in Saving Lives

This PPF shows society's choice between saving lives (measured along the horizontal axis) and all other production (on the vertical axis). Operating on the curve (at points like A, A′, *or* A″*) would be productively efficient. But if the life-saving industry is not efficient, then society is operating inside the PPF (at a point like* B*). Eliminating the inefficiency would enable us to save more lives, or have more of other goods, or both.*



But what if there is productive *in*efficiency in the economy? And what if the source of the inefficiency is in the lifesaving industry itself? More specifically, what if we could save more lives using the current quantity of resources used by the industry, simply by reallocating those resources among different types of lifesaving activities? In that case, we would be currently operating at a point like *B*, *inside* the PPF. By eliminating the inefficiency, we could move *to* the frontier. For example, we could save more lives with no sacrifice of other goods (a move from point *B* to point *A*) or have more of other goods while saving the same number of lives (a move to point *A″*) or have more of both (point *A′*).

Economists argue that the United States and most other countries do, in fact, operate at a point like *B* because of productive inefficiency in saving lives. How have they come to such a conclusion?

The first step in the analysis is to remember that, in a market economy, resources sell at a price. We can use the dollar cost of a lifesaving method to measure the value of the resources used up by that method.

Moreover, we can compare the "cost per year of life saved" of different methods. For example, in the United States we currently spend about $253 million on heart transplants each year and thereby add about 1,600 years to the lives of heart patients. Thus, the cost per year of life saved from heart transplants is $253,000,000/1,600 = $158,000 (rounded to the nearest thousand).

Table 6 lists several of the methods we currently use to save lives in the United States. Some of these methods reflect legal or regulatory decisions (such as the ban on asbestos) and others are standard medical practices (such as annual mammograms for women over 50). Other methods effectively save lives only sporadically (such as seat belts in school buses). You can see that the cost per life saved ranges widely—from $150 per year of life saved for a physician warning a patient to quit

TABLE 6

The Cost of Saving Lives

Method	Cost per Life-Year Saved
Brief physician antismoking intervention:	
Single personal warning from physician to stop smoking	$150
Sickle cell screening and treatment for African-American newborns	$236
Replacing ambulances with helicopters for medical emergencies	$2,454
Intensive physician antismoking intervention:	
Physician identification of smokers among their patients; three physician counseling sessions; two further sessions with smoking-cessation specialists; and materials—nicotine patch or nicotine gum	$2,587
Mammograms: Once every three years, for ages 50–64	$2,700
Chlorination of water supply	$4,000
Next step after suspicious lung X-ray:	
PET Scan	$3,742
Exploratory Surgery	$4,895
Needle Biopsy	$7,116
Vaccination of all infants against strep infections	$80,000
Mammograms: Annually, for ages 50–64	$108,401
Exercise electrocardiograms as screening test:	
For 40-year-old males	$124,374
Heart transplants	$157,821
Mammograms: Annually, for ages 40–49	$186,635
Exercise electrocardiograms as screening test:	
For 40-year-old females	$335,217
Seat belts on school buses	$2,760,197
Asbestos ban in automatic transmissions	$66,402,402

Sources: Compiled from various publications. Individual sources available from authors upon request.

smoking, to over $66,000,000 per year of life saved from the ban on asbestos in automatic transmissions.

The table indicates that some lifesaving methods are highly cost effective. For example, our society probably exhausts the potential to save lives from brief physician antismoking intervention. Most doctors *do* warn their smoking patients to quit.

But the table also indicates some serious productive *in*efficiency in lifesaving. For example, screening and treating African-American newborns for sickle cell anemia is one of the least costly ways of saving a year of life in the United States—only $236 per year of life saved. Nevertheless, 20 percent of African-American newborns do *not* get this screening at all. Similarly, intensive intervention to discourage smoking is far from universal in the U.S. health care system, even though it has the relatively low cost of $2,587 per year of life saved.

Why is the less than universal use of these lower cost methods *productively inefficient?* To answer, let's do some thought experiments. First, let's imagine that we

shift resources from heart transplants to *intensive* antismoking efforts. Then for each year of life we decided *not* to save with heart transplants, we would free up $157,821 in medical resources. If we applied those resources toward intensive antismoking efforts, at a cost of $2,587 per year of life saved, we could then save an additional $157,821/$2,587 = 61 life-years. In other words, we could increase the number of life-years saved without any increase in resources flowing to the health care sector, and therefore, without any sacrifice in other goods and services. If you look back at the definition of productive inefficiency given earlier in this chapter, you'll see why this is an example of it.

But why pick on heart transplants? Our ban on asbestos in automobile transmissions—which requires the purchase of more costly materials made with greater quantities of scarce resources—costs us about $66 million for each life-year saved. Suppose these funds were spent instead to buy the resources needed to provide women aged 40 to 49 with annual mammograms (currently *not* part of most physicians' recommendations). Then for each life-year lost to asbestos, we'd save $66 million/$186,635 = 354 life-years from earlier detection of breast cancer.

Of course, allocating lifesaving resources is much more complicated than our discussion so far has implied. For one thing, the benefits of lifesaving efforts are not fully captured by "life-years saved" (or even by an alternative measure, which accounts for improvement in *quality* of life). The cost per life-year saved from mandating seat belts on school buses is extremely high—almost $3 million. This is mostly because very few children die in school bus accidents—about 11 per year in the entire United States—and, according to the National Traffic Safety Board, few of these deaths would have been prevented with seat belts. But mandatory seat belts—rightly or wrongly—might decrease the anxiety of millions of parents as they send their children off to school. How should we value such a reduction in anxiety? Hard to say. But it's not unreasonable to include it as a benefit—at least in some way—when deciding about resources.

SUMMARY

The *production possibilities frontier* (PPF) is a simple model to illustrate the opportunity cost of society's choices. When we are *productively efficient* (operating *on* the PPF), producing more of one thing requires producing less of something else. The *law of increasing opportunity cost* tells us that the more of something we produce, the greater the opportunity cost of producing still more. Even when we are operating inside the PPF—say because of productive inefficiency or a recession—it is not necessarily easy or costless to move to the PPF and avoid opportunity cost.

In a world of scarce resources, each society must have an *economic system*: its way of organizing economic activity. All economic systems feature *specialization*, in which each person and firm concentrates on a limited number of productive activities, and exchange, through which we obtain most of what we desire by trading with others. Specialization and exchange enable us to enjoy higher living standards than would be possible under self-sufficiency.

One way that specialization increases living standards is by allowing each of us to concentrate on tasks in which we have a *comparative advantage*. When individuals within a country produce more of their comparative advantage goods and exchange with others, living standards in that country are higher. Similarly, when individual nations specialize in their comparative advantage goods and trade with other nations, global living standards rise.

Every economic system determines how resources are allocated. In a *market economy,* resources are allocated primarily through markets. Prices play an important role in markets by forcing decision makers to take account of society's opportunity cost when they make choices.

PROBLEM SET

Answers to even-numbered Questions and Problems can be found on the text website at www.cengage.com/economics/hall.

1. Redraw Figure 1, but this time identify a different set of points along the frontier. Starting at point *F* (5,000 tanks, zero production of wheat), have each point you select show equal increments in the quantity of wheat produced. For example, a new point *H* should correspond to 200,000 bushels of wheat, point *J* to 400,000 bushels, point *K* to 600,000 bushels, and so on. Now observe what happens to the opportunity cost of "200,000 more bushels of wheat" as you move leftward and upward along this PPF. Does the law of increasing opportunity cost apply to the production of wheat? Explain briefly.
2. How would a technological innovation in lifesaving—say, the discovery of a cure for cancer—affect the PPF in Figure 6?
3. How would a technological innovation in the production of *other* goods—say, the invention of a new kind of robot that speeds up assembly-line manufacturing—affect the PPF in Figure 6?
4. Suppose the Internet enables more production of other goods *and* helps to save lives (for simplicity, assume proportional increases). Show how the PPF in Figure 6 would be affected.
5. Suppose that one day, Gilligan (the castaway) eats a magical island plant that turns him into an expert at everything. In particular, it now takes him just half an hour to pick a quart of berries, and 15 minutes to catch a fish.
   a. Redo Table 2 in the chapter.
   b. Who—Gilligan or Maryanne—has a comparative advantage in picking berries? In fishing?
   c. Suppose that Gilligan reallocates his time to produce *two more units* of his comparative advantage good, and Maryanne does the same. Construct a new version of Table 3 in the chapter showing how production changes for each castaway, and for the island as a whole.
6. Suppose that two different castaways, Mr. and Mrs. Howell, end up on a different island. Mr. Howell can pick 1 pineapple per hour, or 1 coconut. Mrs. Howell can pick 2 pineapples per hour, but it takes her two hours to pick a coconut.
   a. Construct a table like Table 2 showing Mr. and Mrs. Howell's labor requirements.
   b. Who—Mr. or Mrs. Howell—has a comparative advantage in picking pineapples? In picking coconuts? Which of the two should specialize in which tasks?
   c. [Harder] Assume that Mr. and Mrs. Howell had originally washed ashore on different parts of the island, and that they originally each spent 12 hours per day working, spending 6 hours picking pineapples and 6 hours picking coconuts. How will their total production change if they find each other and begin to specialize?
7. You and a friend have decided to work jointly on a course project. Frankly, your friend is a less-than-ideal partner. His skills as a researcher are such that he can review and outline only two articles a day. Moreover, his hunt-and-peck style limits him to only 10 pages of typing a day. On the other hand, in a day you can produce six outlines or type 20 pages.
   a. Who has an absolute advantage in outlining, you or your friend? What about typing?
   b. Who has a comparative advantage in outlining? In typing?
   c. According to the principle of comparative advantage, who should specialize in which task?
8. Suppose that an economy's PPF is a straight line, rather than a bowed out, concave curve. What would this say about the nature of opportunity cost as production is shifted from one good to the other?

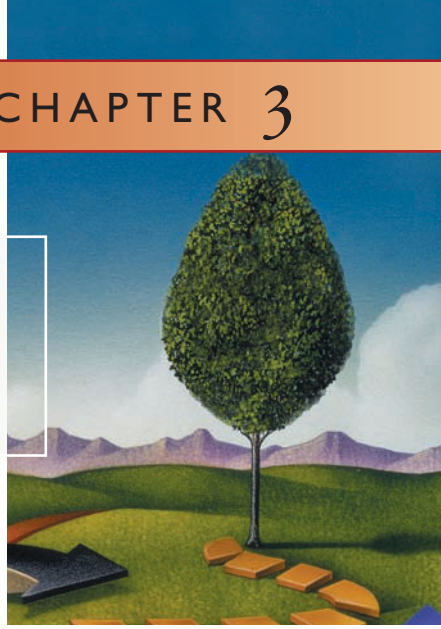
More Challenging

9. Go back to Table 5 in the chapter, which is based on the hours requirements in Table 4. Suppose that when trade opens up between the U.S. and China, the U.S. increases its production of soybeans by 100 bushels (instead of 10 as in the table). China increases its production of T-shirts by 400 (instead of 40). Assume that when the two countries trade with each other, each bushel of soybeans is exchanged for 3 T-shirts. Finally, suppose that the U.S. trades (exports) 90 bushels of soybeans to China.
   a. How many T-shirts from China will the U.S. receive in exchange for its soybeans exports to China?
   b. After trading with China, how many more bushels of soybeans will be available for Americans to consume (compared to the situation before trade)? How many more T-shirts?

c. After trading with the U.S., how many more bushels of soybeans will be available for the Chinese to consume? How many more T-shirts?
d. Based on this example, consider the following statement: "When two countries trade with each other, one country's gain will always be the other country's loss." Is this statement true or false? Explain briefly.

10. One might think that performing a mammogram once each year—as opposed to once every three years—would triple the cost per life saved. But according to Table 6, performing the exam annually raises the cost per life-year saved by about 40 times. Does this make sense? Explain.

CHAPTER 3



Supply and Demand

Father Guido Sarducci, a character on the early *Saturday Night Live* shows, once observed that the average person remembers only about five minutes worth of material from college. He therefore proposed the "Five Minute University," in which you'd learn only the five minutes of material you'd actually remember. The economics course would last only 10 seconds, just enough time for students to learn to memorize three words: "supply and demand."

Of course, there is much more to economics than these three words. But supply and demand does play a central role in economics. What, exactly, does this familiar phrase really mean?

First, supply and demand is an economic *model,* designed to explain *how prices are determined in certain types of markets*.

It's such an important model because prices themselves play such an important role in the economy. In a market system, once the price of something has been determined, only those willing to pay that price will get it. Thus, prices determine which households will get which goods and services and which firms will get which resources. If you want to know why the cell phone industry is expanding while the video rental industry is shrinking, or why homelessness is a more pervasive problem in the United States than hunger, you need to understand how prices are determined. In this chapter, you will learn how the model of supply and demand works and how to use it.

Markets

Put any compound in front of a chemist, ask him what it is and what it can be used for, and he will immediately think of the basic elements—carbon, hydrogen, oxygen, and so on. Ask an economist almost any question about the economy, and he will immediately think about *markets*.

In ordinary language, a market is a specific location where buying and selling take place: a supermarket, a flea market, and so on. In economics, a market is not a place, but rather a collection of *traders*. More specifically,

> *a market is a group of buyers and sellers with the potential to trade with each other.*

Economists think of the economy as a collection of markets. There is a market for oranges, another for automobiles, another for real estate, and still others for corporate stocks, labor services, land, euros, and anything else that is bought and sold.

Characterizing a Market

The first step in analyzing a market is to figure out *which* market we are analyzing. This might seem easy. But we can choose to define a market in different ways, depending on our purpose.

Broad versus Narrow Definition

Suppose we want to study the personal computer industry in the United States. Should we define the market very broadly ("the market for computers"), or very narrowly ("the market for ultra-light laptops"), or something in between ("the market for laptops")? The right choice depends on the problem we're trying to analyze.

For example, if we're interested in why computers *in general* have come down in price over the past decade, we'd treat all types of computers as if they were the same good. Economists call this process **aggregation**—combining a group of distinct things into a single whole.

Aggregation The process of combining distinct things into a single whole.

But suppose we're asking a different question: Why do laptops always cost more than desktops with similar computing power? Then we'd aggregate all laptops together as one good, and all desktops as another, and look at each of these more narrowly defined markets.

We can also choose to define the *geography* of a market more broadly or more narrowly, depending on our purpose. We'd analyze the *national* market for gasoline if we're explaining general nationwide trends in gas prices. But we'd define it more locally to explain, say, why gas prices are rising more rapidly in Los Angeles than in other areas of the country.

Markets A group of buyers and seller with the potential to trade with each other

*In economics, **markets** can be defined broadly or narrowly, depending on our purpose.*

How markets are defined is one of the most important differences between *macro*economics and *micro*economics. In macroeconomics, goods and services are aggregated to the highest levels. Macro models even lump all consumer goods—breakfast cereals, cell phones, blue jeans, and so forth—into the single category "consumption goods" and view them as if they are traded in a single, national "market for consumption goods." Defining markets this broadly allows macroeconomists to take an overall view of the economy without getting bogged down in the details.

In microeconomics, by contrast, markets are defined more narrowly. Instead of asking how much we'll spend on *consumer goods,* a microeconomist might ask how much we'll spend on *health care* or *video games*. Even in microeconomics, there is always some aggregation, but not as much as in macroeconomics.

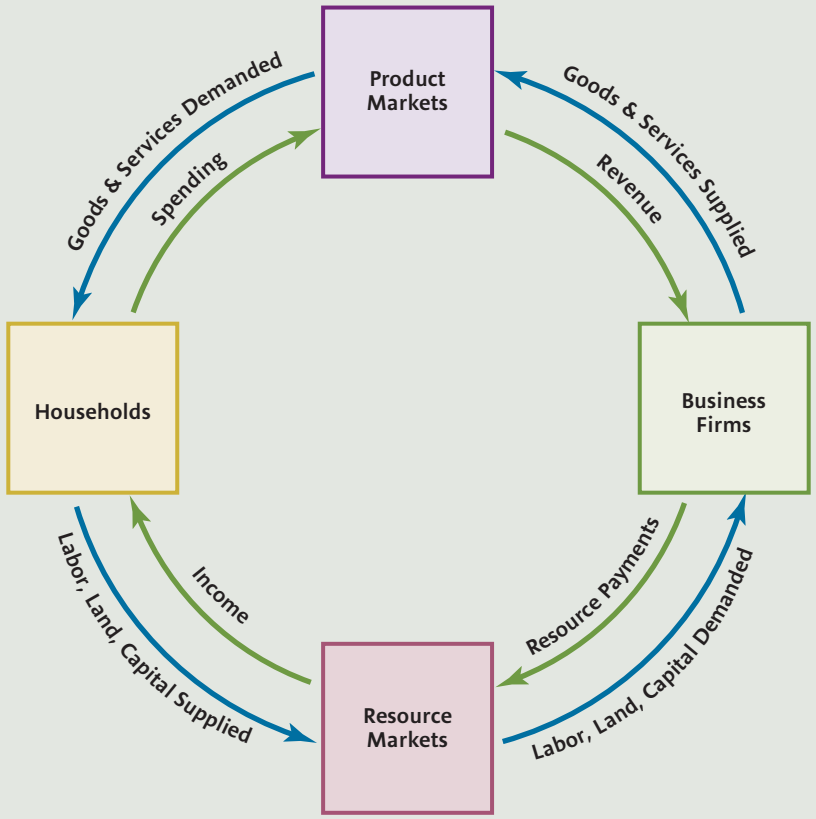
Product and Resource Markets

Circular flow A simple model that shows how goods, resources, and dollar payments flow between households and firms.

Product markets Markets in which firms sell goods and services to households

Figure 1, often called the simple **circular flow** model of the economy, illustrates two different types of markets and how they relate to each other. The upper half illustrates **product markets,** where goods and services are bought and sold. The blue arrows show the flow of products from the business firms who supply them to the households who buy them. The green arrows show the associated flow of dollars from the households who spend the dollars, to the business firms who receive these dollars as revenue. (In the real world, businesses also sell products to the government and to other businesses, but this simple version leaves out these details.)

FIGURE 1 The Circular Flow Model



The outer loop of the diagram shows the flows of goods and resources, and the markets in which they are traded. Households sell resources to firms in resource markets. Business firms use the resources to produce goods and services, which they sell to households in product markets. The inner loop shows money flows. The resource payments made by firms become income to households. Households use the income to purchases goods and services from firms.

The lower half depicts a different set of markets: **resource markets,** where labor, land, and capital are bought and sold. Here, the roles of households and firms are reversed. The blue arrows show resources flowing from the households (who own and supply them) to the business firms (who demand them). The associated flow of dollars is indicated by the green arrows: Business firms pay for the resources they use, and households receive these payments as income.

Resource markets Markets in which households that own resources sell them to firms.

In this chapter, we'll be using supply and demand to analyze product markets—particularly those where business firms supply goods and services to households. In later chapters, we'll be applying supply and demand to resource markets, and to some other types of markets that are not included in the circular flow depicted here. These include markets for existing homes, and markets for financial assets such as stocks and bonds.

Competition in Markets

A final issue in defining a market is how prices are determined. In some markets, individual buyers or sellers have an important influence over the price. For example, in the national market for cornflakes, Kellogg's—an individual *seller*—simply sets its price every few months. It can raise the price and sell fewer boxes of cereal or lower the price and sell more. In a small-town, a major *buyer* of antiques may be

able to negotiate special discount prices with the local antique shops. These are examples of *imperfectly competitive* markets.

Imperfectly competitive market
A market in which a single buyer or seller has the power to influence the price of the product.

*In **imperfectly competitive markets**, individual buyers or sellers can influence the price of the product.*

But now think about the U.S. market for wheat. Can an individual seller have any impact on the market price? Not really. On any given day there is a going price for wheat—say, $5.80 per bushel. If a farmer tries to charge more than that—say, $5.85 per bushel—he won't sell any wheat at all! His customers will instead go to one of his many competitors and buy the identical product from them for less. Each wheat farmer must take the price of wheat as a "given."

The same is true of a single wheat *buyer:* If he tries to negotiate a lower price from a seller, he'd be laughed off the farm. "Why should I sell my wheat to you for $5.75 per bushel, when there are others who will pay me $5.80?" Accordingly, each buyer must take the market price as a given.

The market for wheat is an example of a *perfectly competitive market*.

Perfectly competitive market
(informal definition) A market in which no buyer or seller has the power to influence the price.

*In **perfectly competitive markets** (or just **competitive markets**), each buyer and seller takes the market price as a given.*

What makes some markets imperfectly competitive and others perfectly competitive? You'll learn the complete answer, along with more formal definitions, when you are further into your study of *microeconomics*. But here's a hint: In perfectly competitive markets, there are many small buyers and sellers, each is a small part of the market, and the product is standardized, like wheat. Imperfectly competitive markets, by contrast, have just a few large buyers or sellers, or else the product of each seller is unique in some way.

Understanding competition in markets is important in this chapter for one simple reason:

The supply and demand model is designed to show how prices are determined in perfectly competitive markets.

Competition in the Real World

© JEFF GREENBERG/ALAMY

Markets that are truly perfectly competitive—where no buyer or seller has *any* influence over the price—are rare. Does that mean we can only use supply and demand in those rare cases, such as the market for wheat? Not at all. Supply and demand is useful for many real-world markets, even when the competition is somewhat imperfect.

Consider the market for laptop computers. Laptops made by Lenovo, Hewlett Packard, Toshiba, Apple, and other producers differ in important ways: memory, speed, operating system, and more. And even within a smaller group—say, Windows laptops with the same memory and speed—there are still differences. The keyboards feel different, the reputations for reliability and service are different, and more. For this reason, each producer can charge a different price, even for very similar laptops. Because there is no single market price that all producers take as given, the market is not *strictly* perfectly competitive.

But laptops made by different firms, while not identical, are not *that* different. So the freedom to set price is limited. For example, if other similar laptops are selling for between $900 and $1,000, Lenovo cannot charge $1,400; it would lose almost all of its customers to competitors. While there is no single market price, each producer views the *range* of prices it can charge as given.

Thus, the laptop market—while not perfectly competitive—is still somewhat competitive. And in cases like these, supply and demand can help us see how the price *range* is determined, and what makes that range rise and fall.

More generally,

> *while few markets are* strictly *perfectly competitive, most markets have enough competition for supply and demand to explain broad movements in prices.*

This is why supply and demand has proven to be the most versatile and widely used model in the economist's toolkit. Neither DVDs nor avocados strictly satisfy the requirements of perfect competition. But ask an economist why the price of DVDs has fallen dramatically in recent years, or how wildfires in southern California might affect the price of avocados next month, and he or she will invariably reach for the model of supply and demand.

In the rest of this chapter, we will build the supply and demand model. As the name implies, the model has two major parts. We will first consider each part separately, and then put them together.

Demand

It's tempting to think of the "demand" for a product psychologically—a pure "want" or "desire." But that kind of thinking can lead us astray. For example, you *want* all kinds of things: a bigger apartment, a better car, nicer clothes, more and better vacations. The list is endless. But you don't always *buy* them. Why not?

Because in addition to your wants—which you'd very much like to satisfy—you also face *constraints*. First, you have to *pay*. Second, your spending power is limited, so every decision to buy one thing is also a decision *not* to buy something else (or a decision to save less, and have less buying power in the future). As a result, every purchase confronts you with an opportunity cost. Your "wants," together with the real-world constraints that you face, determine what you will choose to buy in any market. Hence, the following definition:

> *The* ***quantity demanded*** *of a good or service is the number of units that all buyers in a market would choose to buy over a given time period, given the constraints that they face.*

Quantity demanded The quantity of a good that all buyers in a market would choose to buy during a period of time, given their constraints.

Since this definition plays a key role in any supply and demand analysis, it's worth taking a closer look at it.

***Quantity Demanded Implies a* Choice.** Quantity demanded doesn't tell us the amount of a good that households feel they "need" or "desire" in order to be happy. Instead, it tells us how much households would choose to buy *when they take into account the opportunity cost* of their decisions. The opportunity cost arises from the

constraints households face, such as having to pay a given price for the good, limits on spendable funds, and so on.

***Quantity Demanded Is* Hypothetical.** Will households actually be *able* to purchase the amount they want to purchase? As you'll soon see, usually yes. But there are special situations—analyzed in microeconomics—in which households are frustrated in buying all that they would like to buy. Quantity demanded makes no assumptions about the availability of the good. Instead, it's the answer to a hypothetical question: How much would households buy, given the constraints that they face, if the units they wanted to buy were available?

***Quantity Demanded Depends on* Price.** The price of the good is just one variable among many that influences quantity demanded. But because the price is a key variable that our model will ultimately determine, we try to keep that variable front-and-center in our thinking. This is why for the next few pages we'll assume that all other influences on demand are held constant, so we can explore the relationship between price and quantity demanded.

The Law of Demand

How does a change in price affect quantity demanded? You probably know the answer to this already: When something is more expensive, people tend to buy less of it. This common observation applies to air travel, magazines, guitars, and virtually everything else that people buy. For all of these goods and services, price and quantity are *negatively related:* that is, when price rises, quantity demanded falls; when price falls, quantity demanded rises. This negative relationship is observed so regularly in markets that economists call it the *law of demand*.

Law of demand As the price of a good increases, the quantity demanded decreases.

> *The* ***law of demand*** *states that when the price of a good rises and everything else remains the same, the quantity of the good demanded will fall.*

Read that definition again, and notice the very important words, "everything else remains the same." The law of demand tells us what would happen *if* all the other influences on buyers' choices remained unchanged, and only one influence—the price of the good—changed.

This is an example of a common practice in economics. In the real world, many variables change *simultaneously*. But to understand changes in the economy, we must first understand the effect of each variable *separately*. So we conduct a series of mental experiments in which we ask: "What would happen if this one influence—and only this one—were to change?" The law of demand is the result of one such mental experiment, in which we imagine that the price of the good changes, but all other influences on quantity demanded remain constant.

Mental experiments like this are used so often in economics that we sometimes use a shorthand Latin expression to remind us that we are holding all but one influence constant: ***ceteris paribus*** (formally pronounced KAY-ter-is PAR-ih-bus, although it's acceptable to pronounce the first word as SEH-ter-is). This is Latin for "all else the same," or "all else remaining unchanged." Even when it is not explicitly stated, the *ceteris paribus* assumption is virtually always implied. The exceptions are cases where we consider two or more influences on a variable that change simultaneously, as we will do toward the end of this chapter.

Ceteris paribus Latin for "all else remaining the same."

TABLE 1

Demand Schedule for Maple Syrup in the United States

Price (per bottle)	Quantity Demanded (bottles per month)
$1.00	75,000
$2.00	60,000
$3.00	50,000
$4.00	40,000
$5.00	35,000

The Demand Schedule and the Demand Curve

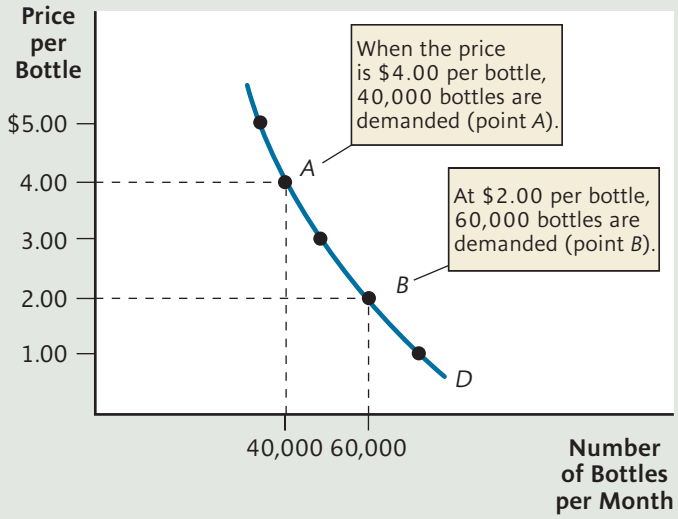
To make our discussion more concrete, let's look at a specific market: the market for real maple syrup in the United States. In this market, we'll view the buyers as U.S. households, whereas the sellers (to be considered later) are maple syrup producers in the United States or Canada.

Table 1 shows a hypothetical **demand schedule** for maple syrup in this market. This is *a list of different quantities demanded at different prices, with all other variables that affect the demand decision assumed constant*. For example, the demand schedule tells us that when the price of maple syrup is $2.00 per bottle, the quantity demanded will be 60,000 bottles per month. Notice that the demand schedule obeys the law of demand: As the price of maple syrup increases, *ceteris paribus,* the quantity demanded falls.

Demand schedule A list showing the quantities of a good that consumers would choose to purchase at different prices, with all other variables held constant.

Now look at Figure 2. It shows a diagram that will appear again and again in your study of economics. In the figure, each price-and-quantity combination in Table 1 is represented by a point. For example, point *A* represents the price $4.00 and quantity 40,000, while point *B* represents the pair $2.00 and 60,000. When we

FIGURE 2 The Demand Curve



connect all of these points with a line, we obtain the famous *demand curve*, labeled with a *D* in the figure.

Demand curve A graph of a demand schedule; a curve showing the quantity of a good or service demanded at various prices, with all other variables held constant.

The ***demand curve*** *shows the relationship between the price of a good and the quantity demanded in the market, holding constant all other variables that influence demand. Each point on the curve shows the total quantity that buyers would choose to buy at a specific price.*

Notice that the demand curve in Figure 2—like virtually all demand curves—*slopes downward*. This is just a graphical representation of the law of demand.

Shifts versus Movements Along the Demand Curve

Markets are affected by a variety of events. Some events will cause us to *move along* the demand curve; others will cause the entire demand curve to *shift*. It is crucial to distinguish between these two very different types of effects.

Let's go back to Figure 2. There, you can see that when the price of maple syrup rises from $2.00 to $4.00 per bottle, the number of bottles demanded falls from 60,000 to 40,000. This is a movement *along* the demand curve, from point *B* to point *A*. In general,

a change in the price of a good causes a movement along *the demand curve.*

In Figure 2, a *fall* in price would cause us to move *rightward* along the demand curve (from point *A* to point *B*), and a *rise* in price would cause us to move *leftward* along the demand curve (from *B* to *A*).

Remember, though, that when we draw a demand curve, we assume all other variables that might influence demand are *held constant* at some particular value. For example, the demand curve in Figure 2 might have been drawn to give us quantity demanded at each price when average household income in the United States remains constant at, say, $40,000 per year.

But suppose average income increases to $50,000. With more income, we'd expect households to buy more of *most* things, including maple syrup. This is illustrated in Table 2. At the original income level, households would choose to buy 60,000 bottles of maple syrup at $2.00 per bottle. But after income rises, they would choose to buy more at that price—80,000 bottles, according to Table 2. A similar

TABLE 2

Increase in Demand for Maple Syrup in the United States

Price (per bottle)	Original Quantity Demanded (average income = $40,000)	New Quantity Demanded (average income = $50,000)
$1.00	75,000	95,000
$2.00	60,000	80,000
$3.00	50,000	70,000
$4.00	40,000	60,000
$5.00	35,000	55,000

change would occur at any other price for maple syrup: After income rises, households would choose to buy more than before. In other words, the rise in income *changes the entire relationship between price and quantity demanded*. We now have a *new* demand curve.

Figure 3 plots the new demand curve from the quantities in the third column of Table 2. The new demand curve lies to the *right* of the old curve. For example, at a price of $2.00, quantity demanded increases from 60,000 bottles on the old curve (point *B*) to 80,000 bottles on the *new* demand curve (point *C*). As you can see, the rise in household income has *shifted* the demand curve to the right.

More generally,

> *a change in any variable that affects demand—except for the good's price—causes the demand curve to shift.*

When buyers would choose to buy a greater quantity at any price, the demand curve shifts *rightward*. If they would decide to buy a smaller quantity at any price, the demand curve shifts *leftward*.

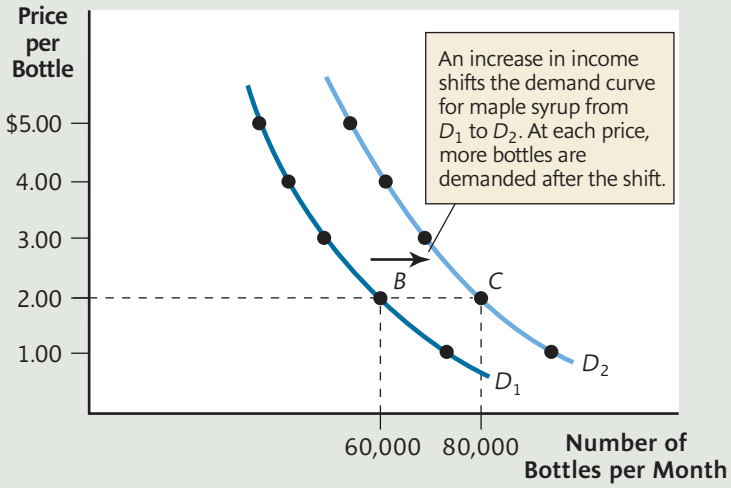
"Change in Quantity Demanded" versus "Change in Demand"

Language is important when discussing demand. The term *quantity demanded* means a *particular amount* that buyers would choose to buy at a specific price, represented by a single point on a demand curve. *Demand*, by contrast, means the *entire relationship* between price and quantity demanded, represented by the entire demand curve.

For this reason, when a change in the price of a good moves us *along* a demand curve, we call it a **change in quantity demanded.** For example, in Figure 2, the movement from point *A* to point *B* is an *increase* in quantity demanded. This is a change from one number (40,000 bottles) to another (60,000 bottles).

Change in quantity demanded
A movement along a demand curve in response to a change in price.

FIGURE 3 A Shift of the Demand Curve



Change in demand A shift of a demand curve in response to a change in some variable other than price.

When something *other* than the price changes, causing the entire demand curve to shift, we call it a **change in demand.** In Figure 3, for example, the shift in the curve would be called an *increase in demand*.

Factors That Shift the Demand Curve

Let's take a closer look at what might cause a change in demand (a shift of the demand curve). Keep in mind that for now, we're exploring *one factor at a time*, always keeping *all other determinants of demand constant*.

Income The amount that a person or firm earns over a particular period.

Normal good A good that people demand more of as their income rises.

Inferior good A good that people demand less of as their income rises.

Income. In Figure 3, an increase in **income** shifted the demand for maple syrup to the right. In fact, a rise in income increases demand for *most* goods. We call these **normal goods.** Housing, automobiles, health club memberships, and real maple syrup are all examples of normal goods.

But not all goods are normal. For some goods—called **inferior goods**—a rise in income would *decrease* demand—shifting the demand curve *leftward*. Regular-grade ground chuck is a good example. It's a cheap source of protein, but not as high in quality as sirloin. With higher income, households could more easily afford better types of meat—ground sirloin or steak, for example. As a result, higher incomes for buyers might cause the demand for ground chuck to *decrease*. For similar reasons, we might expect that Greyhound bus tickets (in contrast to airline tickets) and single-ply paper towels (in contrast to two-ply) are inferior goods.

> *A rise in income will* increase *the demand for a* normal *good, and* decrease *the demand for an* inferior *good.*

Wealth The total value of everything a person or firm owns, at a point in time, minus the total amount owed.

Wealth. Your **wealth** at any point in time is the total value of everything you *own* (cash, bank accounts, stocks, bonds, real estate or any other valuable property) minus the total dollar amount you *owe* (home mortgage, credit card debt, auto loan, student loan, and so on). Although income and wealth are different, (see the nearby Dangerous Curves box), they have similar effects on demand. Increases in wealth among buyers—because of an increase in the value of their stocks or bonds, for example—gives them more funds with which to purchase goods and services. As you might expect,

> *an increase in wealth will* increase *demand (shift the curve rightward) for a normal good, and* decrease *demand (shift the curve leftward) for an inferior good.*

Substitute A good that can be used in place of some other good and that fulfills more or less the same purpose.

Prices of Related Goods. A **substitute** is a good that can be used in place of another good and that fulfills more or less the same purpose. For example, many people use real maple syrup to sweeten their pancakes, but they could use a number of other things instead: honey, sugar, jam, or *artificial* maple syrup. Each of these can be considered a substitute for real maple syrup.

When the price of a substitute rises, people will choose to buy *more* maple syrup. For example, when the price of jam rises, some jam users will switch to maple syrup, and the demand for maple syrup will increase. In general,

> *a rise in the price of a substitute increases the demand for a good, shifting the demand curve to the right.*

Of course, if the price of a substitute falls, we have the opposite result: Demand for the original good decreases, shifting its demand curve to the left.

A **complement** is the opposite of a substitute: It's used *together with* the good we are interested in. Pancake mix is a complement to maple syrup, since these two goods are used frequently in combination. If the price of pancake mix rises, some consumers will switch to other breakfasts—bacon and eggs, for example—that *don't* include maple syrup. The demand for maple syrup will decrease.

A rise in the price of a complement decreases the demand for a good, shifting the demand curve to the left.

To test yourself: How would a lower price for DVDs affect the demand for DVD players? How would it affect the demand for movies in theaters?

Complement A good that is used together with some other good.

Population. As the population increases in an area, the number of buyers will ordinarily increase as well, and the demand for a good will increase. The growth of the U.S. population over the last 50 years has been an important reason (but not the only reason) for rightward shifts in the demand curves for food, housing, automobiles, and many other goods and services.

Expected Price. If buyers expect the price of maple syrup to rise next month, they may choose to purchase more *now* to stock up before the price hike. If people expect the price to drop, they may postpone buying, hoping to take advantage of the lower price later.

In many markets, an expectation that price will rise in the future shifts the current *demand curve rightward, while an expectation that price will fall shifts the current demand curve leftward.*

Expected price changes for goods are especially important for goods that can be purchased and stored until needed later. Expected price changes are also important in the markets for financial assets such as stocks and bonds and in the market for housing, as you'll see in the next chapter.

Tastes. Not everyone likes maple syrup. And among those who do, some *really* like it, and some like it just a little. Buyers' basic attitudes toward a good are based on their *tastes* or *preferences*. Economists are sometimes interested in where these tastes come from or what makes them change. But for the most part, economics deals with the *consequences* of a change in tastes, whatever the reason for its occurrence.

When tastes change *toward* a good (people favor it more), demand increases, and the demand curve shifts to the right. When tastes change *away* from a good, demand decreases, and the demand curve shifts to the left. An example of this is the change in tastes away from cigarettes over the past several decades. The cause may have been an aging population, a greater concern about health among people of *all* ages, or successful antismoking advertising. But regardless of the cause, the effect has been to decrease the demand for cigarettes, shifting the demand curve to the left.

dangerous curves

Income versus Wealth It's easy to confuse *income* with *wealth*, because both are measured in dollars and both are sources of funds that can be spent on goods and services. But they are not the same thing. Your income is how much you earn *per period of time* (such as, $20 *per hour*, $3,500 *per month*, or $40,000 *per year*). Your wealth, by contrast, is the value of what you own minus the value of what you owe at a particular *moment in time*. (Such as, on December 31, 2010, the value of what you own is $12,000, but the value of what you owe is $9,000, so you have $3,000 in wealth.)

Someone can have a high income and low or even negative wealth (such as college students who get good jobs after graduation but still owe a lot on their student loans). And a person with great wealth could have little or no income (for example, if they make especially bad investment choices and earn little or no income from their wealth).

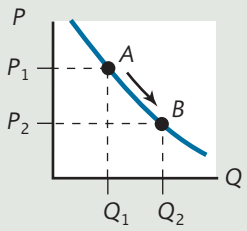
Other Shift Variables. Many other things, besides those we've discussed, can shift the demand curve. For example, if the government began to offer subsidies to households who buy maple syrup, demand would shift rightward. Also, if business firms (rather than just households) are among the buyers, then changes in the demand for their own products will influence their demand for maple syrup. We'll discuss additional shift-variables in later chapters, as they become relevant.

Demand: A Summary

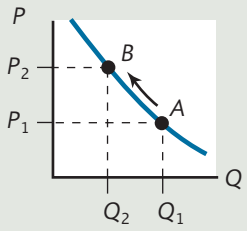
Figure 4 summarizes the key variables we've discussed that affect the demand side of the market and how their effects are represented with a demand curve. Notice the important distinction between events that move us *along* the curve (changes in price) and events that *shift* the curve.

FIGURE 4 **The Demand Curve—A Summary**

price↓ ⇒ move leftward along curve

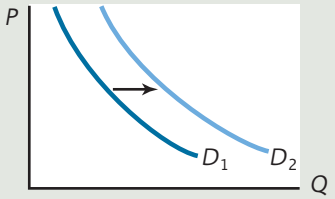


price↑ ⇒ move rightward along curve



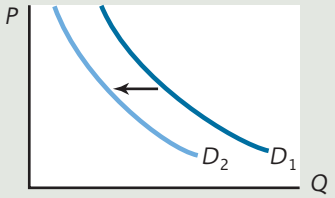
- income or wealth↑
- price of substitute↑
- price of complement↓
- population↑
- expected price↑
- Tastes shift toward good

⇒ demand curve shifts rightward



- income or wealth↓
- price of substitute↓
- price of complement↑
- population↓
- expected price↓
- Tastes shift away from good

⇒ demand curve shifts leftward



Supply

When most people hear the word *supply*, their first thought is that it's the amount of something "available," as if this amount were fixed in stone. For example, someone might say, "We can only drill so much oil from the ground," or "There are only so many apartments for rent in this town." And yet, the world's known oil reserves—as well as yearly production of oil—have increased dramatically over the last half century, as oil companies have found it worth their while to look harder for oil. Similarly, in most towns and cities, short buildings have been replaced with tall ones, and the number of apartments has increased. Supply, like demand, can change, and the amount of a good supplied in a market depends on the *choices* made by those who produce it.

What governs these choices? We assume that those who supply goods and services have a goal: to earn the highest profit possible. But they also face constraints. First, in a competitive market, the price they can charge for their product is a *given*—the market price. Second, firms have to pay the *costs* of producing and selling their product. These costs will depend on the production process they use, the prices they must pay for their inputs, and more. Business firms' desire for profit, together with the real-world constraints that they face, determines how much they will choose to sell in any market. Hence, the following definition:

> ***Quantity supplied*** *is the number of units of a good that* all *sellers in the market would choose to sell over some time period, given the constraints that they face.*

Quantity supplied The specific amount of a good that all sellers in a market would choose to sell over some time period, given their constraints.

Let's briefly go over the notion of quantity supplied to clarify what it means and doesn't mean.

***Quantity Supplied Implies a* Choice.** Quantity supplied doesn't tell us the amount of, say, maple syrup that sellers would like to sell *if* they could charge a thousand dollars for each bottle, and *if* they could produce it at zero cost. Instead, it's the quantity that firms *choose* to sell—the quantity that gives them the highest profit given the constraints they face.

***Quantity Supplied Is* Hypothetical.** Will firms actually be *able* to sell the amount they want to sell at the going price? You'll soon see that they usually can. But the definition of quantity supplied makes no assumptions about firms' ability to sell the good. Quantity supplied answers the hypothetical question: How much *would* suppliers sell, given their constraints, if they were able to sell all that they wanted to.

***Quantity Supplied Depends on* Price.** The price of the good is just one variable among many that influences quantity supplied.

© JAMES MARSHALL/THE IMAGE WORKS

dangerous curves

Does Supply Affect Demand? A troubling thought may have occurred to you. Among the variables that shift the demand curve in Figure 4, shouldn't we include the amount of syrup available? Or to put the question another way, doesn't supply influence demand?

No—at least not directly. We can think of the demand curve as the answers to a series of hypothetical questions about how much people *would like* to buy at different prices. A change in the amount available would not affect the answers to these questions, and so wouldn't affect the curve itself. As you'll see later in this chapter, events relating to supply *will* affect the *price* of the good, but this causes a movement along—not a shift of—the demand curve.

But—as with demand—we want to keep that variable foremost in our thinking. This is why for the next couple of pages we'll assume that all other influences on supply are held constant, so we can explore the relationship between price and quantity supplied.

The Law of Supply

How does a change in price affect quantity supplied? When a seller can get a higher price for a good, producing and selling it become more profitable. Producers will devote more resources toward its production—perhaps even pulling resources from other goods they produce—so they can sell more of the good in question. For example, a rise in the price of laptop (but not desktop) computers will encourage computer makers to shift resources out of the production of other things (such as desktop computers) and toward the production of laptops.

In general, price and quantity supplied are *positively related:* When the price of a good rises, the quantity supplied will rise as well. This relationship between price and quantity supplied is called the law of supply, the counterpart to the law of demand we discussed earlier.

Law of supply As the price of a good increases, the quantity supplied increases.

*The **law of supply** states that when the price of a good rises, and everything else remains the same, the quantity of the good supplied will rise.*

Once again, notice the very important words "everything else remains the same"—*ceteris paribus*. Although many other variables influence the quantity of a good supplied, the law of supply tells us what would happen if all of them remained unchanged and only one—the price of the good—changed.

The Supply Schedule and the Supply Curve

Let's continue with our example of the market for maple syrup in the United States. Who are the suppliers in this market? Maple syrup producers are located mostly in the forests of Vermont, upstate New York, and Canada. The market quantity supplied is the amount of syrup all of these producers together would offer for sale at each price for maple syrup in the United States.

Supply schedule A list showing the quantities of a good or service that firms would choose to produce and sell at different prices, with all other variables held constant.

Table 3 shows the **supply schedule** for maple syrup—a *list of different quantities supplied at different prices, with all other variables held constant*. As you can see, the supply schedule obeys the law of supply: As the price of maple syrup rises, the quantity supplied rises along with it. But how can this be? After all, maple trees must be about 40 years old before they can be tapped for syrup, so any rise in quantity supplied now or in the near future cannot come from an increase in planting. What, then, causes quantity supplied to rise as price rises?

Many things. With higher prices, firms will find it profitable to tap existing trees more intensively. Evaporating and bottling can be done more carefully, so that less maple syrup is spilled and more is available for shipping. Or the product can be diverted from other areas and shipped to the United States instead. For example, if the price of maple syrup rises in the United States but not in Canada, producers would shift deliveries away from Canada so they could sell more in the United States.

TABLE 3

Supply Schedule for Maple Syrup in the United States

Price (per bottle)	Quantity Supplied (bottles per month)
$1.00	25,000
$2.00	40,000
$3.00	50,000
$4.00	60,000
$5.00	65,000

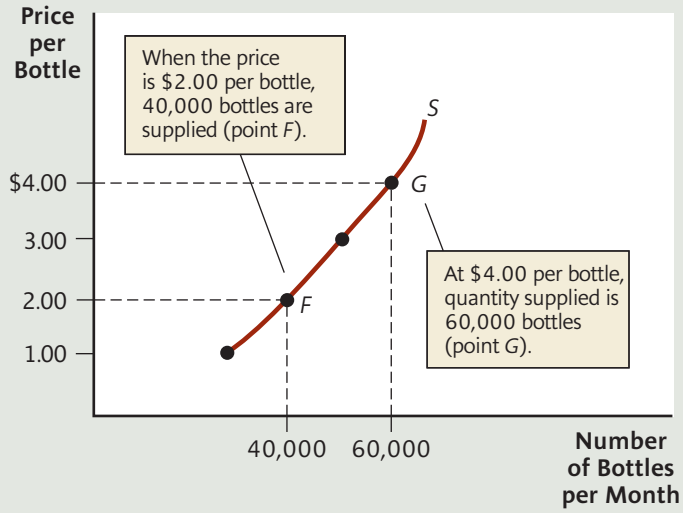
Now look at Figure 5, which shows a very important curve—the counterpart to the demand curve we drew earlier. In Figure 5, each point represents a price-quantity pair taken from Table 3. For example, point *F* in the figure corresponds to a price of $2.00 per bottle and a quantity of 40,000 bottles per month, while point *G* represents the price-quantity pair $4.00 and 60,000 bottles. Connecting all of these points with a solid line gives us the *supply curve* for maple syrup, labeled with an *S* in the figure.

*The **supply curve** shows the relationship between the price of a good and the quantity supplied in the market, holding constant the values of all other variables that affect supply. Each point on the curve shows the quantity that sellers would choose to sell at a specific price.*

Supply curve A graph of a supply schedule, showing the quantity of a good or service supplied at various prices, with all other variables held constant.

Notice that the supply curve in Figure 5—like all supply curves for goods and services—is *upward sloping*. This is the graphical representation of the law of supply.

FIGURE 5 The Supply Curve



Shifts versus Movements Along the Supply Curve

As with the demand curve, it's important to distinguish those events that will cause us to *move along* a given supply curve for the good, and those that will cause the entire supply curve to *shift*.

If you look once again at Figure 5, you'll see that if the price of maple syrup rises from $2.00 to $4.00 per bottle, the number of bottles supplied rises from 40,000 to 60,000. This is a movement *along* the supply curve, from point *F* to point *G*. In general,

a change in the price of a good causes a movement along *the supply curve.*

In the figure, a *rise* in price would cause us to move *rightward* along the supply curve (from point *F* to point *G*) and a *fall* in price would move us *leftward* along the curve (from point *G* to point *F*).

But remember that when we draw a supply curve, we assume that all other variables that might influence supply are *held constant* at some particular values. For example, the supply curve in Figure 5 might tell us the quantity supplied at each price when the cost of an important input—transportation from the farm to the point of sale—remains constant.

But suppose the cost of transportation drops. Then, at any given price for maple syrup, firms would find it more profitable to produce and sell it. This is illustrated in Table 4. With the original transportation cost, and a selling price of $4.00 per bottle, firms would choose to sell 60,000 bottles. But after transportation cost falls, they would choose to produce and sell more—80,000 bottles in our example—assuming they could still charge $4.00 per bottle. A similar change would occur for any other price of maple syrup we might imagine: After transportation costs fall, firms would choose to sell more than before. In other words, *the entire relationship between price and quantity supplied has changed,* so we have a *new* supply curve.

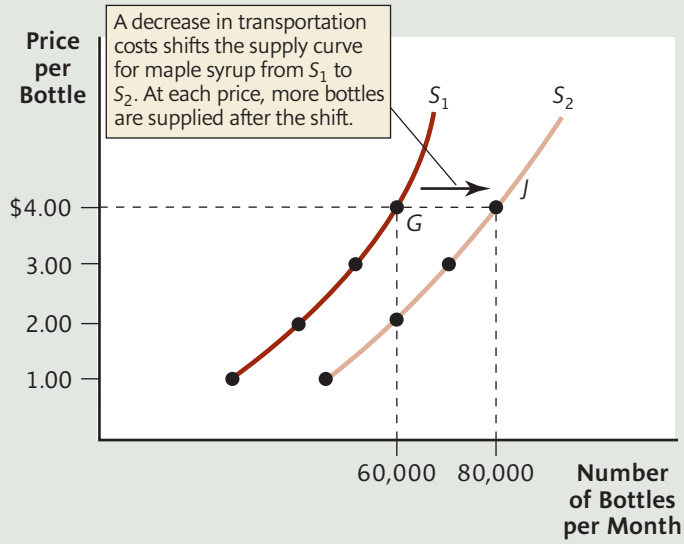
Figure 6 plots the new supply curve from the quantities in the third column of Table 4. The new supply curve lies to the *right* of the old one. For example, at a price of $4.00, quantity supplied increases from 60,000 bottles on the old curve (point *G*) to 80,000 bottles on the *new* supply curve (point *J*). The drop in the transportation costs has *shifted* the supply curve to the right.

TABLE 4

Increase in Supply of Maple Syrup in the United States

Price (per bottle)	Original Quantity Supplied	Quantity Supplied After Decrease in Transportation Cost
$1.00	25,000	45,000
$2.00	40,000	60,000
$3.00	50,000	70,000
$4.00	60,000	80,000
$5.00	65,000	90,000

FIGURE 6 A Shift of the Supply Curve



In general,

a change in any variable that affects supply—except for the good's price—causes the supply curve to shift.

If sellers want to sell a greater quantity at any price, the supply curve shifts *rightward*. If sellers would prefer to sell a smaller quantity at any price, the supply curve shifts *leftward*.

Change in Quantity Supplied versus Change in Supply

As we stressed in our discussion of the demand side of the market, be careful about language when thinking about supply. The term *quantity supplied* means a *particular amount* that sellers would choose to sell at a *particular* price, represented by a single point on the supply curve. The term *supply,* however, means the *entire relationship* between price and quantity supplied, as represented by the entire supply curve.

For this reason, when the price of the good changes, and we move *along* the supply curve, we have a **change in quantity supplied.** For example, in Figure 5, the movement from point *F* to point *G* is an *increase* in quantity supplied.

When something *other* than the price changes, causing the entire supply curve to shift, we call it a **change in supply.** The shift in Figure 6, for example, would be called an *increase in supply*.

Change in quantity supplied A movement along a supply curve in response to a change in price.

Change in supply A shift of a supply curve in response to a change in some variable other than price.

Factors That Shift the Supply Curve

Let's look at some of the *causes* of a change in supply (a shift of the supply curve). As always, we're considering *one* variable at a time, keeping all other determinants of supply constant.

Input Prices. In Figure 6, we saw that a drop in transportation costs shifted the supply curve for maple syrup to the right. But producers of maple syrup use a variety of other inputs besides transportation: land, maple trees, sap pans, evaporators, labor, glass bottles, and more. A lower price for any of these means a lower cost of producing and selling maple syrup, making it more profitable. As a result, we would expect producers to shift resources into maple syrup production, causing an increase in supply.

In general,

> *a fall in the price of an input causes an increase in supply, shifting the supply curve to the right.*

Similarly, a rise in the price of an input causes a decrease in supply, shifting the supply curve to the left. If, for example, the wages of maple syrup workers rose, the supply curve in Figure 6 would shift to the left.

Price of Alternatives. Many firms can switch their production rather easily among several different goods or services, each of which requires more or less the same inputs. For example, a dermatology practice can rather easily switch its specialty from acne treatments for the young to wrinkle treatments for the elderly. An automobile producer can—without too much adjustment—switch to producing light trucks. And a maple syrup producer could dry its maple syrup and produce maple *sugar* instead. Or it could even cut down its maple trees and sell maple wood as lumber. These other goods that firms *could* produce are called **alternate goods** and their prices influence the supply curve.

Alternate goods Other goods that firms in a market could produce instead of the good in question.

For example, if the price of maple *sugar* rose, then at any given price for maple *syrup,* producers would shift some production from syrup to sugar. This would be a decrease in the supply of maple syrup.

Another alternative for the firm is to sell the *same* good in a *different* market, which we'll call an **alternate market.** For example, since we are considering the market for maple syrup in the United States, the maple syrup market in Canada is a different market for producers. For any given price in the United States, a rise in the price of maple syrup in Canada will cause producers to shift some sales from the United States to Canada. In the U.S. market, this will cause the supply curve to shift leftward.

Alternate market A market other than the one being analyzed in which the same good could be sold.

> *When the price for an alternative rises—either an alternate good or the same good in an alternate market—the supply curve shifts leftward.*

Similarly, a decrease in the price of an alternate good (or a lower price in an alternate market) will shift the supply curve rightward.

Technology. A *technological advance* in production occurs whenever a firm can produce a given level of output in a new and cheaper way than before.

Examples would include a new, more efficient tap that draws more maple syrup from each tree, or a new bottling method that reduces spillage. Advances like these would reduce the cost of producing maple syrup, making it more profitable, and producers would want to make and sell more of it at any price.

In general,

> *cost-saving technological advances increase the supply of a good, shifting the supply curve to the right.*

Number of Firms. A change in the number of firms in a market will change the quantity that all sellers together would want to sell at any given price. For example, if—over time—more people decided to open up maple syrup farms because it was a profitable business, the supply of maple syrup would increase. And if maple syrup farms began closing down, their number would be reduced and supply would decrease.

An increase in the number of sellers—with no other change—shifts the supply curve rightward.

Expected Price. Imagine you're the president of Sticky's Maple Syrup, Inc., and you expect that the market price of maple syrup—over which you, as an individual seller, have no influence—to rise next month. What would you do? You'd certainly want to postpone selling some of your maple syrup until the price is higher and your profit greater. Therefore, at any given price *now,* you might slow down production, or just slow down sales by warehousing more of what you produce. If other firms have similar expectations of a price hike, they'll do the same. Thus, an expectation of a *future* price hike will decrease supply *in the present*.

Suppose instead you expect the market price to *drop* next month. Then—at any given price—you'd want to sell more *now,* by stepping up production and even selling out of your inventories. So an expected future drop in the price would cause an increase in supply in the present.

Expected price is especially important when suppliers can hold inventories of goods for later sale, or when they can easily shift production from one time period to another.

In many markets, an expectation of a future price rise *shifts the current supply curve* leftward. *Similarly, an expectation of a* future price drop *shifts the current supply curve* rightward.

Changes in Weather and Other Natural Events. Weather conditions are an especially important determinant of the supply of agricultural goods.

Favorable weather *increases crop yields, and causes a* rightward *shift of the supply curve for that crop*. Unfavorable weather *destroys crops and shrinks yields, and shifts the supply curve* leftward.

In addition to bad weather, natural disasters such as fires, hurricanes, and earthquakes can destroy or disrupt the productive capacity of *all* firms in a region. If many sellers of a particular good are located in the affected area, the supply curve for that good will shift leftward. For example, after Hurricanes Katrina and Rita struck the U.S. Gulf Coast in August and September of 2005, 20 percent of the nation's oil refining capacity was taken out for several weeks, causing a sizable leftward shift of the supply curve for gasoline.

Other Shift Variables. Many other things, besides those listed earlier, can shift the supply curve. For example, a government tax imposed on maple syrup producers would raise the cost of making and selling maple syrup. To suppliers, this would

dangerous curves

Does Demand Affect Supply? In the list of variables that shift the supply curve in Figure 7 we've left out the amount that buyers would like to buy. Is this a mistake? Doesn't demand affect supply?

The answer is no—at least, not directly. The supply curve tells us how much sellers would like to sell at each different price. Demand doesn't affect this hypothetical quantity, so demand has no direct influence on the position of the supply curve. As you'll soon see, demand *can* affect the price of a good. But this causes a movement *along* the supply curve—not a shift.

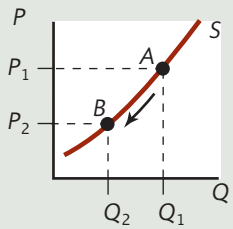
have the same effect as a higher price for transportation: it would shift the supply curve leftward. We'll discuss other shift variables for supply as they become relevant in later chapters.

Supply—A Summary

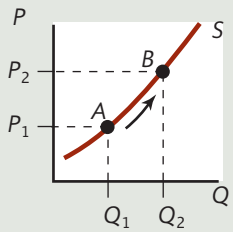
Figure 7 summarizes the various factors we've discussed that affect the supply side of the market, and how we illustrate them using a supply curve. As with demand, notice which events move us along the supply curve (changes in price) and which shift the curve. To test yourself, you might want to create a list of the shift variables in Figure 4 and Figure 7, in random order. Then explain, for each item, which curve shifts, and in which direction.

FIGURE 7 The Supply Curve—A Summary

price↓ ⇒ move leftward along curve

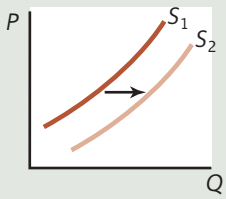


price↑ ⇒ move rightward along curve



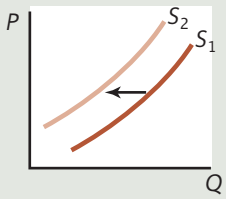
- price of input↓
- price of alternatives↓
- number of firms↑
- expected price↓
- technological advance
- favorable weather

⇒ supply curve shifts rightward



- price of input↑
- price of alternatives↑
- number of firms↓
- expected price↑
- unfavorable weather

⇒ supply curve shifts leftward



Putting Supply and Demand Together

What happens when buyers and sellers, each having the desire and the ability to trade, come together in a market? The two sides of the market certainly have different agendas. Buyers would like to pay the lowest possible price, while sellers would like to charge the highest possible price. Is there chaos when they meet, with buyers and sellers endlessly chasing after each other or endlessly bargaining for advantage, so that trade never takes place? A casual look at the real world suggests not. In most markets, most of the time, there is order and stability in the encounters between buyers and sellers. In most cases, prices do not fluctuate wildly from moment to moment but seem to hover around a stable value. Even when this stability is short-lived—lasting only a day, an hour, or even a minute in some markets—for this short-time the market seems to be at rest. Whenever we study a market, therefore, we look for this state of rest—a price and quantity at which the market will settle, at least for a while.

Economists use the word *equilibrium* when referring to a state of rest. When a market is in equilibrium, both the price of the good and the quantity bought and sold have settled into a state of rest. More formally,

*the **equilibrium price** and **equilibrium quantity** are values for price and quantity in the market that, once achieved, will remain constant—unless and until the supply curve or the demand curve shifts.*

Equilibrium price The market price that, once achieved, remains constant until either the demand curve or supply curve shifts.

Equilibrium quantity The market quantity bought and sold per period that, once achieved, remains constant until either the demand curve or supply curve shifts.

Finding the Equilibrium Price and Quantity

Look at Table 5, which combines the supply and demand schedules for maple syrup from Tables 1 and 3. We'll use Table 5 to find the equilibrium price in this market through the process of elimination.

Prices below the Equilibrium Price

Let's first ask what would happen if the price were less than $3.00 per bottle—say, $1.00. At this price, Table 5 tells us that buyers would want to buy 75,000 bottles each month, while sellers would offer to sell only 25,000. There would be an **excess demand** of 50,000 bottles. What would happen in this cases? Buyers would compete with each other to get more maple syrup than was available, and would

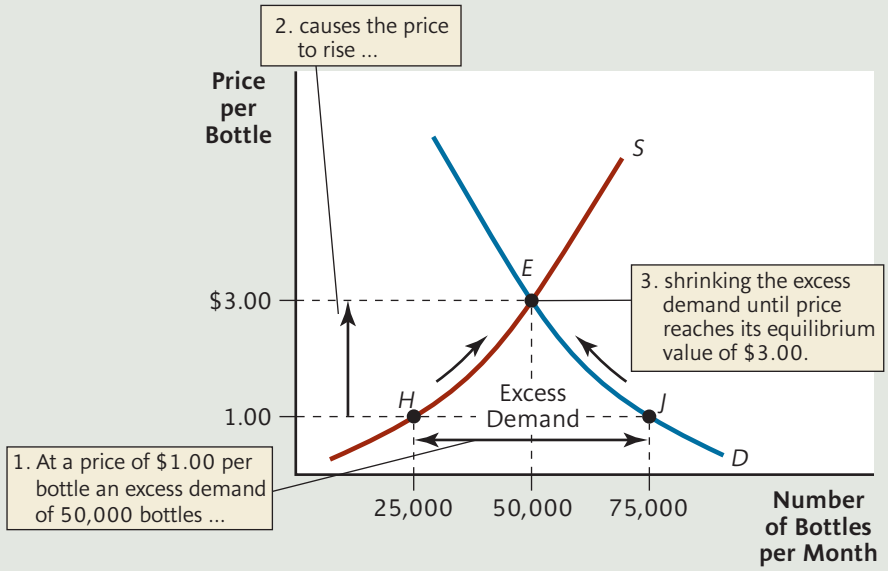
Excess demand At a given price, the amount by which quantity demanded exceeds quantity supplied.

TABLE 5

Finding the Market Equilibrium

Price (per bottle)	Quantity Demanded (bottles per month)	Quantity Supplied (bottles per month)	Excess Demand or Supply?	Consequence
$1.00	75,000	25,000	Excess Demand	Price will Rise
$2.00	60,000	40,000	Excess Demand	Price will Rise
$3.00	**50,000**	**50,000**	**Neither**	**No Change in price**
$4.00	40,000	60,000	Excess Supply	Price will Fall
$5.00	35,000	65,000	Excess Supply	Price will Fall

FIGURE 8 **Excess Demand Causes Price to Rise**



offer to pay a higher price rather than do without. The price would then rise. The same would occur if the price were $2.00, or any other price below $3.00.

We conclude that any price less than $3.00 cannot be an equilibrium price. If the price starts below $3.00, it would start rising—*not* because the supply curve or the demand curve had shifted, but from natural forces within the market itself. This directly contradicts our definition of equilibrium price.

Figure 8 illustrates the same process by putting the supply and demand curves together on the same graph. As you can see, at a price of $1.00, quantity supplied of 25,000 bottles is found at point *H* on the supply curve, while quantity demanded is at point *J* on the demand curve. The horizontal difference between the two curves at $1.00 is a graphical representation of the excess demand at that price.

At this point, we should ask another question: If the price were initially $1.00, would it ever *stop* rising? Yes. Since excess demand is the reason for the price to rise, the process will stop when the excess demand is gone. And as you can see in Figure 8, the rise in price *shrinks* the excess demand in two ways. First, as price rises, buyers demand a smaller quantity—a leftward movement along the demand curve. Second, sellers increase supply to a larger quantity—a rightward movement along the supply curve. Finally, when the price reaches $3.00 per bottle, the excess demand is gone and the price stops rising.

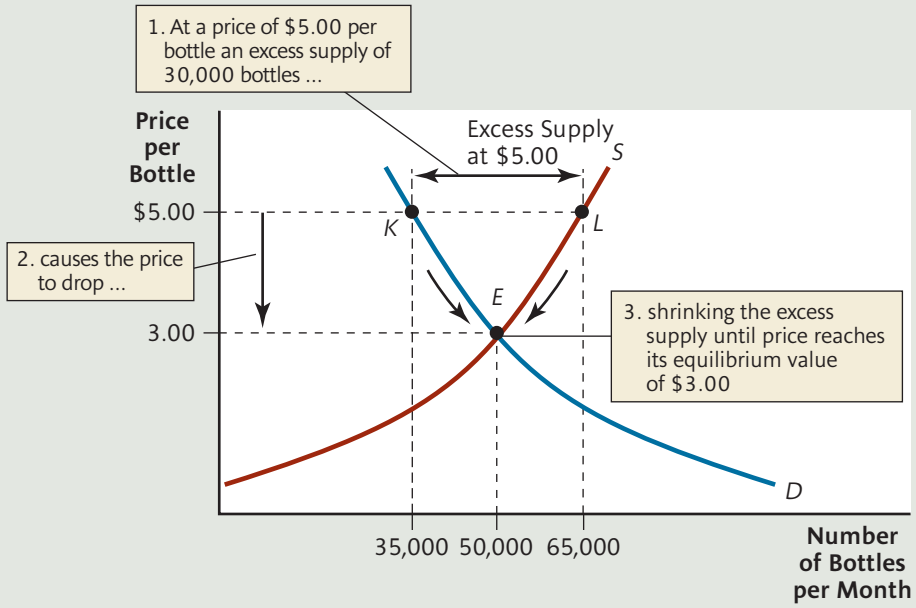
This logic tells us that $3.00 is an *equilibrium* price in this market—a value that won't change as long as the supply and demand curves stay put. But is it the *only* equilibrium price?

Prices above the Equilibrium Price

We've shown that any price *below* $3.00 is not an equilibrium, but what about a price *greater* than $3.00? Let's see.

Suppose the price of maple syrup was, say, $5.00 per bottle. Look again at Table 5 and you'll find that, at this price, quantity supplied would be 65,000 bottles

FIGURE 9 **Excess Supply Causes Price to Fall**



per month, while quantity demanded would be only 35,000 bottles. There is an **excess supply** of 30,000 bottles. Sellers would compete with each other to sell more maple syrup than buyers wanted to buy, and the price would fall. Thus, $5.00 cannot be the equilibrium price.

Excess supply At a given price, the amount by which quantity supplied exceeds quantity demanded.

Figure 9 provides a graphical view of the market in this situation. With a price of $5.00, the excess supply is the horizontal distance between points *K* (on the demand curve) and *L* (on the supply curve). In the figure, the resulting drop in price would move us along both the supply curve (leftward) and the demand curve (rightward). As these movements continued, the excess supply of maple syrup would shrink until it disappeared, once again, at a price of $3.00 per bottle. Our conclusion: If the price happens to be above $3.00, it will fall to $3.00 and then stop changing.

You can see that $3.00 is the equilibrium price—and the *only* equilibrium price—in this market. Moreover, at this price, sellers would want to sell 50,000 bottles—the same quantity that households would want to buy. So, when price comes to rest at $3.00, quantity comes to rest at 50,000 per month—the *equilibrium quantity*.

Equilibrium on a Graph

No doubt, you have noticed that $3.00 happens to be the price at which the supply and demand curves cross. This leads us to an easy, graphical technique for locating our equilibrium:

> *To find the equilibrium in a competitive market, draw the supply and demand curves. Market equilibrium occurs where the two curves cross. At this crossing point, the equilibrium price is found on the vertical axis, and the equilibrium quantity on the horizontal axis.*

Notice that in equilibrium, the market is operating on *both* the supply curve *and* the demand curve so that—at a price of $3.00—quantity demanded and quantity supplied are equal. There are no dissatisfied buyers unable to find goods they want to purchase, nor are there any frustrated sellers unable to sell goods they want to sell. Indeed, this is why $3.00 is the equilibrium price. It's the only price that creates consistency between what buyers choose to buy and sellers choose to sell.

But we don't expect a market to stay at any particular equilibrium forever, as you're about to see.

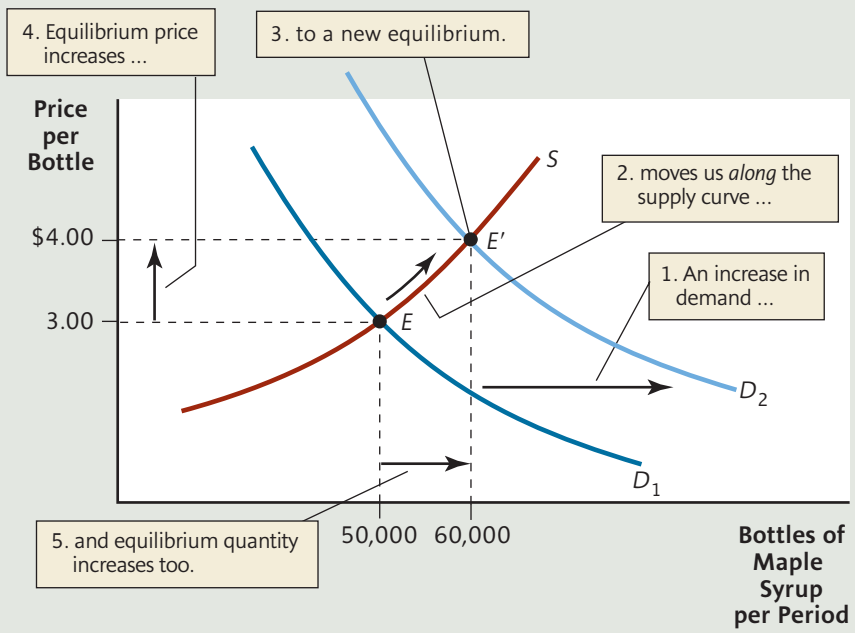
What Happens When Things Change?

Remember that in order to draw the supply and demand curves in the first place, we had to assume particular values for all the other variables—besides price—that affect demand and supply. If one of these variables changes, then either the supply curve or the demand curve will shift, and our equilibrium will change as well. Let's look at some examples.

Example: Income Rises, Causing an Increase in Demand

In Figure 10, point *E* shows an initial equilibrium in the U.S. market for maple syrup, with an equilibrium price of $3.00 per bottle, and equilibrium quantity of 50,000 bottles per month. Suppose that the incomes of buyers rise because the U.S. economy recovers rapidly from a recession. We know that income is one of the

FIGURE 10 **A Shift in Demand and a New Equilibrium**



shift-variables in the demand curve (but not the supply curve). We also can reason that maple syrup is a *normal good,* so the rise in income will cause the demand curve to shift rightward. What happens then?

The old price—$3.00—is no longer the equilibrium price. How do we know? Because if the price *did* remain at $3.00 after the demand curve shifted, there would be an excess demand that would drive the price upward. The new equilibrium—at point E'—is the new intersection point of the curves *after* the shift in the demand curve. Comparing the original equilibrium at point E with the new one at point E', we find that the shift in demand has caused the equilibrium price to rise (from $3.00 to $4.00) and the equilibrium quantity to rise as well (from 50,000 to 60,000 bottles per month).

Notice, too, that in moving from point E to point E', we move *along* the supply curve. That is, a shift of the demand curve has caused a movement along the supply curve. Why is this? The demand shift causes the *price* to rise, and a rise in price always causes a movement *along* the supply curve. But the supply curve itself does not shift because none of the variables that affect sellers—other than the price of the good—has changed.

dangerous curves

The Endless Loop of Erroneous Logic In trying to work out what happens after, say, a rise in income, you might find yourself caught in an endless loop. It goes something like this: "A rise in income causes an increase in demand. An increase in demand causes the price to rise. A higher price causes supply to increase. Greater supply causes the price to fall. A lower price increases demand . . ." and so on, without end. The price keeps bobbing up and down, forever.

What's the mistake here? The first two statements ("a rise in income causes an increase in demand" and "an increase in demand causes price to rise") are entirely correct. But the next statement ("a higher price causes an increase in supply") is flat wrong, and so is everything that follows. A higher price does *not* cause an "increase in supply" (a shift of the supply curve). It causes an increase in *quantity supplied* (a movement *along* the supply curve).

Here's the correct sequence of events: "A rise in income causes an increase in demand. An increase in demand causes price to rise. A higher price causes an increase in *quantity supplied,* moving us along the supply curve until we reach the new equilibrium, with a higher price and greater quantity." End of story.

In this example, income rose. But *any* event that shifted the demand curve rightward would have the same effect on price and quantity. For example, if tastes changed in favor of maple syrup, or a substitute good like jam rose in price, or a complementary good like pancake mix became cheaper, the demand curve for maple syrup would shift rightward, just as it did in Figure 10. So, we can summarize our findings as follows:

> *A rightward shift in the demand curve causes a rightward movement* along *the supply curve. Equilibrium price and equilibrium quantity both rise.*

Example: Bad Weather, Supply Decreases

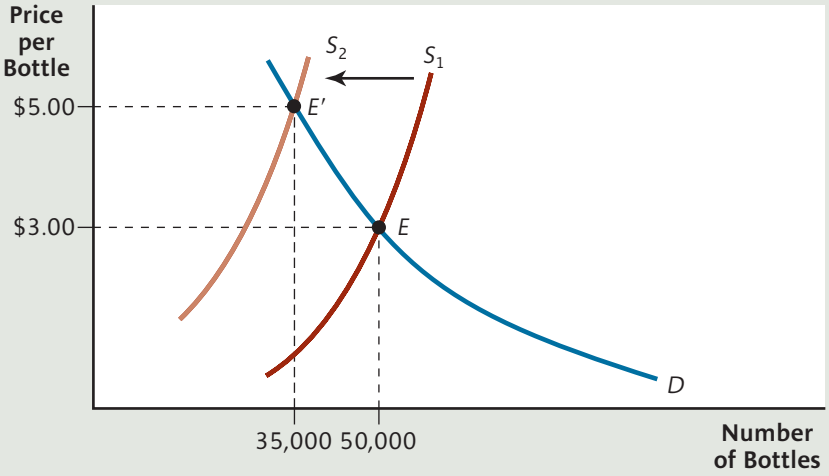
Bad weather can affect supply for most agricultural goods, including maple syrup. An example occurred in January 1998, when New England and Quebec were struck by a severe ice storm. Hundreds of thousands of maple trees were downed, and many more were damaged. In Vermont alone, 10 percent of the maple trees were destroyed. How did this affect the market for maple syrup?

As you've learned, weather can be shift-variable for the supply curve. Look at Figure 11. Initially, the supply curve for maple syrup is S_1, with the market in equilibrium at Point E. When bad weather hits, the supply curve shifts leftward—say, to S_2. The result: a rise in the equilibrium price of maple syrup (from $3.00 to $5.00 in the figure) and a fall in the equilibrium quantity (from 50,000 to 35,000 bottles).

Any event that shifts the supply curve leftward would have similar effects. For example, if the wages of maple syrup workers increase, or some maple syrup

FIGURE 11 A Shift of Supply and a New Equilibrium

An ice storm causes supply to decrease from S_1 to S_2. At the old equilibrium price of $3.00, there is now an excess demand. As a result, the price increases until excess demand is eliminated at point E′. In the new equilibrium, quantity demanded again equals quantity supplied. The price is higher, and fewer bottles are produced and sold.



producers go out of business and sell their farms to housing developers, the supply curve for maple syrup would shift leftward, just as in Figure 11.

More generally,

> *A leftward shift of the supply curve causes a leftward movement* along *the demand curve. Equilibrium price rises, but equilibrium quantity falls.*

Example: Higher Income and Bad Weather Together

So far, we've considered examples in which just one curve shifts due to a change in a single variable that influences *either* demand or supply. But what would happen if two changes affected the market simultaneously? Then both curves would shift.

Figure 12 shows what happens when we take the two factors we've just explored separately (a rise in income and bad weather) and combine them together. The rise in income causes the demand curve to shift rightward, from D_1 to D_2. The bad weather causes the supply curve to shift leftward, from S_1 to S_2. The result of all this is a change in equilibrium from point E to point E', where the new demand curve D_2 intersects the new supply curve S_2.

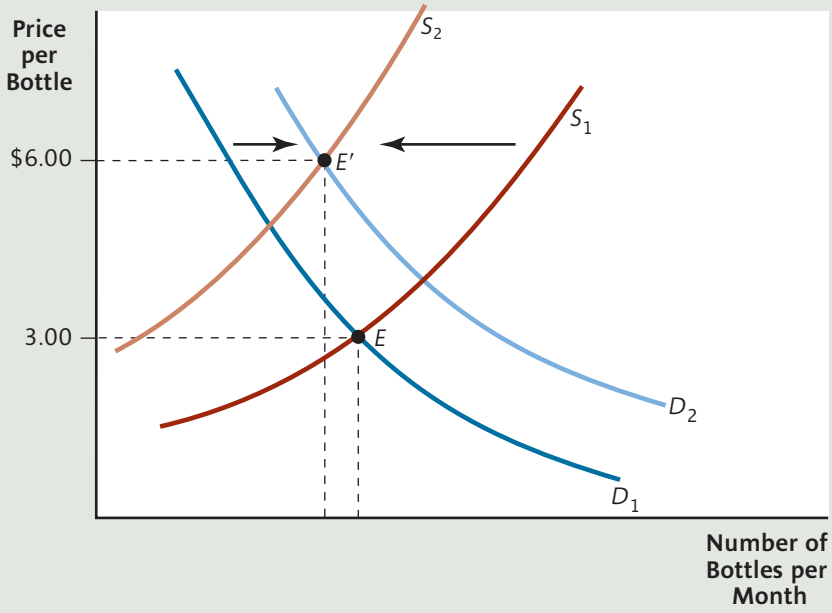
Notice that the equilibrium price rises from $3.00 to $6.00 in our example. This should come as no surprise. A rightward shift in the demand curve, with no other change, causes price to rise. And a leftward shift

dangerous curves

Do Curves Shift Up and Down? Or Right and Left? when describing an increase in demand or supply, it's tempting to say "upward" instead of "rightward." Similarly, for a decrease, it's tempting to say "downward." But be careful! While this interchangeable language works for the demand curve, it does *not* work for the supply curve.

To prove this to yourself, look at Figure 6. There you can see that a rightward shift of the supply curve (an increase in supply) is also a *downward* shift of the curve. In later chapters, it will sometimes make sense to describe shifts as upward or downward. For now, it's best to avoid these terms and stick with *rightward* and *leftward*.

FIGURE 12 A Shift in Both Curves and a New Equilibrium



An increase in income shifts the demand curve rightward from D_1 to D_2. At the same time, bad weather shifts the supply curve leftward from S_1 to S_2. The equilibrium moves from point E to point E′. While the price must rise after these shifts, quantity could rise or fall or remain the same, depending on the relative sizes of the shifts. In the figure, quantity happens to fall.

in the supply curve, with no other change, causes price to rise. So when we combine the two shifts together, the price must rise. In fact, the increase in the price will be greater than would be caused by either shift alone.

But what about equilibrium quantity? Here, the two shifts work in *opposite* directions. The rightward shift in demand works to increase quantity, while the leftward shift in supply works to decrease quantity. We can't say what will happen to equilibrium quantity until we know which shift is greater and thus has the greater influence. Quantity could rise, fall, or remain unchanged.

In Figure 12, it just so happens that the supply curve shifts more than the demand curve, so equilibrium quantity falls. But you can easily prove to yourself that the other outcomes are possible. First, draw a graph where the demand curves shifts rightward by more than the supply curve shifts leftward. In your graph, you'll see that equilibrium quantity rises. Then, draw one where both curves shift (in opposite directions) by equal amounts, and you'll see that equilibrium quantity remains unchanged.

We can also imagine other combinations of shifts. A rightward or leftward shift in either curve can be combined with a rightward or leftward shift in the other.

Table 6 lists all the possible combinations. It also shows what happens to equilibrium price and quantity in each case, and when the result is ambiguous (a question mark). For example, the top left entry tells us that when both the supply and demand curves shift rightward, the equilibrium *quantity* will always rise, but the equilibrium price could rise, fall, or remain unchanged, depending on the relative *size* of the shifts.

Do *not* try to memorize the entries in Table 6. Instead, remember the advice in Chapter 1: to study economics actively, rather than passively. This would be a good time to put down the book, pick up a pencil and paper, and see whether you can

TABLE 6

Effect of Simultaneous Shifts in Supply and Demand

	Increase in Demand (Rightward Shift)	No Change in Demand	Decrease in Demand (Leftward Shift)
• **Increase in Supply (Rightward Shift)**	$P\uparrow$ $Q\uparrow$	$P\downarrow$ $Q\uparrow$	$P\downarrow$ Q?
• **No change in Supply**	$P\uparrow$ $Q\uparrow$	No change in P or Q	$P\downarrow$ $Q\downarrow$
• **Decrease in Supply (Leftward Shift)**	$P\uparrow$ Q?	$P\uparrow$ $Q\downarrow$	P? $Q\downarrow$

draw a graph to illustrate each of the nine possible results in the table. When you see a question mark (?) for an ambiguous result, determine which shift would have to be greater for the variable to rise or to fall.

The Three-Step Process

In this chapter, we built a model—a supply and demand model—and then used it to analyze price changes in several markets. You may not have noticed it, but we took three distinct steps as the chapter proceeded. Economists take these same three steps to answer many questions about the economy, as you'll see throughout this book.

Let's review these steps:

Step 1—Characterize the Market: *Decide which market or markets best suit the problem being analyzed, and identify the decision makers (buyers and sellers) who interact there.*

In economics, we make sense of the very complex, real-world economy by viewing it as a collection of *markets*. Each of these markets involves a group of *decision makers*—buyers and sellers—who have the potential to trade with each other. At the very beginning of our analysis, we must decide which market or markets to look at (such as the U.S. market for maple syrup).

Step 2—Find the Equilibrium: *Describe the conditions necessary for equilibrium in the market, and a method for determining that equilibrium.*

Once we've defined a market, and put buyers and sellers together, we look for the point at which the market will come to rest—the equilibrium. In this chapter, we used supply and demand to find the equilibrium price and quantity in a perfectly competitive market, but this is just one example of how economists apply Step 2.

Step 3—What Happens When Things Change: *Explore how events or government policies change the market equilibrium.*

Once you've found the equilibrium, the next step is to ask how different events will *change* it. In this chapter, for example, we explored how rising income or bad weather (or both together) would affect the equilibrium price and quantity for maple syrup.

Economists follow this same three-step procedure to analyze important questions in both microeconomics and macroeconomics. In this book, we'll be taking these three steps again and again, and we'll often call them to your attention.

Using the Theory

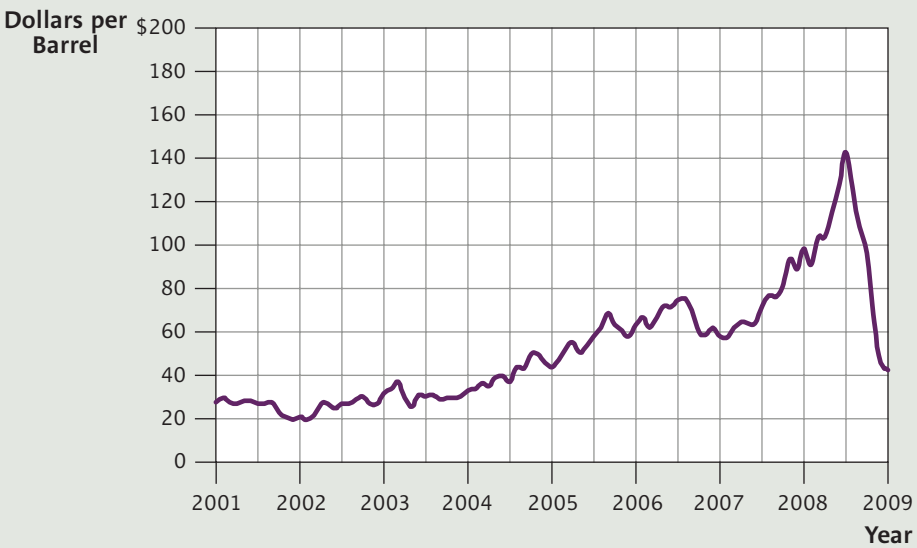
THE OIL PRICE SPIKE OF 2007–2008

Everyday, the world produces more than 80 million barrels of oil and uses up the same amount to produce almost every good and service we enjoy. So when the price of oil changes, every part of the economy is affected.

Figure 13 plots the average monthly price of oil from January 2001 through early 2009. These are prices in the *spot market*—where crude oil is bought and sold for delivery either immediately or within a few weeks after being pumped out of the ground. Notice the price spike from January 2007 to mid-2008. During this short period, oil prices rose from $58 to $143 per barrel. As a result, gasoline prices shot up around the globe. In the U.S., for example, the price of a gallon of regular gas almost doubled from $2.27 to $4.11 per gallon, making drivers unhappy and causing various culprits to be blamed on blogs, cable TV talk shows, and even U.S. Congressional hearings.

One popular culprit was *OPEC* (The Organization of Petroleum Exporting Countries). The 12 OPEC nations were accused of limiting production to drive up the market price. Another culprit was *speculators* (traders who buy something now expecting the price to rise so they can sell at a profit). Blame was even

FIGURE 13 Crude Oil Prices: 2001–2009



Average Price of West Texas Intermediate crude in Cushing, OK in first week of each calendar month.
Source: Energy Information Administration (http://www.eia.doe.gov/)

spread to a *conspiracy* of speculators, who must have been coordinating their actions to drive the price higher.

Many economists were skeptical about these explanations. To understand why, we'll use supply and demand, along with the three-step process discussed in this chapter.

Characterizing the Market

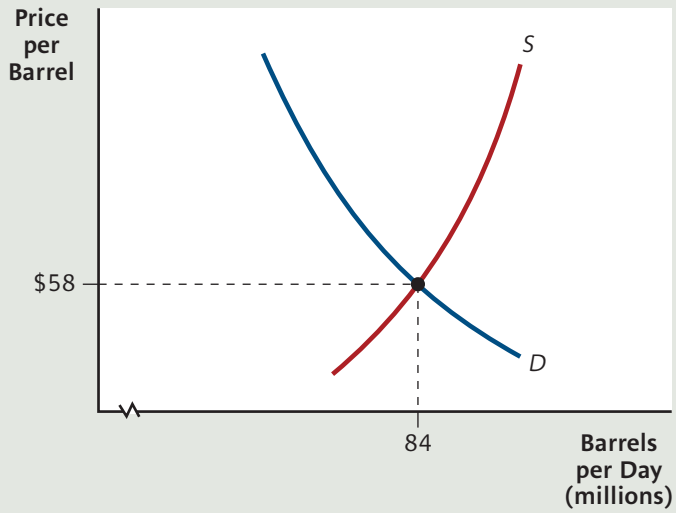
How should we characterize the market? First, we'll view it as a *global market*. It costs just a couple of dollars to ship a barrel of oil halfway around the world, so oil sellers and oil buyers around the globe can easily trade with each other. And our goal is to explain why oil prices rose worldwide, rather than in any particular country or region.

Second, because we want to explain why prices for *all* types of oil rose together, we'll look at the market for *all types* of crude oil—regardless of weight or sulfur content or any other quality. Finally, we'll regard the market as *competitive,* enabling us to use supply and demand. True, about one-third of the world's oil is produced by OPEC, which manipulates its members' total quantity to influence the world price. This part of the market is not competitive. But the remaining two-thirds of the world's oil *is* traded under competitive conditions—with many buyers and sellers, each of whom takes the market price as given. We'll regard OPEC's supply decisions as a shift variable for the market supply curve.

Finding the Equilibrium

Figure 14 shows market supply and demand curves for oil. The supply curve slopes upward: A higher price, *ceteris paribus,* increases the total quantity supplied by private and state-owned oil-producing firms in about 50 countries. (Even when OPEC's quantity is fixed, the *total* quantity supplied will rise with the price.) The demand curve slopes downward: A higher price, *ceteris paribus,* reduces the quantity

FIGURE 14 **Equilibrium in the Oil Market: January 2007**



demanded by oil-buying firms around the world, including oil refineries, electric power plants, plastics makers, and others. These firms use oil to make goods and services (gasoline, jet fuel, electricity, plastic bags) that they sell to households or other firms.

The equilibrium price occurs where quantity supplied and demanded are equal. In Figure 14, the equilibrium price is the one in early January 2007, when it averaged $58 per barrel, and world oil production averaged 84 million barrels per day.

What Happens When Things Change?

When we looked at the market for maple syrup, we observed how some event (such as an ice storm) changed the equilibrium. Here, we'll do the reverse. We know *what* happened (the equilibrium price rose) and we ask: What caused it?

Let's first consider the three most popular media explanations at the time.

Popular Explanation #1: Speculation in the Futures Market

In the oil *futures* market, traders sign contracts promising to buy or sell oil months or even years into the future, at a price stated in the contract. Starting in the mid-2000s, the oil futures market became increasingly popular for speculators. Pension funds, hedge funds, and even ordinary households began buying oil futures contracts in huge quantities, betting that the price of oil would continue rising. Could this sort of speculation explain the oil price spike in the *spot* market, as seen in Figure 13?

Not directly. Futures contracts are virtually always settled in cash, with no one actually buying or selling actual oil to settle a contract. A futures contract is like a bet on tomorrow's football game. The "game" in this analogy is the spot market for oil—where barrels of physical oil are actually bought and sold. Just as a football bet doesn't affect the outcome of tomorrow's game, a bet on the future price of oil has no direct impact on what that price will actually be when buyers and sellers do their trading.

But the futures market can *indirectly* influence spot market oil prices. For example, suppose that higher futures-market prices make everyone *think* that oil prices are, in fact, heading up. As you've learned, in many markets such an *expected* price increase can cause supply and demand curves to shift *now*. Buyers might want to purchase more oil than they currently need, shifting the demand curve rightward. Or producers might want to cut back on sales, shifting the supply curve leftward. So the next step in our analysis is to ask: Did producers cut back production? Did buyers purchase and hoard oil for future use? Let's look first at producers.

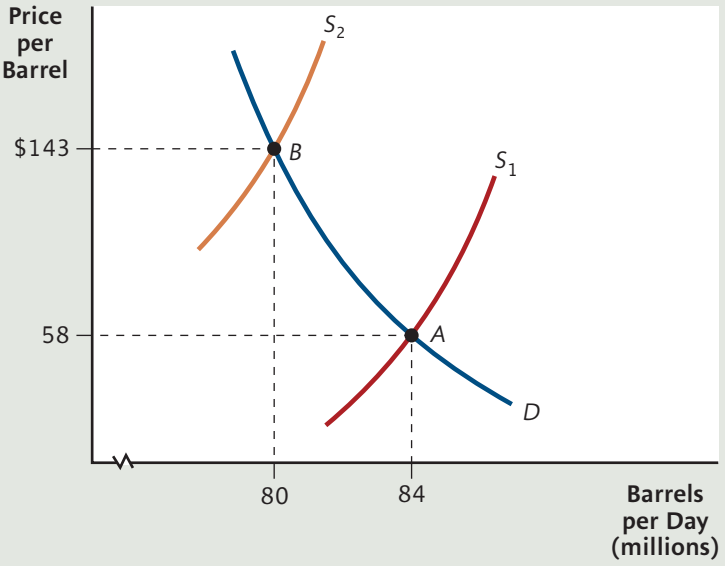
Popular Explanation #2: Cutbacks by Producers

Figure 15 shows what would have happened if, *ceteris paribus,* OPEC or some other group of suppliers decreased supply (say, because they expected higher prices later on or were actually trying to manipulate the price). The market supply curve shifts leftward—from S_1 to S_2—and the new equilibrium price rises to $143. Such a leftward shift of the supply curve could explain the rise in price, but notice that when the supply curve shifts leftward, equilibrium quantity falls (down to 80 million barrels). Does this match the facts?

No. From 2007 to 2008, the quantity of oil produced and sold *rose,* in contradiction to the prediction in Figure 15. Even OPEC raised its production during this period. Thus a leftward shifting supply curve *alone* cannot explain what happened.

FIGURE 15 A Hypothetical Decrease in Oil Supply

If a leftward shift of the supply curve had been the reason for higher oil prices, the equilibrium quantity would have decreased. As an example, the equilibrium would have moved from point A to a point like B, with equilibrium quantity falling to 80 million barrels per day.



While we cannot rule out a leftward shift in the supply curve, it can—at best—be only *part* of the explanation.

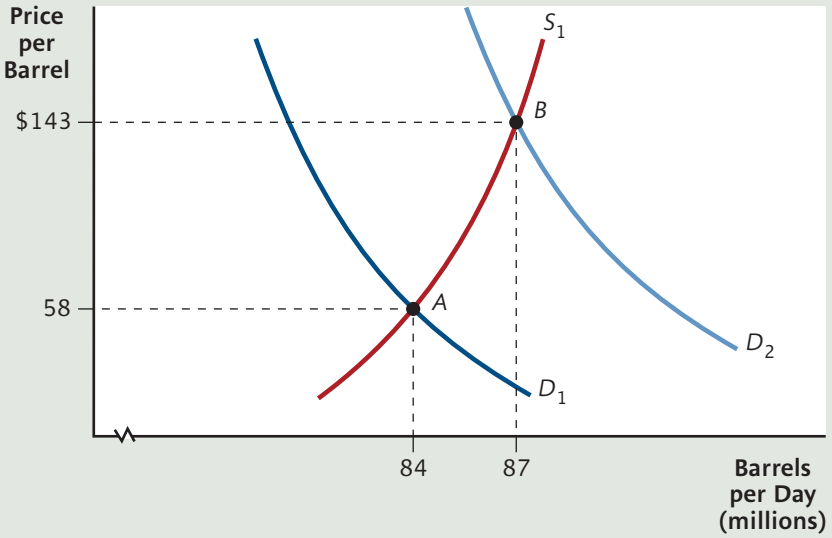
Popular Explanation #3: Manipulation, Speculation, or Hoarding by Buyers

Figure 16 shows what would have happened if, *ceteris paribus,* buyers increased their demand for oil, because they expected higher prices later. Oil users might have decided to buy more than they needed at the time, stocking up for the future. Or speculators might have hoarded oil in order to profit by selling it later. Or maybe buyers were conspiring to drive the price up by taking oil off the market and storing it somewhere. In any of these scenarios, the demand curve shifts rightward, from D_1 to D_2. Equilibrium price rises from $58 to $143, and equilibrium quantity rises too (from 84 million to 87 million barrels per day). This matches what actually happened to price and quantity from January 2007 to July 2008.

In fact, a rightward shift in the demand curve *must* be part of our explanation. (Look back at Table 6 in the chapter, which tells us that—regardless of what happens to supply—demand must increase in order for both price and quantity to rise.) But could buyer manipulation, speculation, or hoarding in the spot market have been the reason?

Not likely. First, a conspiracy large enough to actually manipulate the market price would have to involve many buyers, and the secret would be hard to keep. But second—and more importantly—to increase demand in the *spot* market, the manipulators, speculators, or hoarders would have to buy actual oil and store it somewhere for later sale or use. Storing oil is both expensive and difficult. And the data for the U.S., Japan, and Europe show no buildup of oil inventories during this period. Since we have no inventory evidence to support manipulation, speculation, or hoarding, this explanation is not very satisfying.

FIGURE 16 An Increase in Oil Demand



An increase in the demand for oil would raise both price and *quantity. The graph shows the rise in equilibrium price (to $143) and equilibrium quantity (to 87 million barrels per day) that actually occurred from January 2007 to July 2008.*

The Least-Interesting but Most Logical Answer

We've concluded that the demand curve for oil in the spot market *must* have shifted rightward (either by itself or in combination with a supply shift). Otherwise, both price and quantity could not have risen. We've also concluded that speculation, manipulation and buyer-hoarding are unlikely explanations for the demand shift. What else is left?

The answer would not make a good premise for a Hollywood thriller. It is far less interesting than crazed speculators or villainous manipulators. The demand for oil seems to have increased because—quite simply—firms wanted to *use* more oil at any given price. And they wanted to *use* more because households around the world—with rapidly growing incomes—wanted to buy more goods and services produced from oil.

Indeed, global income growth had been increasing the demand for oil throughout the 2000s. But in 2007 and early 2008, income grew even more rapidly, and so did oil demand. This rapid increase in demand by oil-using firms appears to have raised price and quantity, without increasing oil inventories.

Now, if rightward shifts in the supply curve had kept pace with the rightward shifts in the demand curve, more oil could have been produced and sold without any rise in price. But in 2007 and early 2008, no major new oil field was brought into production, and no new cost-saving technology was discovered. Nothing happened that might have shifted the supply curve significantly rightward.

Now look back at Figure 13, where you can see that the price of oil *dropped* rapidly after mid-2008. In this case, a spreading global recession slowed growth and actually *reduced* economic activity in some of the largest oil-consuming countries, shifting the demand curve leftward. The OPEC countries tried to prevent the price from falling too rapidly by coordinating a decrease in production (shifting the supply curve leftward), but they seemed unable to shift it fast enough or far enough to prevent a rapid price drop.

In mid-2008, as oil prices fell, the public had no difficulty understanding and accepting economists' central explanation: that falling incomes were decreasing the demand for oil, causing the price to drop. Yet just six months earlier, many could not accept the same logic in the other direction: that rising incomes were increasing the demand for oil, causing the price to rise.

SUMMARY

In a market economy, prices are determined through the interaction of buyers and sellers in *markets*. *Perfectly competitive* markets have many buyers and sellers, and none of them individually can affect the market price. If an individual, buyer, or seller has the power to influence the price of a product, the market is *imperfectly competitive*.

The model of *supply and demand* explains how prices are determined in perfectly competitive markets. The *quantity demanded* of any good is the total amount buyers would choose to purchase given the constraints that they face. The *law of demand* states that quantity demanded is negatively related to price; it tells us that the *demand curve* slopes downward. The demand curve is drawn for given levels of income, wealth, tastes, prices of substitute and complementary goods, population, and expected future price. If any of those factors changes, the demand curve will shift. A change in price, however, moves us *along* the demand curve.

The *quantity supplied* of a good is the total amount sellers would choose to produce and sell given the constraints that they face. According to the *law of supply,* supply curves slope upward. The supply curve will shift if there is a change in the price of an input, the price of an alternate good, the price in an alternate market, the number of firms, expectations of future prices, or (for some goods) a change in weather. A change in the price of the good, by contrast, moves us *along* the supply curve.

Equilibrium price and quantity in a market are found where the supply and demand curves intersect. If either or both of these curves shift, price and quantity will change as the market moves to a new equilibrium.

Economists frequently use a three-step process to answer questions about the economy. The three steps—taken several times in this chapter—are to (1) characterize the market or markets involved in the question; (2) find the equilibrium in the market; and (3) ask what happens when something changes.

PROBLEM SET

Answers to even-numbered questions and problems can be found on the text website at www.cengage.com/economics/hall.

1. Consider the following statement: "In late 2008, as at other times in history, oil prices came down at the same time as the quantity of oil produced fell. Therefore, one way for us to bring down oil prices is to slow down oil production." True or false? Explain.
2. In the late 1990s and through 2000, the British public became increasingly concerned about "Mad Cow Disease," which could be deadly to humans if they ate beef from these cattle. Fearing the disease, many consumers switched to other meats, like chicken, pork, or lamb. At the same time, the British government ordered the destruction of thousands of head of cattle. Illustrate the effects of these events on the equilibrium price and quantity in the market for British beef. Can we determine with certainty the direction of change for the quantity? For the price? Explain briefly.
3. Discuss, and illustrate with a graph, how each of the following events will affect the market for coffee:
   a. A blight on coffee plants kills off much of the Brazilian crop.
   b. The price of tea declines.
   c. Coffee workers organize themselves into a union and gain higher wages.
   d. Coffee is shown to cause cancer in laboratory rats.
   e. Coffee prices are expected to rise rapidly in the near future.

4. The following table gives hypothetical data for the quantity of two-bedroom rental apartments demanded and supplied in Peoria, Illinois:

Monthly Rent	Quantity Demanded (thousands)	Quantity Supplied (thousands)
$800	30	10
$1,000	25	14
$1,200	22	17
$1,400	19	19
$1,600	17	21
$1,800	15	22

   a. Graph the demand and supply curves.
   b. Find the equilibrium price and quantity.
   c. Explain briefly why a rent of $1,000 cannot be the equilibrium in this market.
   d. Suppose a tornado destroys a significant number of apartment buildings in Peoria, but doesn't affect people's desire to live there. Illustrate on your graph the effects on equilibrium price and quantity.

5. The following table gives hypothetical data for the quantity of alarm clocks demanded and supplied per month.

Price per Alarm Clock	Quantity Demanded	Quantity Supplied
$ 5	3,500	700
$10	3,000	900
$15	2,500	1,100
$20	2,000	1,300
$25	1,500	1,500
$30	1,000	1,700
$35	500	1,900

   a. Graph the demand and supply curves.
   b. Find the equilibrium price and quantity.
   c. Illustrate on your graph how a decrease in the price of telephone wake-up services would affect the market for alarm clocks.
   d. What would happen if there was a decrease in the price of wake-up services at the same time that the price of the plastic used to manufacture alarm clocks rose?

6. The table at the end of this problem gives hypothetical data for the quantity of electric scooters demanded and supplied per month.
   a. Graph the demand and supply curves.
   b. Find the equilibrium price and quantity.
   c. Illustrate on your graph how an increase in the wage rate paid to scooter assemblers would affect the market for electric scooters.
   d. What would happen if there was an increase in the wage rate paid to scooter assemblers at the same time that tastes for electric scooters increased?

Price per Electric Scooter	Quantity Demanded	Quantity Supplied
$150	500	250
$175	475	350
$200	450	450
$225	425	550
$250	400	650
$275	375	750

7. The following table gives hypothetical data for the quantity of gasoline demanded and supplied in Los Angeles per month.

Price per Gallon	Quantity Demanded (millions of gallons)	Quantity Supplied (millions of gallons)
$1.20	170	80
$1.30	156	105
$1.40	140	140
$1.50	123	175
$1.60	100	210
$1.70	95	238

   a. Graph the demand and supply curves.
   b. Find the equilibrium price and quantity.
   c. Illustrate on your graph how a rise in the price of automobiles would affect the gasoline market.

8. How would each of the following affect the market for blue jeans in the United States? Illustrate each answer with a supply and demand diagram.
   a. The price of denim cloth increases.
   b. An economic slowdown in the United States causes household incomes to decrease.

9. Indicate which curve shifted—and in which direction—for each of the following. Assume that only one curve shifts.
   a. The price of furniture rises as the quantity bought and sold falls.
   b. Apartment vacancy rates increase while average monthly rent on apartments declines.
   c. The price of personal computers continues to decline as sales skyrocket.

10. Consider the following forecast: "Next month, we predict that the demand for oranges will continue to increase, which will tend to raise the price of oranges. However, the higher price will increase supply, and a greater supply tends to lower prices. Accordingly, even though we predict that demand will increase next month, we cannot predict whether the price of oranges will rise or fall." There is a serious mistake of logic in this forecast. Can you find it? Explain.

11. A couple of months after Hurricane Katrina, an article in *The New York Times* contained the following passage: "Gasoline prices—the national average is now $2.15, according to the Energy Information Administration—have fallen because higher prices held down demand and Gulf Coast supplies have been slowly restored."[1] The statement about supply is entirely correct and explains why gas prices came down. But the statement about demand confuses two concepts you learned about in this chapter.
    a. What two concepts does the statement about demand seem to confuse? Explain briefly.
    b. On a supply and demand diagram, show what most likely caused gasoline prices to rise when Hurricane Katrina shut down gasoline refineries on the Gulf Coast.
    c. On another supply and demand diagram, show what most likely happened in the market for gasoline as Gulf Coast refineries were repaired—and began operating again—after the Hurricane.
    d. What role did the *demand* side of the market play in explaining the rise and fall of gas prices?

12. Draw supply and demand diagrams for market *A* for each of the following. Then use your diagrams to illustrate the impact of the following events. In each case, determine what happens to price and quantity in each market.
    a. *A* and *B* are substitutes, and the price of good *B* rises.
    b. *A* and *B* satisfy the same kinds of desires, and there is a shift in tastes away from *A* and toward *B*.
    c. *A* is a normal good, and incomes in the community increase.
    d. There is a technological advance in the production of good *A*.
    e. *B* is an input used to produce good *A*, and the price of *B* rises.

13. The Using the Theory section of this chapter points out that when we observe an increase in both price and quantity, we know that the demand curve must have shifted rightward. However, we cannot rule a shift in the supply curve as well. Prove this by drawing a supply and demand graph for each of the following cases:
    a. Demand curve shifts rightward, supply curve shifts leftward, equilibrium price and quantity both rise.
    b. Demand and supply curves both shift rightward, equilibrium price and quantity both rise.
    c. Evaluate the following statement: "During the oil price spike from 2007 to mid-2008, we know the supply curve could not have shifted leftward, because quantity supplied rose." True or False? Explain.
    d. "During the oil price spike from 2007 to mid-2008, the supply curve may have shifted leftward (say, because a rise in expected price), but the demand curve must have shifted rightward as well." True of False? Explain.

14. In the second half of 2008, as the equilibrium price and quantity of oil both decreased, OPEC succeeded in reducing oil production by its members, but could not prevent the dramatic drop in price. Illustrate both the original equilibrium (before the price drop) and the final equilibrium (after the price drop) on a supply and demand diagram.

More Challenging

15. Suppose that demand is given by the equation $Q^D = 500 - 50P$, where Q^D is quantity demanded, and P is the price of the good. Supply is described by the equation $Q^S = 50 + 25P$, where Q^S is quantity supplied. What is the equilibrium price and quantity? (See Appendix.)

16. While crime rates fell across the country over the past few decades, they fell especially rapidly in Manhattan. At the same time, there were some neighborhoods in the New York metropolitan area in which the crime rate remained constant. Using supply and demand diagrams for rental housing, explain how a falling crime rate in Manhattan could make the residents in *other* neighborhoods *worse off*. (Hint: As people from around the country move to Manhattan, what happens to rents there? If people already living in Manhattan cannot afford to pay higher rent, what might they do?)

17. An analyst observes the following equilibrium price-quantity combinations in the market for restaurant meals in a city over a four-year period:

Year	*P*	*Q* (thousands of meals per month)
1	$12	20
2	$15	30
3	$17	40
4	$20	50

She concludes that the market defies the law of demand. Is she correct? Why or why not?

[1] "Economic Memo: Upbeat Signs Hold Cautions for the Future," *New York Times*, November 30, 2005.

APPENDIX

Solving for Equilibrium Algebraically

In the body of this chapter, notice that the supply and demand curves for maple syrup were *not* graphed as straight lines. This is because the data they were based on (as shown in the tables) were not consistent with a straight-line graph. You can verify this if you look back at Table 1: When the price rises from $1.00 to $2.00, quantity demanded drops by 15,000 (from 75,000 to 60,000). But when the price rises from $2.00 to $3.00, quantity demanded drops by 10,000 (from 60,000 to 50,000). Since the change in the independent variable (price) is $1.00 in both cases, but the change in the dependent variable (quantity demanded) is different, we know that when the relationship between quantity demanded and price is graphed, it will not be a straight line.

We have no reason to expect demand or supply curves in the real world to be straight lines (to be *linear*). However, it's often useful to approximate a curve with a straight line that is reasonably close to the original curve. One advantage of doing this is that we can then express both supply and demand as simple equations, and solve for the equilibrium using basic algebra.

For example, suppose the demand for take-out pizzas in a modest-size city is represented by the following equation:

$$Q^D = 64{,}000 - 3{,}000\,P$$

where Q^D stands for the quantity of pizzas demanded per week. This equation tells us that every time the price of pizza rises by $1.00, the number of pizzas demanded each week *falls* by 3,000. As we'd expect, there is a negative relationship between price and quantity demanded. Moreover, since quantity demanded always falls at the same rate (3,000 fewer pizzas for every $1.00 rise in price), the equation is linear.[2]

Now we'll add an equation for the supply curve:

$$Q^S = -20{,}000 + 4{,}000\,P$$

where Q^S stands for the quantity of pizzas supplied per week. This equation tells us that when the price of pizza rises by $1.00, the number of pizzas supplied per week *rises* by 4,000—the positive relationship we expect of a supply curve.[3] And like the demand curve, it's linear: Quantity supplied continues to rise at the same rate (4,000 more pizzas for every $1.00 increase in price).

We know that if this market is in equilibrium, quantity demanded (Q^D) will equal quantity supplied (Q^S). So let's *impose* that condition on these curves. That is, let's require $Q^D = Q^S$. This allows us to use the definitions for Q^D and Q^S that have price as a variable, and set those equal to each other in equilibrium:

$$64{,}000 - 3{,}000\,P = -20{,}000 + 4{,}000\,P$$

This is one equation with a single unknown—P—so we can use the rules of algebra to isolate P on one side of the equation. We do this by adding 3,000 P to both sides, which isolates P on the right, and adding 20,000 to both sides, which moves everything that *doesn't* involve P to the left, giving us:

$$84{,}000 = 7{,}000\,P$$

Finally, dividing both sides by 7,000 gives us

$$84{,}000/7{,}000 = P$$

or

$$P = 12$$

We've found our equilibrium price: $12.

[2] If you try to graph the demand curve, don't forget that supply and demand graphs reverse the usual custom of where the independent and dependent variables are plotted. Quantity demanded is the dependent variable (it *depends* on price), and yet it's graphed on the *horizontal* axis.

[3] Don't be troubled by the negative sign (−20,000) in this equation. It helps determine a minimum price that suppliers must get in order to supply any pizza at all. Using the entire equation, we find that if price were $5.00, quantity supplied would be zero, and that price has to rise *above* $5.00 for any pizzas to be supplied in this market. But since a "negative supply" doesn't make sense, this equation is valid only for prices of $5.00 or greater.

What about equilibrium quantity? In equilibrium, we know quantity demanded and quantity supplied are equal, so we can *either* solve for Q^D using the demand equation, or solve for Q^S using the supply equation, and we should get the same answer. For example, using the demand equation, and using the equilibrium price of $12:

$$Q^D = 64{,}000 - 3{,}000\ (12)$$

or

$$Q^D = 28{,}000$$

To confirm that we didn't make any errors, we can also use the supply equation.

$$Q^S = -20{,}000 + 4{,}000\ (12)$$

or

$$Q^S = 28{,}000$$

We've now confirmed that the equilibrium quantity is 28,000.

CHAPTER 4

Working with Supply and Demand

In the last chapter, we used supply and demand for a basic but important purpose: to explain how the interactions of buyers and sellers determine the price of a good or service, and how that price changes when the market is free to adjust to events.

But supply and demand can help us answer many other important questions. For example, can the government change the price in a market? And if so, are some ways of doing this more effective than others?

Supply and demand can also help us analyze other types of markets, and understand some of the events that have shaken economies around the globe. What caused the housing boom from the late 1990s to 2006? And why did the boom end so suddenly?

This chapter is all about working with supply and demand in new ways, and in different contexts. As you will see, the model—with some modifications—can help us answer all of these questions.

Government Intervention in Markets

The forces of supply and demand deserve some credit. They force the market price to adjust until something remarkable happens: The quantity that sellers want to sell is also the quantity that buyers want to buy. Thus, every buyer and seller can turn their intentions into actual market trades.

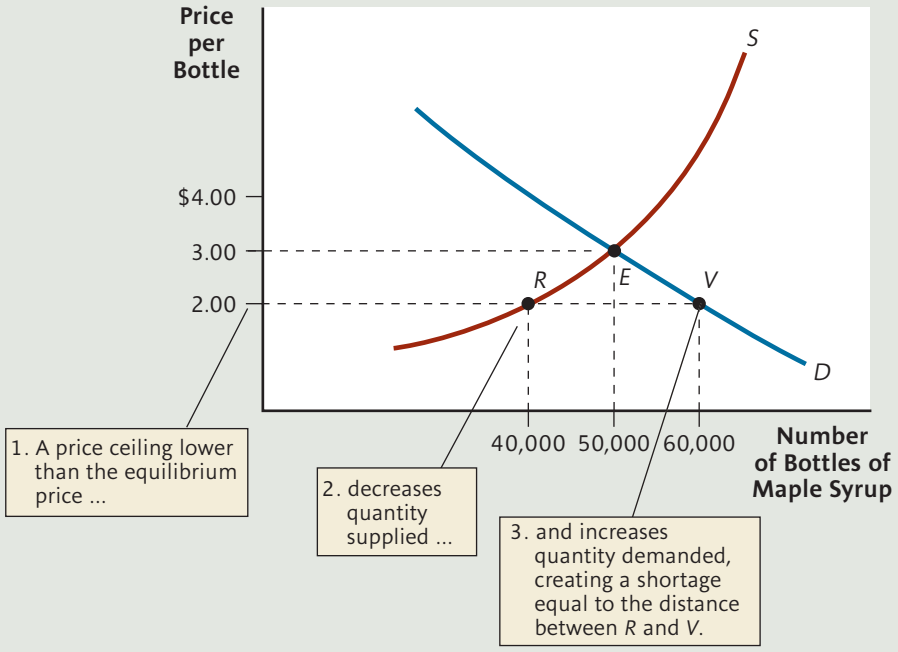
So, three cheers for supply and demand! Or better make that *two* cheers. Because while everyone agrees that having prices is necessary for the smooth functioning of our economy, not everyone is happy with the prices that supply and demand give us. Apartment dwellers often complain that their rent is too high, and farmers complain that the price of their crops is too low.

We can also be dissatisfied with market *quantities*. We might ask government to help increase the number of people attending college, or help decrease the quantity of gasoline that we use.

Responding to these dissatisfactions and desires, governments sometimes intervene to change the market outcome. And government can do so in a variety of ways.

We will first look at two policies in which the government tries to *fight* the market—that is, to prevent the price from reaching its equilibrium value. Economists are generally skeptical about the effectiveness and efficiency of these policies. Then we'll turn to methods government uses to *manipulate* markets—changing the equilibrium itself.

FIGURE 1 **A Price Ceiling in the Market for Maple Syrup**



Fighting the Market: Price Ceilings

Price ceiling A government-imposed maximum price in a market.

Figure 1 shows our familiar market for maple syrup, with an equilibrium price of $3.00 per bottle. Suppose that maple syrup buyers complain to the government that this price is too high. And suppose the government responds by imposing a **price ceiling** in this market—a regulation preventing the price from rising above, say, $2.00 per bottle.

If the ceiling is enforced, then producers will no longer be able to charge $3.00 for maple syrup but will have to content themselves with $2.00 instead. In Figure 1, we will move down along the supply curve, from point *E* to point *R*, decreasing quantity supplied from 50,000 bottles to 40,000. At the same time, the decrease in price will move us along the demand curve, from point *E* to point *V*, increasing quantity demanded from 50,000 to 60,000. Together, these changes in quantities supplied and demanded create an *excess demand* for maple syrup of 60,000 − 40,000 = 20,000 bottles each month. Ordinarily, the excess demand would force the price back up to $3.00. But now the price ceiling prevents this from occurring. What will happen?

A practical observation about markets can help us arrive at an answer:

Short side of the market The smaller of quantity supplied and quantity demanded at a particular price.

> *When quantity supplied and quantity demanded differ, the* ***short side of the market****—whichever of the two quantities is smaller—will prevail.*

Shortage An excess demand not eliminated by a rise in price, so that quantity demanded continues to exceed quantity supplied.

This simple rule follows from the voluntary nature of exchange in a market system: No one can be forced to buy or sell more than they want to. With an excess demand quantity supplied is less than quantity demanded, so *sellers* are on the short side of the market. Since we cannot force them to sell any more than they want to (40,000 units) the result is a **shortage** of maple syrup—not enough available to satisfy demand at the going price.

But this is not the end of the story. Because of the shortage, all 40,000 bottles produced each month will quickly disappear from store shelves, and many buyers will be disappointed. The next time people hear that maple syrup has become available, everyone will try to get there first, and we can expect long lines at stores. Those who really crave maple syrup may have to go from store to store, searching for that rare bottle. When we include the *opportunity cost* of the time spent waiting in line or shopping around, the ultimate effect of the price ceiling may be a *higher* cost of maple syrup for many consumers.

A price ceiling creates a shortage and increases the time and trouble required to buy the good. While the price decreases, the opportunity cost may rise.

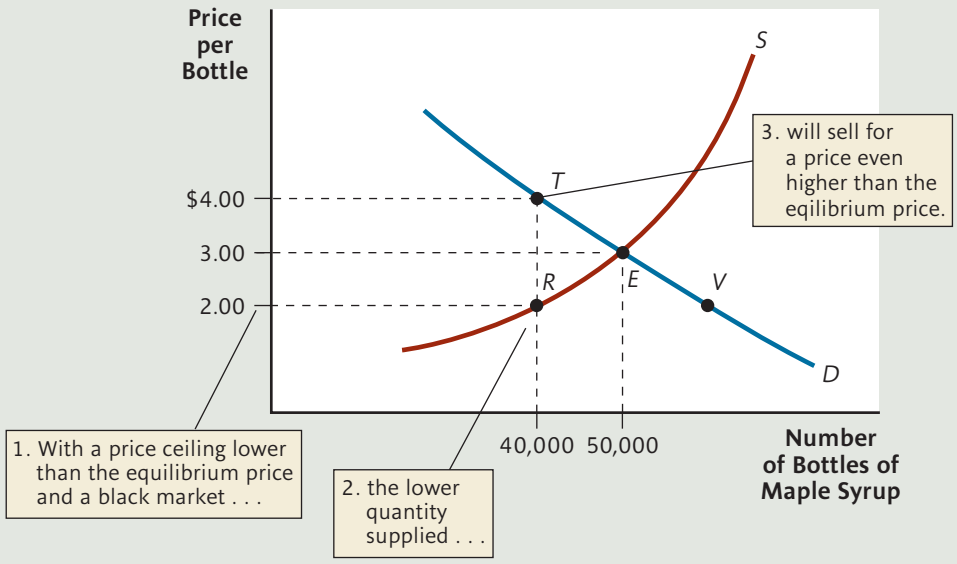
And there is still more. The government may be able to prevent maple syrup *producers* from selling above the price ceiling. But it may not be able to prevent a **black market**, where goods are sold illegally at prices higher than the legal ceiling.

Black market A market in which goods are sold illegally at a price above the legal ceiling.

Ironically, the black market price will typically exceed the original, freely determined equilibrium price—$3.00 per bottle in our example. To see why, look at Figure 2. With a price ceiling of $2.00, sellers supply 40,000 bottles per month. Suppose all of this is bought by people—maple syrup scalpers, if you will—who then sell it at the highest price they can get.

What price can they charge? We can use the demand curve to find out. At $4.00 per bottle (point *T*), the scalpers would just be able to sell all 40,000 bottles each month. If they charge more than $4, they would be stuck with unsold bottles every month, so they'd learn to lower the price. If they charged less than $4, quantity demanded would exceed the 40,000 bottles they are trying to sell, so they'd run out of maple syrup and see that they could safely raise the price. Thus, if the price ceiling remains in place, we'd expect the black market price to settle at $4 per bottle.

FIGURE 2 A Price Ceiling with a Black Market



The unintended consequences of price ceilings—long lines, black markets, and, often, higher prices—explain why they are generally a poor way to bring down prices. Experience with price ceilings has generally confirmed this judgment. Many states do have laws to limit price hikes during declared emergencies, thereby creating temporary price ceilings. But permanent or semipermanent price ceilings are exceedingly rare.

There is, however, one type of market in which several cities have imposed long-lasting price ceilings: the market for apartment rentals.

An Example: Rent Controls

Rent controls Government-imposed maximum rents on apartments and homes.

A price ceiling imposed in a rental housing market is called **rent control**. Most states have laws *prohibiting* rent control. But more than a dozen states do allow it. And in four of these states (New York, California, Maryland, and New Jersey), some form of rent control has existed in several cities and towns for decades.

In theory, rent control is designed to keep housing affordable, especially for those with low incomes. But for this purpose, it's a very blunt instrument because it doesn't target those with low incomes. Rather, *anyone* who was lucky enough to be living in one of the affected units when rent control was first imposed gets to pay less than market rent, as long as he or she continues to hold the lease on the unit. Many renters in cities such as New York and Santa Monica have higher incomes and living standards than do the owners from whom they rent.

Second, rent control causes the same sorts of problems as did our hypothetical price ceiling on maple syrup. It creates a persistent excess demand for rental units, so renters must spend more time and trouble finding an apartment. Typically, something akin to the "black market" develops: Real estate brokers quickly "snap up" the rent-controlled apartments (either because of their superior knowledge of the market, or their ability to negotiate exclusive contracts with the owners). Apartment seekers, who don't want to spend months searching on their own, will hire one of these brokers. Alternatively, one can sublet from a leaseholder, who will then charge market rent for the sublet and pocket the difference. Either way, many renters end up paying a higher cost for their apartment than the rent-controlled price.

Finally, rent controls cause a decrease in the quantity of apartments supplied (a movement along the supply curve). This is because lower rents reduce the incentives for owners to maintain existing apartments in rentable condition, and also reduce incentives to build new ones.

In our example of the market for maple syrup, the decrease in quantity supplied—combined with a black market—caused the average buyer to end up paying a higher price than before the ceiling was imposed (see point *T* in Figure 2). The same thing can happen in the apartment rental market. As supply decreases, the total price of renting an apartment for a few years can rise above the market equilibrium price—if we include real estate commissions or the unofficial, higher rents paid by those who sublet.

Fighting the Market: Price Floors

Price floor A government-imposed minimum price in a market.

Sometimes, governments try to help sellers of a good by establishing a **price floor**—a minimum amount below which the price is not permitted to fall. The most common use of price floors around the world has been to raise prices (or prevent prices from

falling) in agricultural markets. Price floors for agricultural goods are commonly called *price support programs*.

In the United States, price support programs began during the Great Depression, after farm prices fell by more than 50 percent between 1929 and 1932. The Agricultural Adjustment Act of 1933, and an amendment in 1935, gave the president the authority to intervene in markets for a variety of agricultural goods. Over the next 60 years, the United States Department of Agriculture (USDA) put in place programs to maintain high prices for cotton, wheat, rice, corn, tobacco, honey, milk, cheese, butter, and many other farm goods. Although some of these supports were removed in recent years, many remain. For example, government policy still maintains price floors for peanuts, sugar, and dairy products.

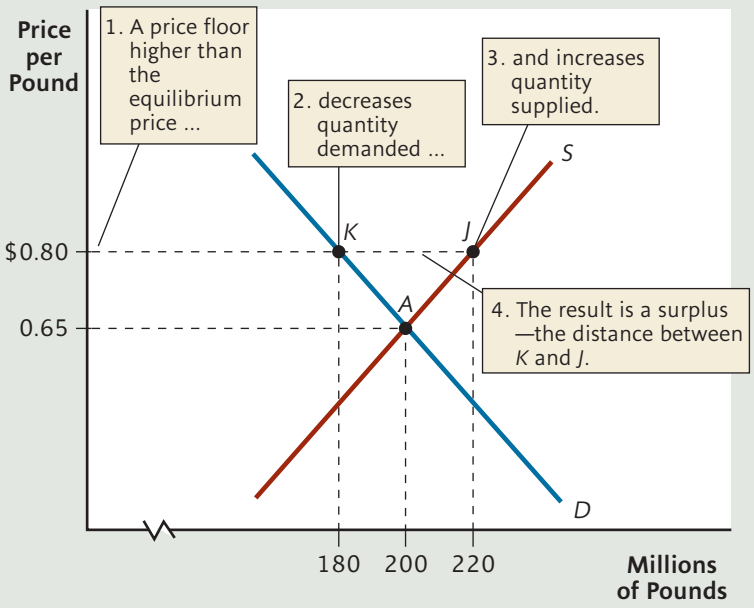
To see how price floors work, let's look at the market for nonfat dry milk—a market in which the USDA has been supporting prices continually since 1933. Figure 3 shows that—before any price floor is imposed—the market is in equilibrium at point *A*, with an equilibrium price of 65 cents per pound and an equilibrium quantity of 200 million pounds per month.

Now let's examine the impact of the price floor, recently set at $0.80 per pound. At this price, producers want to sell 220 million pounds, while consumers want to purchase only 180 million pounds. There is an excess supply of 220 million − 180 million = 40 million pounds. Our short-side rule tells us that buyers determine the amount actually traded. They purchase 180 million of the 220 million pounds produced, and producers are unable to sell the remainder.

The excess supply of 40 million pounds would ordinarily push the market price down to its equilibrium value: $0.65. But now the price floor prevents this from happening. The result is a **surplus**—continuing extra production of nonfat dry milk that no one wants to buy at the going price.

Surplus An excess supply not eliminated by a fall in price, so that quantity supplied continues to exceed quantity demanded.

FIGURE 3 A Price Floor in the Market for Nonfat Dry Milk



dangerous curves

Floor Above, Ceiling Below! It's tempting to think that a price floor set *under* the equilibrium price, or a price ceiling is *above* it. After all, a floor is usually on the bottom of something, and a ceiling is on the top. Right? In this case, wrong! A price floor *below* the equilibrium price would have no impact, because the market price would *already* satisfy the requirement that it be higher than the floor. Similarly, a price ceiling set *above* the equilibrium price would have no impact (make sure you understand why). So remember: Always draw an effective price floor *above* the equilibrium price and an effective price ceiling *below* the equilibrium price.

Maintaining a Price Floor

If the government merely *declared* a price floor of $0.80 per pound, many farmers who are unable to sell all of their product would be tempted to sell some illegally at a price below the floor. This would take sales away from other farmers trying to sell at the higher price, so they, too, would feel pressure to violate the floor. Soon, the price floor would collapse.

To prevent this, governments around the world have developed a variety of policies designed to prevent surplus goods from forcing down the price. One method, frequently used in the United States, is for the government to promise to buy any unsold product at a guaranteed price. In the market for nonfat dry milk, for example, the government agrees to buy any unsold supplies from sellers at a price of $0.80 per pound. With this policy, no supplier would ever sell at any price *below* $0.80, since it could always sell to the government instead. With the price effectively stuck at $0.80, private buyers buy 180 million pounds—point *K* on the demand curve in Figure 3. But since quantity supplied is 220 million, at point *J*, the government must buy the excess supply of 40 million pounds per year. In other words, the government maintains the price floor by *buying up* the entire excess supply. This prevents the excess supply from doing what it would ordinarily do: drive the price down to its equilibrium value.

A price floor creates a surplus of a good. In order to maintain the price floor, the government must prevent the surplus from driving down the market price. In practice, the government often accomplishes this goal by purchasing the surplus itself.

In 2009, for example, the USDA expected to purchase more than 100 million pounds of non-fat dry milk, at a cost of almost $100 million.

However, purchasing surplus food is expensive, so price floors are usually accompanied by government efforts to *limit* any excess supplies. In the dairy market, for example, the U.S. government has developed a complicated management system to control the production and sale of milk to manufacturers and processors, which helps to limit the government's costs. In other agricultural markets, the government has ordered or paid farmers *not* to grow crops on portions of their land and has imposed strict limits on imports of food from abroad. As you can see, price floors often get the government deeply involved in production decisions, rather than leaving them to the market.

Price floors have certainly benefited farmers and helped them in times of need. But this market intervention has many critics—including most economists. They have argued that the government spends too much money buying surplus agricultural products, and the resulting higher prices distort the public's buying and eating habits—often to their nutritional detriment. For example, the General Accounting Office estimated that from 1986 to 2001, price supports for dairy products cost American consumers $10.4 billion in higher prices. And this does not include the cost of the health effects—such as calcium and protein deficiencies

among poor children—due to decreased milk consumption. The irony is that many of the farmers who benefit from price floors are wealthy individuals or large, powerful corporations that do not need the assistance. Economists argue that assistance to farmers would be more cost-effective if given directly to those truly in need, rather than supporting all farmers—rich and poor alike—with artificially high prices.

With price floors and price ceilings, the government tries to *prevent* the market price from reaching its equilibrium value. As you've seen, these efforts often backfire or have serious unintended consequences. But government can also intervene in a different way: It can try to influence the market outcome using *taxes* or *subsidies*—then stand out of the way and let the market help to achieve the desired outcome.

Manipulating the Market: Taxes

Taxes are imposed on markets to give the government revenue so it can provide public goods and services, to correct inequities in the distribution of income and wealth, or—our focus in this chapter—to change the price or quantity in a market.

A tax on a specific good or service is called an **excise tax**, which can be collected from either sellers or buyers. As you're about to see, the impact on the market and each of its participants is the same, regardless of which party is legally obligated to pay the tax, or from which party it is actually collected.

Excise tax A tax on a specific good or service.

An Excise Tax on Sellers

Let's explore the impact of an excise tax imposed on sellers with an example: Gasoline taxes. In the United States, the gasoline tax—collected from gasoline sellers—originated to fund the building and maintenance of the national highway system. But in recent years, many economists and others have wanted to increase this tax, for both geopolitical goals (to decrease U.S. imports of foreign oil) as well as environmental goals (to help fight pollution, traffic congestion, and climate change). Those who advocate a higher tax point out that gasoline taxes are much higher in Europe (approaching $4 per gallon or more in several European countries), while federal and state gasoline taxes in the U.S. average only 47 cents per gallon.

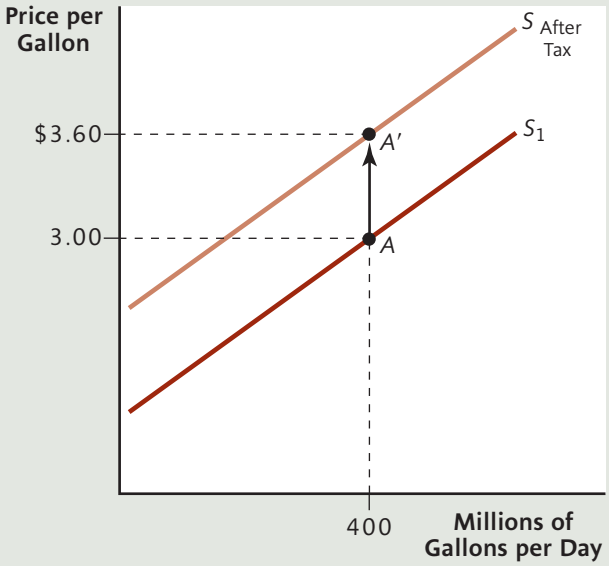
In our example, we'll assume that initially there is no gas tax at all, and then we'll impose a tax of 60 cents per gallon on *sellers*. How would such a tax affect the market for gasoline?

Suppose that before the tax is imposed, the supply curve for gasoline is S_1 in Figure 4. (Ignore the curve above it for now). Point A on this curve tells us that 400 million gallons will be supplied if the price is $3 per gallon. Let's rephrase this another way: *In order for the gasoline industry to supply 400 million gallons they must get $3 per gallon.*

What happens when our tax collector gets $0.60 for each gallon sold? What price must the industry charge now to supply the same 400 million gallons? The answer is $3.60, at point A'. Only by charging $0.60 more for each gallon could they continue to get the amount ($3) that makes it just worth their while to supply 400 million gallons. The same is true at any other quantity we might imagine: The price would have to be $0.60 more than before to get the industry to supply that same quantity.

FIGURE 4 A Tax on Sellers Shifts the Supply Curve Upward

After a $0.60 per gallon tax is imposed on sellers, the price at which any given quantity would be supplied is $0.60 greater than before, so the supply curve shifts upward. For example, before the tax, 400 million gallons would be supplied at $3 per gallon (point A); after the tax, to get that same quantity supplied requires a price of $3.60 (point A′).



So imposing a $0.60 tax on gas suppliers shifts the entire supply curve *upward* by $0.60, to $S_{\text{After Tax}}$.

A tax collected from sellers shifts the supply curve upward by the amount of the tax.

Now look at Figure 5, which shows the market for gasoline. Before the tax is imposed, with supply curve S_1 and demand curve D_1, the equilibrium is at point A, with price at $3 and quantity at 400 million. After the $0.60 tax is imposed and the supply curve shifts up to $S_{\text{After Tax}}$, the new equilibrium price is $3.40, with 300 million gallons sold.

Tax incidence The division of a tax payment between buyers and sellers, determined by comparing the new (after tax) and old (pretax) market equilibriums.

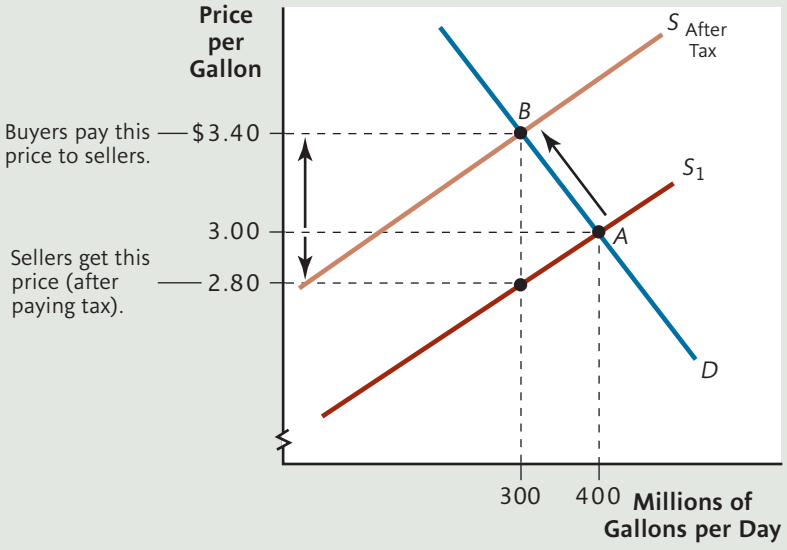
Who is paying this tax? Let's take a step back and think about it. Although the tax is collected from gas sellers, who really *pays*—that is, who sacrifices funds they would otherwise have if not for the tax—is an entirely different question. Economists call the distribution of this sacrifice the **tax incidence**.

© DAVID KARP/BLOOMBERG NEWS/LANDOV

In our example, buyers paid $3 for each gallon before the tax, and $3.40 after. So buyers are really paying $0.40 of this tax each time they buy a gallon of gas in the form of a higher price.

What about sellers? Before the tax, they got $3.00 per gallon. After the tax, they collect $3.40 from drivers but $0.60 of that goes to the government. If we want to know how much sellers get after taxes, we have to go back to the old supply curve S_1, which lies below the new supply curve by exactly $0.60. In effect, the old supply curve deducts the tax and shows us what the sellers really receive. When the sellers charge $3.40, the original supply curve S_1 shows us that they receive only $2.80. This is $0.20 less than they received before, so sellers end up paying $0.20 of the tax.

FIGURE 5 The Effect of an Excise Tax Imposed on Sellers



After a $0.60 excise tax is imposed on sellers, the market equilibrium moves from point A *to point* B, *with buyers paying sellers $3.40 per gallon. But sellers get only $3.40 − $0.60 after paying the tax. Thus, the tax causes buyers to pay $0.40 more per gallon, and sellers to get $0.20 less.*

In general,

The incidence of a tax that is collected from sellers generally falls on both sides of the market. Buyers pay more, and sellers receive less, for each unit sold.

An Excise Tax on Buyers

Suppose that, instead of collecting the $0.60 tax from the sellers, the tax was collected directly from buyers. Before the tax is imposed, the demand curve for gasoline is D_1 in Figure 6. Point *A* on this curve tells us that 400 million gallons will be demanded by buyers each day if the price they have to pay is $3.00. Or, rephrased, *in order for buyers to demand 400 million gallons, each gallon must cost them $3.00*. If the cost per gallon is any more than that, buyers will not buy all 400 million gallons.

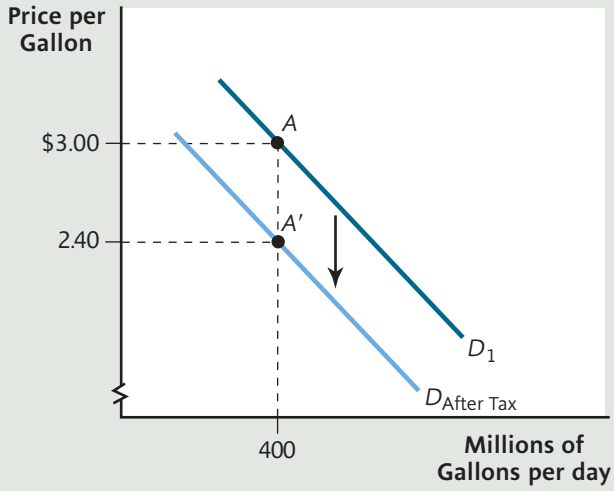
Now let's impose the $0.60 tax on buyers. (Imagine a government tax collector standing at the pump, requiring each buyer to hand over $0.60 for every gallon they buy.) What price will buyers now be willing to pay and still buy all 400 million gallons? The answer is $2.40, at point *A′*. We know this because only if they pay $2.40 will gasoline continue to cost them the $3.00 per gallon which makes it just worth their while to demand all 400 million gallons. The same is true at any other quantity we might imagine: The price would have to be $0.60 less than before to induce buyers to demand that same quantity. So imposing a $0.60 tax on buyers shifts the entire demand curve *downward* by $0.60, to $D_{\text{After Tax}}$.

A tax collected from buyers shifts the demand curve downward by the amount of the tax.

Figure 7 shows the impact on the market. Before the tax is imposed, with demand curve D_1 and supply curve *S*, the equilibrium is at point *A*, with price at $3.00 and

FIGURE 6 **A Tax on Buyers Shifts the Demand Curve Downward**

After a $0.60 per gallon tax is imposed on buyers, the price at which any given quantity would be demanded is $0.60 less than before, so the demand curve shifts downward. For example, before the tax, 400 million gallons would be demanded at $3 per gallon (point A); after the tax, that same quantity would be demanded at a price of $2.40 (point A′).

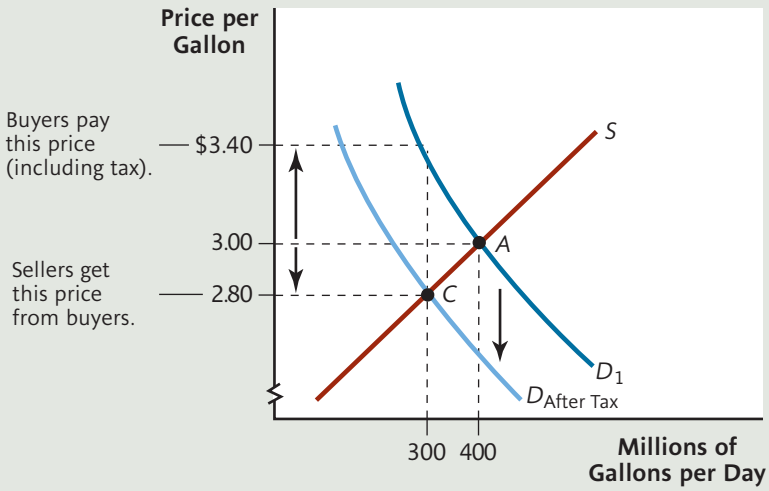


quantity at 400 million. After the $0.60 tax is imposed, and the demand curve shifts down to $D_{\text{After Tax}}$, the new equilibrium price is $2.80, with 300 million gallons sold. With the tax imposed on buyers this time, the supply curve is not affected.

What is the incidence of this tax? Let's see . . . Buyers paid $3.00 for each gallon of gas before the tax, and $2.80 after. But they also have to pay $0.60 to the government. If we want to know how much buyers pay *including* the tax, we have to go back to the old demand curve D_1, which lies above the new demand curve by exactly $60. As you can see, when buyers pay $2.80 to the gas station they pay a total of $3.40. This is $0.40 more than they paid in total before, so buyers end up paying $0.40 of the tax.

FIGURE 7 **The Effect of an Excise Tax Imposed on Buyers**

After a $0.60 excise tax is imposed on buyers, the market equilibrium moves from point A to point C, with buyers paying sellers $2.80 per gallon. But buyers pay a total of $2.80 + $0.60 = $3.40 per gallon when the tax is included. Thus, the tax causes buyers to pay $0.40 more, and sellers to get $0.20 less, just as when the tax is imposed on sellers.



What about sellers? Sellers received $3.00 for each gallon before the tax, and $2.80 after. So sellers are really paying $0.20 of this tax, in the form of a lower price.

The incidence of a tax that is collected from buyers falls on both sides of the market. Buyers pay more, and sellers receive less, for each unit sold.

Tax Incidence Versus Tax Collection

The numerical incidence of any tax will depend on the shapes of the supply and demand curves. But you may have noticed that the incidence in our example is the same whether the tax is collected from buyers or sellers. In both cases, buyers pay $0.40 of the tax per gallon and sellers pay $0.20. If you'll excuse the rhyme, this identical incidence is no coincidence.

The incidence of a tax (the distribution of the burden between buyers and sellers) is the same whether the tax is collected from buyers or sellers.

Why? Because the two methods of collecting taxes are not really different in any important economic sense. Whether the tax collector takes the $0.60 from the gas station owner when the gas is sold, or takes $0.60 from the driver when the gas is sold, one fact remains: Buyers will pay $0.60 more than the sellers receive. The market finds a new equilibrium reflecting this. In our example, this new equilibrium occurs where each gallon costs drivers $3.40 in total, and the gas suppliers receive $2.80 of this, because that is the only incidence at which quantity demanded and supplied are equal.[1]

Manipulating the Market: Subsidies

A **subsidy** is the opposite of a tax. Instead of the government demanding a payment *from* the buyer or seller, the government makes a payment *to* the seller or buyer. And whereas a tax raises the price to the buyer and decreases purchases of the product, a subsidy does the reverse: It lowers prices to buyers and *encourages* people to buy it. In the United States, federal, state, and local governments subsidize a variety of goods and services, including medical care for the poor and elderly, energy-saving equipment, smoking-cessation programs, and college education.

Subsidy A government payment to buyers or sellers on each unit purchased or sold.

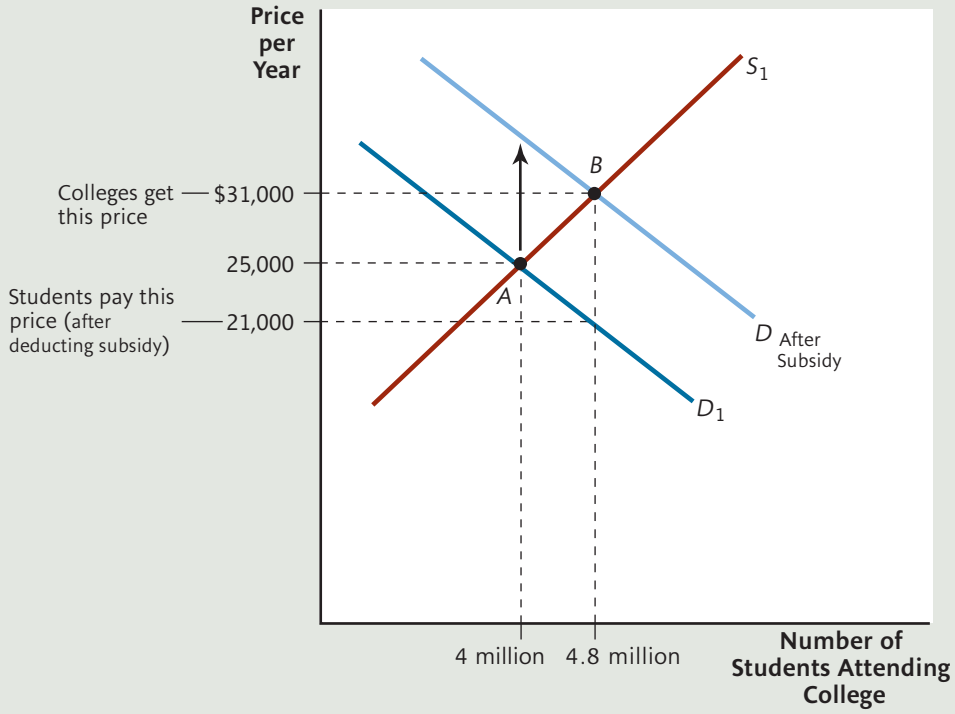
A Subsidy to Buyers

Let's explore the impact of a subsidy given to buyers with an example: Tuition assistance to college students.

Every year in the United States, federal and state governments provide more than $100 billion in subsidies—scholarships and other assistance—to help students pay the cost of a college education. The reasons for these policies are clear: We want to encourage more people to get college degrees. A college education gives substantial benefits to the degree-holders themselves, in the form of higher future incomes. And a more educated population creates benefits for society as a whole. There are also important equity considerations: Financial assistance increases the number of students from poor households, many of whom might otherwise have to forgo the benefits of college.

[1] In this chapter, we've considered only one type of burden caused by the tax: changes in price. But another burden of the tax is a decrease in quantity. In Chapter 14, you'll learn a more comprehensive method of measuring the burden of a tax that takes into account changes in both price and quantity.

FIGURE 8 A Subsidy for Students Attending College



After a $10,000 subsidy is given to college students, the market equilibrium moves from point A *to point* B, *with students paying colleges $31,000 per year. But students pay a total of $31,000 − $10,000 = $21,000 when their subsidy is deducted. Thus, the subsidy causes students to pay $4,000 less per year, and causes colleges to get $6,000 more per year.*

Figure 8 shows the market for attending college. Initially, without any government involvement, demand curve D_1 intersects supply curve S at point A. Four million students attend college each year, with each paying a total of $25,000 in tuition.

Now, let's introduce financial assistance: a subsidy of $10,000 per year for each student, and we'll assume here that the subsidy is paid out to buyers—college students—to help them pay tuition.

The subsidy shifts the demand curve *upward* by $10,000, to $D_{\text{After Subsidy}}$. Why? The old demand curve told us that if the price were $25,000, 4 million students would attend college. After a subsidy, the same 4 million students would choose to attend only if what they paid *from their own pockets* is still $25,000. A price of $35,000, with $10,000 kicked in from the government, would give us the same attendance of 4 million.

A subsidy paid to buyers shifts the demand curve upward by the amount of the subsidy.

However, the subsidy *changes* the market equilibrium from point A to point B. Now, 4.8 million students decide to attend college. But notice that the price is higher as well: Colleges are now charging $31,000 per year.

In general,

A subsidy paid to buyers benefits both sides of a market. Buyers pay less and sellers receive more for each unit sold.

Who benefits from the subsidy? Colleges benefit: They get more for each student who attends ($31,000 instead of $25,000). Students benefit as well: They pay $31,000 to the colleges, but the government pays $10,000 of that, so the cost to students has dropped to $21,000. However, notice that the $10,000 subsidy has *not* reduced the cost of college by a full $10,000. In our example, only $4,000 of the subsidy ends up as a direct benefit to the student, while the other $6,000 goes to the college.

A Subsidy to Sellers

As you learned earlier, the burden (incidence) of a tax is the same, regardless of whether it is collected from sellers or buyers. What about a subsidy? Is the benefit to each side of the market the same, regardless of which side the payment is given to? Indeed it is.

We leave it to you to do the analysis, but here's a hint: If the subsidy in our example had been paid to colleges (the sellers) instead of students (the buyers), the supply curve would shift *downward* by the amount of the subsidy. If you draw the graph, using the same initial supply and demand curves as in Figure 8 and the same $10,000 subsidy, you'll see that students will pay $21,000 (and not receive anything from the government), while colleges will collect $31,000 when we include the subsidy. What buyers end up paying, and what sellers end up receiving, is the same as in Figure 8.

In general,

The distribution of benefits from a subsidy is the same, regardless of whether the subsidy is paid to buyers or sellers.

Supply and Demand in Housing Markets

So far in this text, we've used supply and demand to analyze a variety of markets—for maple syrup, crude oil, higher education, and more. All of these markets have one thing in common: they are markets in which business firms sell currently produced goods or services.

But the supply and demand model is a versatile tool. With a bit of modification, we can use it to analyze almost any market in which something is traded at a price, including markets for labor, foreign currencies, stocks, bonds and more. Our only requirement is that there are many buyers and sellers, and each regards the market price (or a narrow range of prices) as given. In the remainder of this chapter, we'll use supply and demand to understand a type of market that has been at the center of recent economic events: the market for residential housing.

What's Different about Housing Markets

Housing markets differ in an important way from others we've considered so far. When people buy maple syrup, they buy newly produced bottles, not previously owned ones. But when people shop for a home, they generally consider newly constructed homes and previously-owned homes to be very close substitutes. After all, a house is a house. If properly maintained, it can last for decades or even centuries. Indeed, most of the homes that people own, and most that change hands each year, were originally built and sold long before.

This key difference—that housing markets are dominated by previously-owned homes—means we'll need to think about supply and demand in a somewhat different way. To understand this new approach, we first need to take a short detour.

A Detour: Stock and Flow Variables

Many economic variables fall into one of two categories: *stocks* or *flows*.

Stock variable A variable representing a quantity at a moment in time

Flow variable A variable representing a process that takes place over some time period.

A **stock variable** *measures a quantity in existence at a moment in time.*
A **flow variable** *measures a process that takes place over a period of time.*

To understand the difference, think of a bathtub being filled with water. At any given moment, there are a certain number of gallons actually *in* the tub. This volume of water is a *stock variable:* a quantity that exists at a moment in time (such as 15 gallons). But each minute, a certain volume of water flows *into* the tub. This rate of flow is a *flow variable:* a process that takes place over a *period* of time (such as, 2 gallons per minute).

In this book you've encountered both types of variables. In Chapter 3, for example, the quantity of maple syrup demanded or supplied was a flow: a certain number of bottles bought or sold *per month*. Similarly, household income is a flow: so many dollars earned *per month* or *per year*. But household *wealth*—the total value of what someone owns minus the total owed—is a stock variable. Wealth is measured in dollars at a particular point in time.

Now let's think about housing. New home *construction* and new home *purchases* are flow variables: so many homes are built or purchased per month or per year. By contrast, the number of homes that people own at a given time—say, 100 million homes on January 1, 2011—is a stock variable. Indeed, we often refer to this number of homes as the *housing stock*. Of course, as time passes and new homes are built, the housing stock rises. But at any given point in time, the housing stock is a fixed number of homes.

Which of these two concepts—the flow of new homes or the housing stock—should we use in our supply and demand model for housing? Let's see. If we use the flow of new homes, then equilibrium occurs when the number of *new* homes built per period is equal to the number of *new* homes people want to buy per period. But this leaves out the vast majority of homes that people can trade—homes built in *previous* periods and still standing. The most important changes in the housing market originate with these *previously*-built homes. So our concept of equilibrium should involve these homes as well.

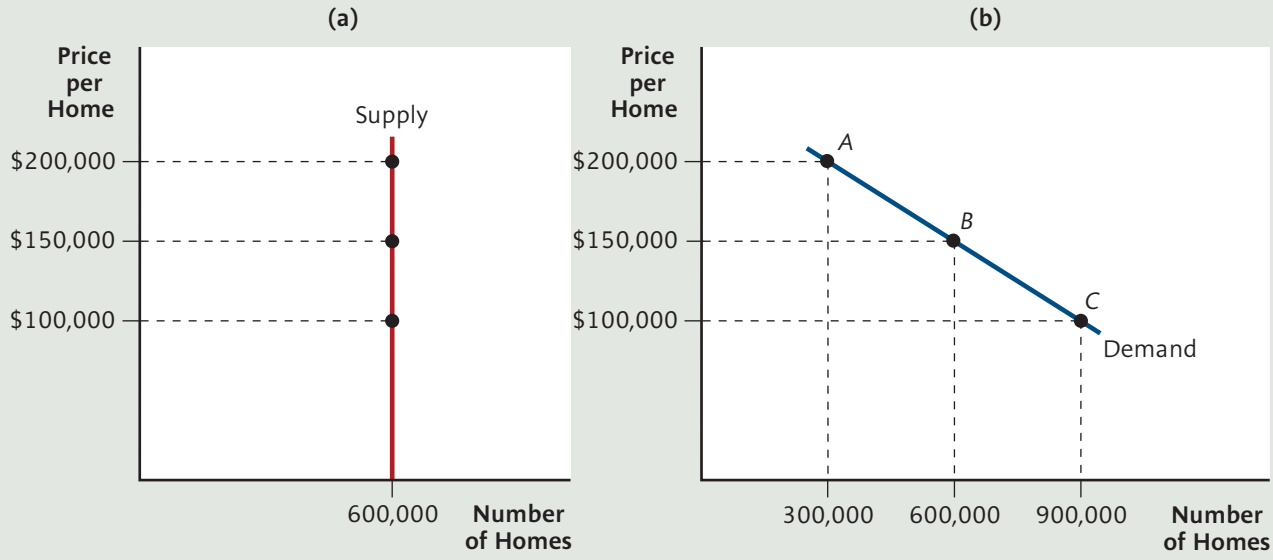
This is why it is best to think of the supply and demand for homes in terms of the housing *stock*. At a given point in time, the housing stock blends together *all* homes built up to that time—whether recently or long ago, and whether previously owned or never occupied. When we view the market in this way, we see that equilibrium occurs when the total number of homes people *want* to own is equal to the total number *available* for ownership—the housing stock.

To see how this works, let's take a closer look at supply and demand as stock variables, and illustrate them graphically.

Supply and Demand Curves in a Housing Market

The two panels in Figure 9 show supply and demand curves for a local housing market in a small city. Look first at the *supply* curve in panel (a). It represents the housing *stock*: the number of homes that exist in the area. The curve is vertical at 600,000 because, at this moment in time, the housing stock is fixed at 600,000.

FIGURE 9 **Supply and Demand Curves in a Housing Market**



The supply curve for housing tells us the number of homes that exist at a particular time, which does not depend on the price. The demand curve tells us how many homes people in the market want to own. The lower the price, the greater the quantity of homes demanded.

Whether the price is $200,000, $150,000 or $100,000, the number of homes in existence at this moment will still be 600,000.

The *demand* curve in panel (b) represents the *demand for the housing stock*. It tells us the number of homes that everyone in the area would *like to own* at each price, holding constant all the *other* variables that influence demand. To keep things simple, we'll imagine for most of our discussion that households or families can own a maximum of one home each. In that case, the demand curve tells us the number of families who want to be homeowners at each price. (Later in the chapter we'll discuss the role of owning multiple homes in the recent housing boom.)

Behind the Demand Curve: Ownership Costs

Notice that the demand curve in panel (b) slopes downward: As the purchase price of a home falls, more people want to own them. Why? Anyone deciding whether they want to become a homeowner, or continue being one, will compare the cost of owning with the cost of the next best alternative: renting. Because rent is paid monthly, it's natural to think about the costs of homeownership on a monthly basis as well.

There are several components to the monthly cost of owning a home, including maintenance, property taxes, and—the largest component—interest. As you are about to see, this monthly cost depends on the price of the home—not just for current buyers, but even for those who bought long ago.

Home Prices and Monthly Costs for Prospective Owners. Let's first consider monthly ownership costs for those thinking of *becoming* owners by buying a home. Suppose first that a buyer pays entirely with his or her own funds. Then those funds can no longer be invested elsewhere (such as, in a bank account), where they would

earn interest. Thus, buying a home means foregoing monthly interest—an implicit cost of owning. The greater the purchase price, the greater the monthly interest foregone.

Most prospective homeowners, however, do *not* pay the entire purchase price themselves. Instead, they pay for a small part of the purchase with their own funds—called the down payment—and *borrow* the rest, getting a housing loan called a **mortgage**. The mortgage is paid back monthly over many years. For most of that time, the monthly payments will consist largely of interest charges. The higher the purchase price of the home, the bigger the mortgage loan, and the greater the monthly mortgage payment.

Mortgage A loan given to a homebuyer for part of the purchase price of the home.

So far, we've seen that for a *prospective* homeowner, the monthly cost of owning the home will vary directly with the price of the home. This remains true whether the buyer is planning to take out a mortgage, or buy the home without borrowing.

But what about someone who *already* owns a home? Would a later change in the home's price affect monthly costs for that owner?

Home Prices and Monthly Costs for Current Owners. You might think that, once someone has purchased a home, monthly ownership costs should depend only on the price paid earlier, at the time of purchase, and be immune from any later price change. But in fact, any change in the home's price—even after it was purchased—will change the owner's monthly costs.

To see why, remember that anyone who owns a home could always choose to sell it. The owner could then pay back any amount that might be remaining on the mortgage and also get some cash back (since the selling price is usually greater than the amount owed on the mortgage). That cash could earn interest. Continued ownership, therefore, means continued foregone interest.

Of course, the higher the selling price for the house, the more cash the owner *would* get from selling, and the more interest the owner sacrifices by *not* selling. So once again, a rise in home prices increases the monthly cost of ownership.

Let's summarize what we've found about ownership costs:

Both current and prospective homeowners face a monthly cost of ownership. This monthly cost rises when home prices rise, and falls when home prices fall.

Now you can understand why the demand curve in Figure 9(b) slopes downward. When housing prices fall and nothing else changes, the monthly cost of owning a home declines as well. With lower monthly costs, more people will prefer to own rather than rent, so the quantity of homes demanded increases.

dangerous curves

Misinterpreting the Supply and Demand Curves for Homes. It's very easy to slip into thinking that the supply and demand curves for housing have the "flow" interpretation of the previous chapter, instead of the proper "stock" interpretation here. So remember: the supply curve does *not* represent the number of homes people want to sell over a period of time. Rather, it represents the number of homes that *exist* at a point in time. Similarly, the demand curve does *not* tell us how many homes people would like to buy over a period of time. Rather, it tells us how many homes people want to *own* at a point in time.

Shifts versus Movements Along the Demand Curve

As with any demand curve, when we change the price and move along the curve, we hold constant all other influences on demand. For example, in Figure 9(b), if the price of a home falls from $150,000 to $100,000, *ceteris paribus*, we move along the demand curve from point *B* to point *C*. The number of families who want to own rises from 600,000 to 900,000. The opposite happens when home prices rise from $150,000 to $200,000: Fewer people want to own, and we move along the demand curve from point *B* to point *A*.

But as we make these movements, we are holding constant the monthly cost of renting a home, interest rates in the economy, tastes for homeownership, average income, population—anything that might affect number of people who want to own homes in the market *other* than the price of a home. If any of these other factors change, the demand curve will shift. We'll look at examples of these shifts a bit later.

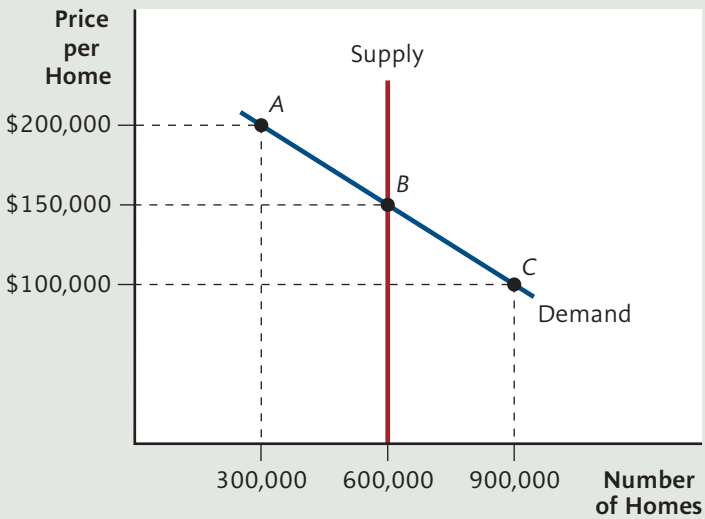
Housing Market Equilibrium

Figure 10 combines the supply and demand curves from Figure 9. The equilibrium, as always, occurs where the two curves intersect, at point *B*, with the price of a house at $150,000. But because this is a new type of market, it's worth spending a little time understanding *why* equilibrium occurs at this price.

Suppose the price of homes in this area was $100,000, which is less than the equilibrium price. The demand curve at point *C* tells us that 900,000 people would want to own homes at this price. But the supply curve tells us that only 600,000 homes are available—that is, only 600,000 homes *can* be owned. So with a price of $100,000, there is an excess demand for homes. What will happen?

People who want homes, but don't yet have them, will try to buy them from the current owners, bidding up prices. Because the housing stock is constant (at least for now), the rise in price will *not* change the quantity of homes available. But it *will* move us along the demand curve (upward and leftward from point *C*), as higher

FIGURE 10 **Equilibrium in a Housing Market**



The equilibrium in this market is at point B, where the price of homes is $150,000. If the price were higher—say $200,000–the number of homes people want to own (300,000 at point A) would be less than the number in existence and currently owned (600,000). Owners would try to sell, and the price would fall until all 600,000 homes were demanded. If the price were lower than the equilibrium price—say $100,000—the number of homes people want to own (900,000 at point C) would be less than the number in existence and currently owned (600,000). People would try to buy homes, and the price would rise until only 600,000 were demanded.

home prices (and higher monthly ownership cost) reduce the number of people who want to own. The price will continue rising until the number of people who *want* to own a home is equal to the number of homes that *can* be owned: the 600,000 that exist. This occurs when the price reaches $150,000, at point *B*.

What if the price started *higher* than $150,000—say, $200,000? Then only 300,000 people would want to own homes, at point *A*. But remember: 600,000 homes exist, and at any time, every one of them must be owned by *someone*. So at a price of $200,000, half of the current homeowners would prefer *not* to own. What will happen?

Those who prefer not to own will try to sell, causing home prices to fall. As the price drops, and we move rightward along the demand curve, more people decide they want to own. The price continues dropping until it reaches $150,000, at point *B*, where people are content to own all 600,000 homes in existence.

The equilibrium price in a housing market is the price at which the quantity of homes demanded (the number that people want to own) and quantity supplied (the housing stock) are equal.

What Happens When Things Change

So far, in Figure 10, we've identified the equilibrium at a particular point in time. We did so by assuming that the housing stock was fixed at 600,000, and all influences on demand other than the price of homes were held constant.

But *over time,* in most housing markets, both the supply and demand curves will shift rightward. The supply curve shifts rightward as the housing stock rises (new homes are built). And the demand curve shifts rightward for a variety of reasons, including population growth and rising incomes. As a result, the market equilibrium will move rightward over time as well. But what happens to home *prices* depends on the *relative* shifts in the supply and demand curves. Let's look at three possibilities.

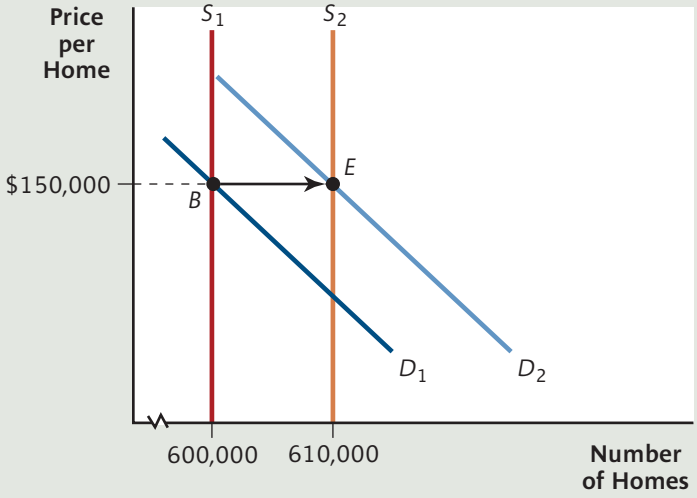
Equal Changes in Supply and Demand: A Stable Housing Market

Figure 11 illustrates a stable housing market, in which increases in the housing stock just keep pace with increases in housing demand over time. Let's start with the initial situation, at point *B*, with a housing stock of 600,000 homes and a price of $150,000. Over the next year, population and income growth shifts the demand curve rightward to D_2. At each price, the demand for homes is 10,000 more than before. New construction increases the housing stock by 10,000 as well, shifting the supply curve to S_2. The equilibrium moves to point *E*, with a new, higher housing stock of 610,000, and an unchanged equilibrium price of $150,000.

When the housing stock grows at the same rate as housing demand, housing prices remain unchanged.

We should note that Figure 11 is not entirely realistic. In most housing markets, construction costs for labor and raw materials tend to rise over time. In order to cover these rising costs and continue increasing the housing stock, average home prices must rise over time, at least modestly. You'll learn more about these types of

FIGURE 11 A Stable Housing Market



When the supply of homes increases at the same rate as demand for them, the equilibrium price remains unchanged. In the figure, the rightward shift in the supply curve (from S_1 to S_2) is equal to the rightward shift in the demand curve (from D_1 to D_2). Equilibrium moves from point B to point E, but the price remains at $150,000.

market adjustments when you study perfectly competitive markets in more detail, in microeconomics. In the figure, we've ignored rising construction costs.

But in some cases, we observe home prices rising *much* faster than can be explained by rising construction costs alone. When home prices rise especially rapidly, we know that increases in demand must be outpacing increases in supply. Let's consider two possible ways this could happen.

Restrictions on New Building: Rapidly Rising Prices

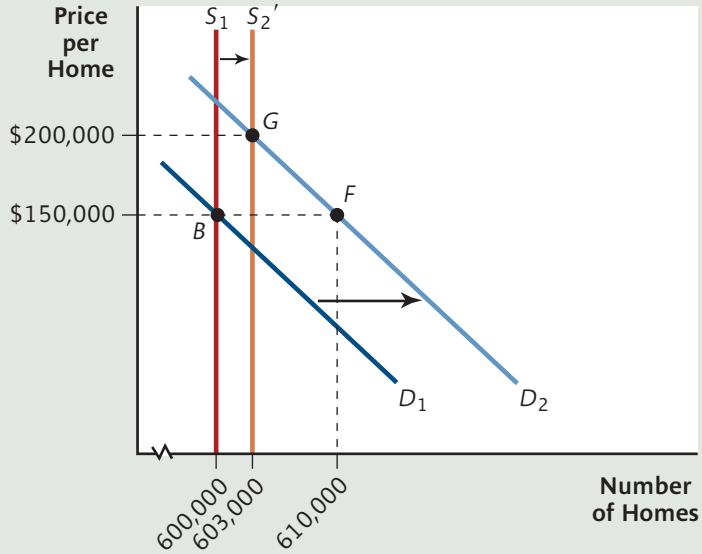
In some housing markets, local restrictions on new building can prevent the housing stock from keeping up with ongoing increases in demand. This is illustrated in Figure 12. As in our previous example, the initial market equilibrium is at point *B*, with price equal to $150,000. And once again, the demand curve shifts rightward over the year by 10,000 units, due to population and income growth. But now, we assume that restrictions allow construction of only 3,000 new homes, so the supply curve shifts by less than in the previous figure: to S_2'. After the demand shift, people want to own 610,000 homes (at point *F*) at a price of $150,000, but only 603,000 exist. The excess demand of 7,000 homes drives up home prices, and we move leftward along the demand curve, until only 603,000 homes are demanded. At the new equilibrium (point *G*), the price is $200,000.

When restrictions on new building prevent the housing stock from growing as fast as demand, housing prices rise.

Restrictions on new building explain why housing prices have risen rapidly for decades in cities like New York and in many areas of California. The *demand* for housing in these areas increases in most years, but various restrictions on new building or the limited supply of coastal land prevent the housing stock from keeping up.

FIGURE 12 A Housing Market with Restricted Supply Growth

When supply is restricted, and cannot increase as fast as demand, housing prices rise. In the figure, the rightward shift in the supply curve (from S_1 to S_2') is less than the rightward shift in the demand curve (from D_1 to D_2). Equilibrium moves from point B to point G, and the price rises from $150,000 to $200,000



Faster Demand Growth: Rapidly Rising Prices

Several factors could cause the demand curve for housing to begin shifting rightward more rapidly than in the past: population shifts (a sudden influx of new residents), rapid income growth (because of a booming industry in the area), or a change in expectations about future prices. Let's take a closer look at this last factor: expectations.

So far, we've considered home ownership as an alternative to renting. That is, whether you own or rent, you get valuable services—a roof over our head and a place to watch TV. And we've discussed these alternatives in terms of their relative monthly costs.

But a house is more than just a place to live. It is also an example of an **asset**—something of value that someone owns. An asset can be sold in the future at whatever price prevails in the market at that time. If the asset is sold at a higher price than the purchase price, the seller enjoys a **capital gain**, If the asset is sold for *less* than the purchase price, the seller suffers a **capital loss**.

Capital gain The gain to the owner of an asset when it is sold for a price higher than its price when originally purchased.

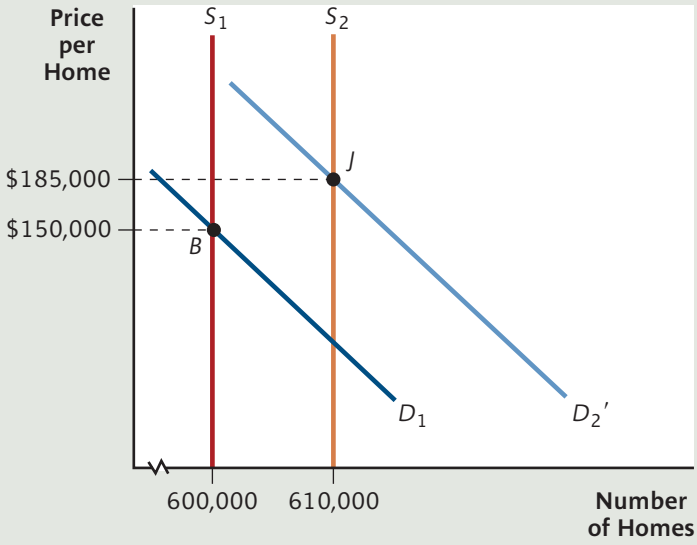
Capital loss The loss to the owner of an asset when it is sold for a price lower than its price when originally purchased.

Anticipated capital gains are one of the reasons that people hold assets, including homes. In fact, capital gains and losses are especially important in the housing market, for one reason: A home is one of the most *leveraged* financial investments that most people ever make.

The appendix to this chapter discusses leverage in more detail, and some special features of leverage in the housing market. But the takeaway from the appendix is this: Leverage *magnifies* the impact of a price change on the rate of return you will get from an asset. When home prices rise, your rate of return from investing in a home can be *several times* the percentage growth in housing prices. This makes the demand for housing particularly sensitive to changing expectations about future home prices.

Let's suppose that people begin to think housing prices will increase more rapidly over the next few years than they've risen in the past. With housing seen as an even more profitable investment than before, more people want to be homeowners

FIGURE 13 Accelerating Demand Growth



When demand begins to increase faster than previously, increases in supply usually lag behind. In the figure, the rightward shift in the supply curve (from S_1 to S_2) is less than the rightward shift in the demand curve (from D_1 to D_2'). Equilibrium moves from point B to point J, with the price rising from $150,000 to $185,000.

at any given price for homes. The demand curve will shift rightward by more than it otherwise would.

Figure 13 illustrates the result. The demand curve—instead of shifting to D_2 as in Figure 12, now shifts further, to D_2'. But notice that the supply curve continues to shift only to S_2 (the housing stock rises by only 10,000 units), just as in Figure 12. Why doesn't the housing stock keep up with the suddenly higher demand?

Because it takes *time* for new construction to be planned and completed. The change in the housing stock over the *current* year is based on how many construction projects were initiated in the *previous* year. And the number of projects started in the previous year was based on prices *then*—before the surge in demand. With the demand curve shifting to D_2', and the supply curve to S_2, the equilibrium moves to point *J*, with a new equilibrium price of $185,000.

Note that with higher housing prices construction firms will have an incentive to increase building. Unless restrictions prevent them from doing so, the housing stock will rise at a faster rate—in *future* years. Eventually, the housing stock can catch up to the higher demand, but that will happen much later. In the meantime, the main impact of the rapid increase in demand is higher home prices.

In summary:

When the demand for housing begins rising faster than previously, the housing stock typically lags behind, and housing prices rise.

Changes in expectations—and rapidly shifting demand curves—were a major cause of the housing bubble from the late 1990s through 2006, and the housing bust that immediately followed, as you are about to see.

Using the Theory

THE HOUSING BOOM AND BUST OF 1997–2008

© ROBIN NELSON/PHOTOEDIT

Figure 14 shows an index measure of inflation-adjusted U.S. housing prices. (The inflation adjustment removes the effects of general inflation from the change in home prices, making comparisons over time more meaningful.) The index begins with a value of 100 in 1975. In any other year, its value tells us the percentage by which the median inflation-adjusted home price exceeded the median price in 1975. For example, in 1991, the index's value was 107.12, telling us that the median home in 1991, after adjusting for inflation, cost about 7% more than the median home in 1975.

The most glaring feature of this graph is the startling price increase from 1997 to 2006. During that period, the housing price index almost doubled. (Remember—this is the rise in the index *after* we remove any increase due to general inflation.) Something special must have been happening with housing.

The rapid rise in housing prices—especially after 2001—has been described as a housing "bubble." The term bubble suggests something that is destined to burst. And Figure 14 shows this is just what happened: In mid-2006, U.S. housing prices began falling. And as the months went by, they continued to fall—faster. A similar pattern of bubble and burst occurred in many other countries. In some of them, such as Britain and Ireland, the boom and bust were even more extreme than in the United States.

What caused housing prices to behave this way? A complete answer requires some additional tools and concepts that you will learn later in microeconomics and macroeconomics. But at the center of this economic storm were the familiar forces of supply and demand.

The Housing Boom: Rapidly Rising Demand

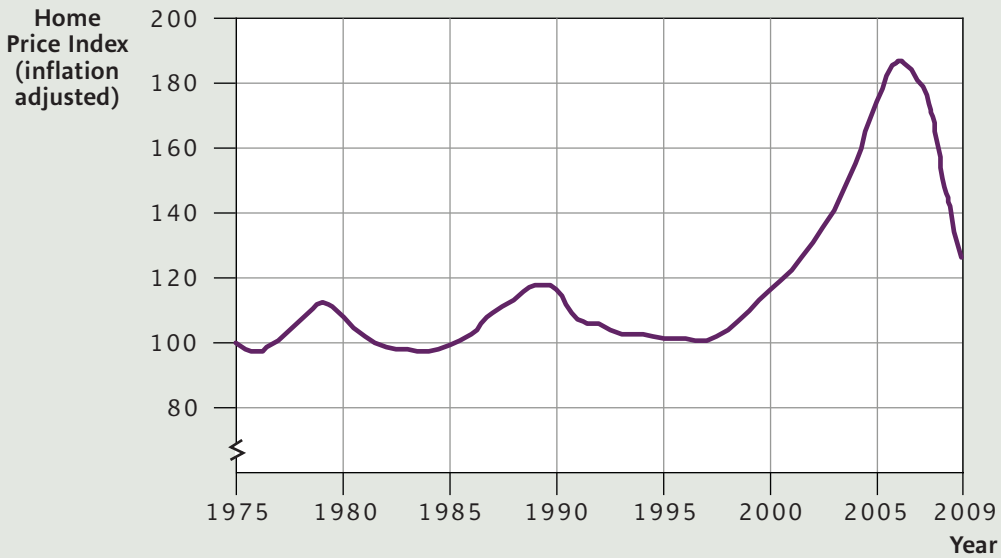
As you've learned, rapid increases in home prices can occur when demand begins increasing faster than supply. From 1997 to 2006, several factors caused the demand for housing to grow more rapidly each year. While supply increased as well, it lagged behind—more and more each year—because of the ever-more-rapid increases in demand. The result was a surge in housing prices.

What caused demand to increase so rapidly during this period?

Economic Growth

In the United States and many other countries, the 1990s were a period of prosperity and rising incomes. Higher incomes increase the demand for most goods, including housing. In addition, after years of high employment, people felt increasingly confident about the future, and more willing to take on the long-term financial obligations associated with homeownership.

FIGURE 14 **Index of Home Prices, Adjusted for Inflation**



After adjusting for price changes from general inflation, the housing boom began in 1997, and home prices increased ever more rapidly until 2006. That marked the beginning of the housing bust, with prices dropping dramatically for the next few years.

Source: S&P/Case-Shiller Home Price Index

Interest Rates

Beginning in 2001, interest rates on many types of loans trended downward, including the interest rate on mortgage loans. The reasons for this general decline in interest rates are somewhat controversial. The policy of the U.S. Federal Reserve played a key role. But global financial forces may have contributed as well. Indeed, the decline in interest rates was observed in many other countries, not just the United States.

In macroeconomics, you'll learn more about how interest rates in the overall economy are determined, and what causes them to change. In this section, we'll just note that a drop in interest rates reduces the monthly cost of homeownership and increases the number of homes demanded *at any given home price*. Thus, the general drop in interest rates contributed to the rightward shift in the demand for housing.

Government Policy

In the United States (more so than most other countries), owning a home has been viewed as a desirable way to promote financial security for individuals and responsible citizenship for local communities. For this reason, the government has long encouraged homeownership in two major ways.

First, it allows homeowners to deduct mortgage interest payments from their taxable income. The government in effect says, "If you shift some of your spending from other things to mortgage payments, we will give you some of your tax dollars back." This amounts to a subsidy: a payment from the government to the borrower for each dollar of interest paid on a mortgage loan, lowering monthly homeownership costs.

Second, government agencies[2] have increased the funds available for mortgage lending by purchasing mortgages from banks and other financial institutions, giving

[2] The two main agencies are the Federal National Mortgage Association, informally known as "Fannie Mae" and the Federal Home Loan Mortgage Corporation, known as "Freddie Mac"

them fresh cash to lend out again for another mortgage. This resulting increase in funding for mortgages has helped to keep mortgage interest rates—and monthly homeownership costs—low.

These policies have caused the demand for housing to be greater than it would otherwise be. But they had been in place for decades, so their mere existence cannot explain the housing boom. However, at the start of the boom, both policies were stepped up significantly. Government agencies expanded their purchases of mortgage loans, helping to push mortgage interest rates even lower. And in 1997, the government added another tax policy for homeowners: It raised the "capital gains exclusion" on home sales from $125,000 to $500,000, and made it easier to apply the exclusion to a second home. This meant that the first $500,000 of capital gains from selling a home would be entirely tax free. Thus, owning one or more homes with a mortgage—already an attractive, highly leveraged investment—became even more attractive.

Financial Innovations

Two types of financial innovation—both of which had existed prior to the boom—became more prevalent as the boom developed. The first type of innovation involved more-attractive terms for borrowers. The adjustable rate mortgage (ARM), for example, offered a very low initial interest rate (and low monthly payments). The interest rate and monthly payments would leap upward later, usually after two years. But during the initial low-payment period, ARMs lowered monthly homeownership costs, and increased the demand for housing.

A second type of innovation made mortgage *lending* more attractive. Traditionally, a mortgage lender such as a bank would hold onto a mortgage and collect the monthly payments from the homeowner for the life of the loan, typically 30 years. But a technique called *securitization* pooled many mortgages together, and then divided them into smaller financial assets called *mortgage backed securities*. Each mortgage backed security promised its holder monthly payments that came not from *one* homeowner's monthly mortgage payments, but from *hundreds* of homeowners monthly payments.

Though mortgage backed securities had been around for decades, they became more popular during the housing boom because of *other* financial innovations. These included new ways for lenders to quantify the risks of individual mortgage loans (called "credit scoring"), which enabled them to quantify the overall risk of any mortgage backed security. Also, new ways of combining and re-dividing the securities themselves were developed, which seemed to reduce their risks further.

Financial institutions in the U.S. and around the world—hungry for new, low-risk opportunities to lend—purchased hundreds of billions of dollars of these new securities each year. In this way, an entirely new group of global investors—and an entirely new source of funds—became available for mortgage lending. Mortgage interest rates—and monthly costs for new home buyers—fell further. The demand curve for homes shifted further rightward.

Lending Standards

Banks and other financial institutions that make mortgage loans take a risk: If housing prices decline, and an owner owes more on the mortgage than the home is worth, the owner might *default* (stop making payments).

When a family defaults, it ultimately moves out of the home—either by walking away or because they are forced out by *foreclosure*. This process is costly, and when the home is resold, it is usually a distress sale, at a bargain price. All in all, 50% or

more of the remaining value of the loan can be lost in a default. These losses are then transmitted to the lender or anyone currently holding the mortgage backed security that contains that mortgage as part of its larger pool.

Traditionally, lenders have guarded against homeowner defaults in two ways (1) lending only to those whose incomes and credit histories suggest a small probability of default; and (2) requiring the borrower to make a sizeable down payment—traditionally 20% of the home's value. The down payment gives the borrower something at stake—and a reason to continue making payments—even if the price of the home declines modestly.

However, as the boom proceeded, with huge amount of funds flowing into the mortgage market and lenders competing with each other to find borrowers, lending standards deteriorated. Lenders began making increasing numbers of so-called *subprime loans:* loans to borrowers who previously would not have qualified due to low or unstable incomes or bad credit histories. (Credit scoring, discussed earlier, suggested that the risks of such loans could be quantified and therefore managed.) In 2006 alone, more than $600 billion in subprime mortgage loans were made—about one-fifth of all mortgage lending.

Down payments began to shrink as well. From 2005 to 2007, more than half of the subprime mortgage loans in the U.S. actually had no down payment at all. In Britain, some large banks offered mortgages for 125% of the value of the home—in effect, a negative down payment!

The decline in lending standards contributed to the housing boom by opening up the prospect of homeownership to millions of people who would not otherwise have qualified. This caused a further rightward shift in the demand curve for housing.

Speculation

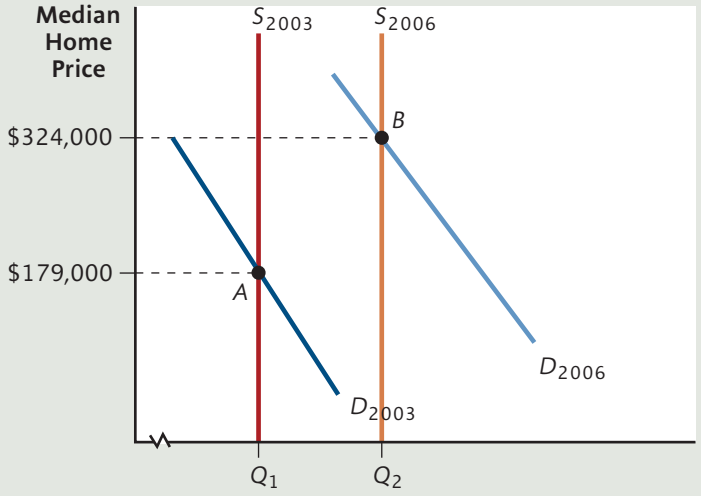
Once housing prices had increased rapidly for several years, and people began expecting them to *keep* rising at those rates, speculation took over. Remember that—even with a 20% down payment—housing is a highly leveraged investment (see the appendix). If the housing market gives you a way to turn a $40,000 investment into $100,000 or more in a few years, why *not* buy a home? And why just one? Why not buy two, three, or as many as you can obtain mortgages for? You can always rent out the ones you aren't using. Your ARMs will have low monthly payments for a couple of years, and when the interest rates reset to higher, unaffordable levels, you can always sell your house for a capital gain.

Once this kind of thinking takes over a market, it feeds on itself. People want to speculate in housing because they expect the price to keep rising, and the price keeps rising because more and more people are speculating and shifting the demand for housing ever rightward.

An Example: The Boom in Las Vegas

Figure 15 shows an example of how all of these forces drove up housing prices in one particular city—Las Vegas—where a full-forced bubble began in mid-2003 (a bit later than in some other cities). Initially, with demand curve D_{2003} and supply curve S_{2003}, the market was in equilibrium at Point *A*, with the median home worth $179,000. Then, for all of the reasons we've discussed, the demand curve for homes in Las Vegas shifted rightward each year, reaching D_{2006} in mid-2006. As housing prices rose, new construction picked up as well, increasing the housing stock and shifting the supply curve rightward, ultimately to S_{2006}. But, as discussed in this chapter, when demand increases rapidly, the housing stock often lags behind. That

FIGURE 15 The Housing Boom in Las Vegas



is what happened in Las Vegas. As shown in the figure, demand increased substantially more than supply, so home prices soared—rising to $324,000 in June 2006.

Unfortunately, what happened in Vegas didn't stay in Vegas. The same bubbles developed in many areas of the country—especially in towns and cities in California, Arizona, and Florida.

The Housing Bust: A Sudden Drop in Demand

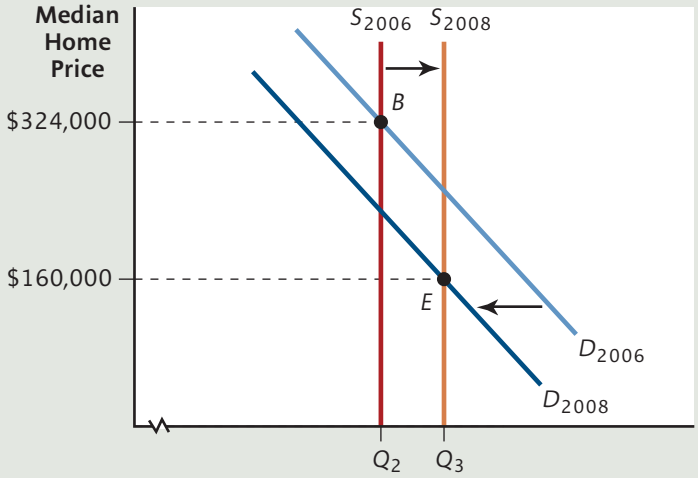
Every bubble bursts at some point. If nothing else, there are natural limits to its growth. If home prices had continued to rise at such a rapid pace, eventually new buyers would not be able to make monthly mortgage payments, even with low interest rates. The speculation would ultimately slow, and then reverse direction when people realized that prices were no longer rising.

But a bubble can burst before it reaches a natural limit. And in U.S. housing markets, problems began to occur in mid-2006, largely due to two simultaneous events: (1) oil and gasoline prices spiked, so many new homeowners were struggling to make ends meet; and (2) interest rates on a large group of adjustable rate mortgages reset to higher levels. Suddenly, people noticed a disturbing rise in defaults—especially on subprime mortgages with no down payments. Around the world, lenders to the U.S. mortgage market began to take a closer look at market conditions, and they did not like what they saw: more ARMs resetting to higher levels over the next several years, and more defaults down the road. Everyone suddenly took notice of how high home prices had risen, and how far they could fall.

For the first time, it seemed, investors began to ask questions about the statistical analysis used by financial institutions to measure the risk of ARMs and subprime mortgages, and the risks of their associated mortgage backed securities. And it turned out that every financial institution had made the same assumption: That housing prices would continue rising or, at worst, fall only modestly. No one seemed to consider what would happen if housing fell more dramatically.

Now, with the prospect of higher default rates, and the possibility of falling home prices, the former flood of funding for new mortgages turned into a drought. Interest rates on new mortgages—to the extent they were available at all—rose. The

FIGURE 16 The Housing Bust in Las Vegas



demand curve for housing shifted leftward. And the fears of mortgage lenders became self-fulfilling: Housing prices fell.

Moreover, once housing prices began to fall, speculative fever worked in reverse. By 2007, what had been a near-certain, highly-leveraged gain for home-buyers turned into the prospect of a highly-leveraged loss. Anyone buying a home with a traditional down payment risked losing the entire investment in a few years or less. And many of those who already owned homes suddenly wished they didn't. The demand curve for housing shifted further leftward, and housing prices fell even more rapidly.

An Example: The Bust in Las Vegas

Figure 16 illustrates how these events affected the housing market in Las Vegas. Initially, at the peak of the bubble in mid-2006, with demand curve D_{2006} and supply curve S_{2006}, the market was at point *B*, with the average price of a home at $324,000. But then, the demand curve shifted leftward, and by late 2008, it reached a location like that of D_{2008}. The supply curve, of course, did *not* shift leftward. Houses generally last a long time, and the homes built during the bubble had now become permanent additions to the housing stock. Moreover, the housing stock *continued* to grow—even as home prices dropped. (Remember: Construction projects are started long in advance. Also, housing prices—while dropping—still remained high by historical standards for many months after the bubble burst, providing continued incentives to build.)

In the figure, the new equilibrium price is depicted at $160,000 for 2008. But in reality, the average home price in Las Vegas at the end of 2008, while still heading downward, had fallen only to $178,000. Why does the figure show an equilibrium price lower than the actual price at the time? The reason is: While housing prices tend to *rise* rapidly to a new, higher equilibrium, they tend to *fall* to a lower equilibrium very slowly. Thus, the market price at the end of 2008 had probably not yet reached its equilibrium. One reason that prices fall so slowly is psychological: sellers who bought at higher prices resist selling at a lower price and acknowledging that they've lost money on their most significant investment. Instead, they may hang on and refuse to sell for months or years, waiting in vain for a higher offer that never

comes. An unfortunate consequence is that a housing bust can last for years, continuing to affect the economy long after the bubble initially bursts.

By the end of 2008, the average price of a home in the U.S. had fallen almost 30% from the peak of the bubble. (Las Vegas home prices were down about 45%). Few economists at the time believed that home prices had reached their ultimate lows. But in early 2009, new government programs to increase the demand for housing were put in place. The goal was to raise the equilibrium price, so that *actual* home prices would have less to fall before reaching that equilibrium. By mid-2009, there were some encouraging signs these policies might be having an effect.

SUMMARY

The model of supply and demand is a powerful tool for understanding all sorts of economic events. For example, governments often intervene in markets through *price ceilings* or *price floors,* designed to prevent the market from reaching equilibrium. Economic analysis shows that these policies are often ineffective in achieving their goal of helping one side of the market, and often create additional problems.

Governments also intervene in markets with *taxes* and *subsidies*. Taxes increase the equilibrium price and decrease the equilibrium quantity, while subsidies do the opposite: decreasing price and increasing quantity. Taxes and subsidies have the same effect on the market regardless of whether the tax is imposed on (or the subsidy given to) buyers or sellers.

Supply and demand can also be used to understand markets other than those for currently produced goods and services. One important example is the market for residential housing, which is usefully analyzed as *stock variables* (quantities that exist at a moment in time), rather than *flow variables* (processes that take place over a period of time). In the housing market, the supply curve tells us the quantity of homes in existence, and the demand curve tells us the number of homes that the population would like to own. The demand curve slopes downward because housing entails an ongoing ownership cost, with interest cost (paid or foregone) one of its major components. The lower the price of a home, the lower the monthly ownership cost, and the more attractive owning is compared to renting.

In a stable housing market, the housing stock keeps pace with demand growth, so prices remain stable. But restrictions on new building, or a sudden acceleration of demand, can cause housing prices to soar. Because people usually buy homes with *mortgage loans,* housing is a highly *leveraged* financial investment: the value of the home is a multiple of the funds invested. As a result, the demand for housing is especially sensitive to changes in expected prices. This played a role in the most recent housing boom, and the housing bust that followed.

PROBLEM SET

Answers to even-numbered Questions and Problems can be found on the text website at www.cengage.com/economics/hall

1. The market for rice has the following supply and demand schedules:

P (per ton)	Q^D (tons)	Q^S (tons)
$10	100	0
$20	80	30
$30	60	40
$40	50	50
$50	40	60

To support rice producers, the government imposes a price floor of $50 per ton.
   a. What quantity will be traded in the market? Why?
   b. What steps might the government have to take to enforce the price floor?

2. In Figure 2, a price ceiling for maple syrup caused a shortage, which led to a black market price ($4) higher than the initial equilibrium price ($3). Suppose that the price ceiling remains in place for years. Over time, some maple syrup firms go out of business. With fewer firms, the supply curve in the figure shifts leftward by 10,000 bottles per month. After the shift in the supply curve:
   a. What is the shortage caused by the $2 price ceiling?
   b. If all of the maple syrup is once again purchased for sale on the black market, how will the black market price be greater, less than, or the same as that in Figure 2? Explain briefly.

3. In the chapter, you learned that one way the government enforces agricultural price floors is to buy up the excess supply itself. If the government wanted to follow a similar kind of policy to enforce a price *ceiling* (such as rent control), and thereby prevent black-market-type activity, what would it have to do? Is this a sensible solution for enforcing rent control? Briefly, why or why not?

4. In Figure 5, prove that the incidence of a $0.60 tax imposed on sellers could not be split equally between buyers and sellers, given the supply and demand curves as drawn. [Hint: What price would gasoline sellers have to charge after the tax for an even split? What would happen in the market if sellers charged this price?]

5. Figure 8 shows the impact of a $10,000 subsidy on the market for college education when the subsidy is paid to college students. Starting with the same initial supply and demand curves, show what happens when the same $10,000 subsidy per student is paid to the *colleges* they attend. Suggestion: Trace the relevant curves from the figure on your own sheet of paper. (Hint: If a subsidy is paid directly to the colleges, which curve will shift? In what direction?)

6. State whether each of the following is a stock variable or a flow variable, and explain your answer briefly.
   a. Total farm acreage in the U.S.
   b. Total spending on food in China
   c. The total value of U.S. imports from Europe
   d. Worldwide iPhone sales
   e. The total number of parking spaces in Los Angeles
   f. The total value of human capital in India
   g. Investment in new human capital in India

7. Suppose you buy a home for $200,000 with a $20,000 down payment and finance the rest with a home mortgage.
   a. Suppose that if you default on your mortgage loan, you lose the home, but nothing else. By what percentage would housing prices have to fall to create an economic incentive for you to default on the loan? Explain briefly.
   b. Suppose that if you default on your mortgage loan, you not only lose the home, but also $10,000 in moving and relocating expenses. By what percentage would housing prices have to fall now to create an economic incentive for default?

8. Every year, the housing market in Monotone, Arizona, has the same experience: The demand curve for housing shifts rightward by 500 homes, 500 new homes are built, and the price of the average home doesn't change. Using supply and demand diagrams, illustrate how each of the following new events, *ceteris paribus,* would affect the price of homes in Monotone during the current year, and state whether the price rises or falls.
   a. Because of special tax breaks offered to Monotone home builders, 800 new housing units are built during the current year.
   b. Because of events in the overall economy, interest rates fall.
   c. The Monotone city council passes a new zoning law that prevents *any* new home construction in Monotone during the year.
   d. Because of the new zoning law, and the resulting change in home prices, people begin to think that homes in Monotone are a better investment than they had thought before.
   e. 500 new homes are built in Monotone during the year, but that same year, an earthquake destroys 2,000 preexisting homes. As a result of the earthquake, 3,000 homeowners decide they no longer want to live or own homes in Monotone.

9. Every year in Houseville, California, builders construct 2,000 new homes—the most the city council will allow them to build. And every year, the demand curve for housing shifts rightward by 2,000 homes as well. Using supply and demand diagrams, illustrate how each of the following new events, *ceteris paribus,* would affect the price of homes in Houseville during the current year, and state whether home prices would rise or fall.
   a. Houseville has just won an award for the most livable city in the United States. The publicity causes the demand curve for housing to shift rightward by 5,000 this year.
   b. Houseville's city council relaxes its restrictions, allowing the housing stock to rise by 3,000 during the year.
   c. An earthquake destroys 1,000 homes in Houseville. There is no affect on the demand for housing, and the city council continues to allow only 2,000 new homes to be built during the year.
   d. The events in a., b., and c. all happen at the same time.

10. [Requires appendix] Suppose you buy a home for $400,000 with a $100,000 down payment and finance the rest with a home mortgage.
   a. Immediately after purchasing your home, before any change in price, what is the value of your *equity* in the home?
   b. Immediately after purchasing your home, before any change in price, what is your simple leverage ratio on your investment in the home?
   c. Now suppose that over the next three years, the price of your home has increased to $500,000. Assuming you have not borrowed any additional funds using the home as collateral, but you still

owe the entire mortgage amount, what is the new value of your equity in the home? Your new simple leverage ratio?

d. Evaluate the following statement: "An increase in the value of a home, with no additional borrowing, increases the degree of leverage on the investment in the home." True or false? Explain.

11. [Requires appendix] Suppose, as in the previous problem, you buy a home for $400,000 with a down payment of $100,000 and take out a mortgage for the remainder. Over the next three years, the price of the home rises to $500,000. However, during those three years, you also borrow $50,000 in *additional* funds using the home as collateral (called a "home equity loan"). Assume that, at the end of the three years, you still owe the $50,000 as well as your original mortgage.
    a. What is your equity in the home at the end of the three years?
    b. How many times are you leveraged on your invest ment in the home at the end of the three years?
    c. By what percentage could your home's price fall (after it reaches $500,000) before your equity in the home is wiped out?

More Challenging

12. [Requires Appendix] Suppose, as in the previous problem, you buy a home for $400,000 with a down payment of $100,000 and take out a mortgage for the remainder. Over the next three years, the price of the home rises to $500,000. However, during those three years, you borrow the *maximum* amount you can borrow without changing the value of your home equity. Assume that, at the end of the three years, you still owe all that you have borrowed, including your original mortgage.
    a. How much do you borrow (beyond the mortgage) over the three years?
    b. What is your simple leverage ratio at the end of the three years?
    c. By what percentage could your home's price fall (after it reaches $500,000) before your equity in the home is wiped out?

13. [Requires appendix] Could any combination of home price, mortgage, or further borrowing on a home result in a simple leverage ratio of 1/2? If yes, provide an example. If no, briefly explain why.

Understanding Leverage

This appendix discusses the concept of *leverage:* what it means, how it can be measured, and its implications for owning an asset. Our focus here is on the housing market. But leverage can be applied to many other markets, as you will see later in this text.

Let's start by exploring how the housing market would operate *without* leverage. Imagine that you had to pay for a home in full, using only your own funds. In that case, if you have $100,000 available for buying a home, you could buy a home worth $100,000, and no more.

Suppose you bought a home for $100,000, and over the year, housing prices rose 10%. The home would then be worth $110,000. If you then sold it (and if we ignore selling costs and maintenance) your $100,000 investment would have turned into $110,000—a capital gain of $10,000. This gives you a 10% rate of return on your financial investment of $100,000. We'll also note that, if the price of the home *fell* by 10%, down to $90,000, you would have a capital *loss* of $10,000–again, 10% of your financial investment. Notice that, when you use only your own funds, your rate of return on your investment (+10% or −10%) is the same as the rate of change for the home's price.

This example is *not* how most people buy a home. In the United States and many other countries, if you have $100,000 to invest in a home, you will use it for just *part* of the purchase—called the *down payment*—and borrow the remainder. This allows you to buy a home worth substantially more than $100,000. Using borrowed money to buy a home is an example of a *leveraged* financial investment.

To see how this works in practice, let's once again assume you have $100,000 of your own funds available. But now, you'll use it as a down payment, equal to 20% of the home's purchase price. You'll buy a home for $500,000, and take out a mortgage loan for the amount not covered by your down payment: $400,000.

Panel (a) of Figure A.1 illustrates how this works: You use your own $100,000, plus $400,000 from the mortgage lender, to purchase a home worth $500,000. In return, you sign a mortgage contract, promising to pay back $400,000 over time.

Now let's suppose, once again, that housing prices rise by 10% over the year. Because you own a $500,000 home, its price has risen to $550,000. Panel (b) shows what happens when you sell this home. You sell the house for $550,000, pay back what you owe the bank ($400,000), and the mortgage contract is paid in full. You now have $150,000 left over. Remember: Your original investment was $100,000, and you now have $150,000, for a capital gain of $50,000. Thus, a 10% rise in housing prices has given you a 50% rate of return on your investment.

Of course, if the price *fell* by 10%, your home would be worth only $450,000. If you sold it at that price, and paid off the $400,000 loan, you'd be left with only $50,000 of the $100,000 you started with. In that case, you'd have *lost* 50% of your initial investment.

As you can see, when you borrow to buy a home, the potential capital gains and losses on your original investment are magnified. This magnification of gains and losses through borrowing is called leverage.

Measuring Leverage

For many purposes, it's useful to calculate the *degree* of leverage associated with an investment, such as a home purchase. There are various ways of measuring leverage, but all of them rely on the concept of *equity*:

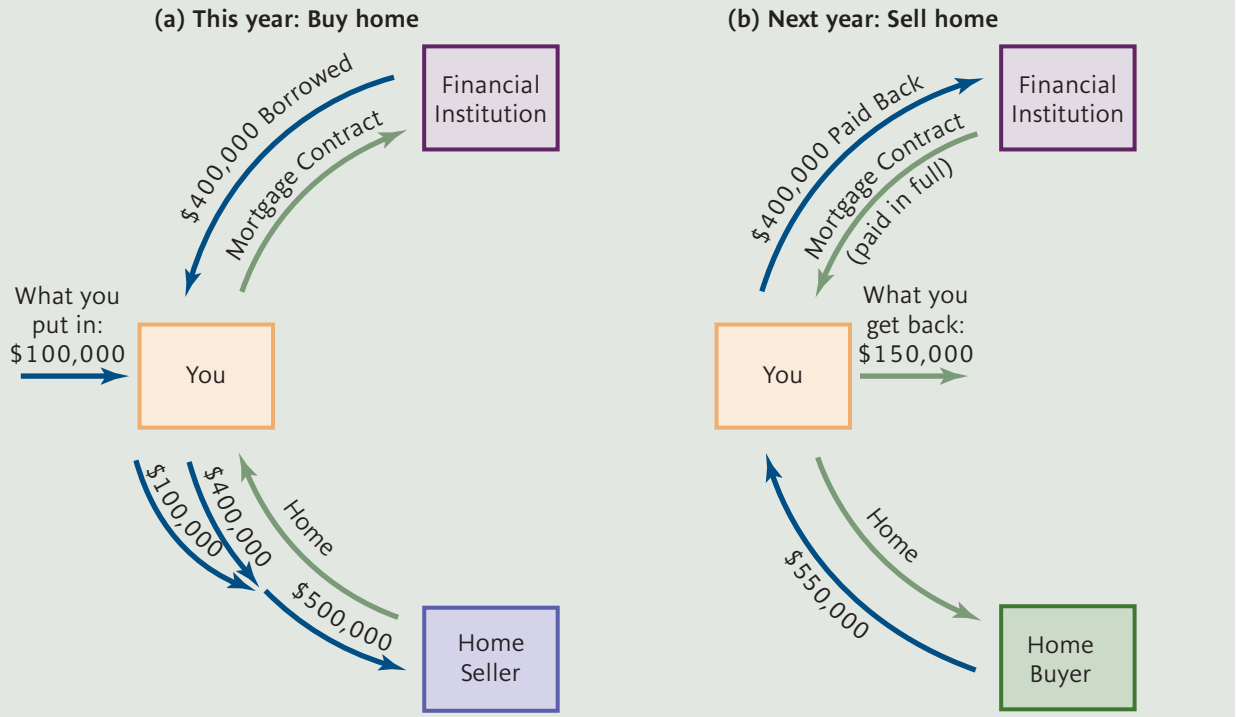
> An owner's equity *in an asset is the difference between the asset's value and any unpaid debts on the asset (that is, debts for which the asset was used as collateral):*

$$\text{Equity in Asset} = \text{Value of Asset} - \text{Debt Associated with Asset}$$

Notice that an owner's equity depends on the asset's *value*. For assets owned by individuals or families, we use the *current market value* of the asset—the price at which it could be sold.[3] So, for example, a homeowner's equity is the price at which the home could be sold, minus any

[3] Owner's equity for a business firm is defined very much like equity for a household: The total value of its assets minus what it owes to others. But for firms, asset values are typically based on historical prices paid, rather than current market value, until the asset is actually sold or officially revalued. So, for example, a building purchased long ago and still owned by a firm will be valued at its original purchase price, with some adjustment for depreciation.

FIGURE A.1 Leveraged Buying and Selling



debts for which the home was used as collateral. The equity represents the part of the asset's value that truly belongs to its owner. It is what the owner would get if the asset were sold, after paying back the associated debt.

The concept of equity leads directly to one way of measuring leverage, which we'll call the *simple leverage ratio:*

The ***simple leverage ratio*** *is the ratio of an asset's value to the owner's equity in the asset:*

$$\text{Simple Leverage Ratio} = \frac{\text{Value of Asset}}{\text{Equity in Asset}}$$

Now, let's apply these concepts to home ownership. Suppose, as in our first example, you use only your own funds to buy a $100,000 home, and never use the home as collateral for a loan. Using the definitions above, your equity in the home is $100,000 − $0 = $100,000. Your simple leverage ratio is $100,000/$100,000 = 1. A leverage ratio of 1 means no leverage at all: There is no borrowing to magnify capital gains and losses.

Now let's calculate the leverage ratio in our second example, in which you make a $100,000 down payment on a $500,000 home, and borrow the remaining $400,000. Your equity in the home is $500,000 − $400,000 = $100,000. And your simple leverage ratio is $500,000/$100,000 = 5. In words, we'd say you are "leveraged five times."

Leverage and Rate of Return

The simple leverage ratio serves as a "rate-of-return multiplier." That is, we can multiply the rate of change in a home's price by the leverage ratio to get the rate of return on the (leveraged) investment. For example, we found earlier that, when you buy a $500,000 home and are leverage 5 times, a 10% rise in housing prices gives you a 50% rate of return on your investment—five times the percentage increase in the home's price.

As you can see, when asset prices rise, leverage can increase your rate of return dramatically. But when asset prices fall, leverage increases the chance of wiping out your entire investment. With no leverage, your home's price would have to fall by 100% before your owner's equity would disappear. If you are leveraged 5 times, a drop of 20% eliminates your equity. And if you're leveraged 20 times, it takes only a 5% drop in prices to wipe out your entire investment.

One last word. In this appendix, we've been applying the concept of leverage, and the simple leverage ratio, to a single asset. But leverage can also refer to the *combined* assets of a household or business firm, or even to an entire sector of the economy. If you are studying macroeconomics, you'll see in a later chapter that high degrees of financial sector leverage in the U.S. and several other countries played a crucial role in the recent global recession.

CHAPTER 5

Elasticity

Imagine that you are the mayor of one of America's large cities. Every day, the headlines blare about local problems—poverty, crime in the streets, the sorry state of public education, roads and bridges that are falling apart, traffic congestion—and you, as mayor, are held accountable for all of them. Of course, you could help alleviate these problems, if only you had more money to spend on them. You've already raised taxes as much as you can, so where to get the money?

One day, an aide bounds into your office. "I've got the perfect solution," he says, beaming. "We raise mass transit fares." He shows you a sheet of paper on which he's done the calculation: Each year, city residents take 500 million trips on public transportation. If fares are raised by 50 cents, the transit system will take in an additional $250 million—enough to make a dent in some of the city's problems.

You stroke your chin and think about it. So many issues to balance: fairness, practicality, the political impact. But then another thought occurs to you: Your aide has made a serious mistake! Public transportation—like virtually everything else that people buy—obeys the law of demand. A rise in price—with no other change—will cause a decrease in quantity demanded. If you raise fares, each *trip* will bring in more revenue, but there will be *fewer trips* taken. Mass transit revenue might rise or it might fall. How can you determine which will happen?

To answer that question, you need to know how *sensitive* mass transit ridership is to a change in price.

In this chapter, you will learn about *elasticities:* measures of the sensitivity of one variable to another. As you'll see, economists use a variety of different types of elasticities to make predictions and to recommend policy changes.

Price Elasticity of Demand

At the most general level, an *elasticity* measures the sensitivity of one market variable to another. One of the most important elasticities is the *price elasticity of demand,* which measures the sensitivity of quantity demanded to the price of the good itself. But how should we measure this sensitivity?

One obvious candidate is the *slope* or *steepness* of the demand curve. After all, for any given rise in price, the flatter the demand curve, the greater will be the decrease in quantity demanded along the curve. So it seems that the flatter the demand curve (the smaller the absolute value of its slope), the greater the sensitivity of quantity demanded to price. But that reasoning is only partially correct, and it can get us into trouble.

Problems with Slope

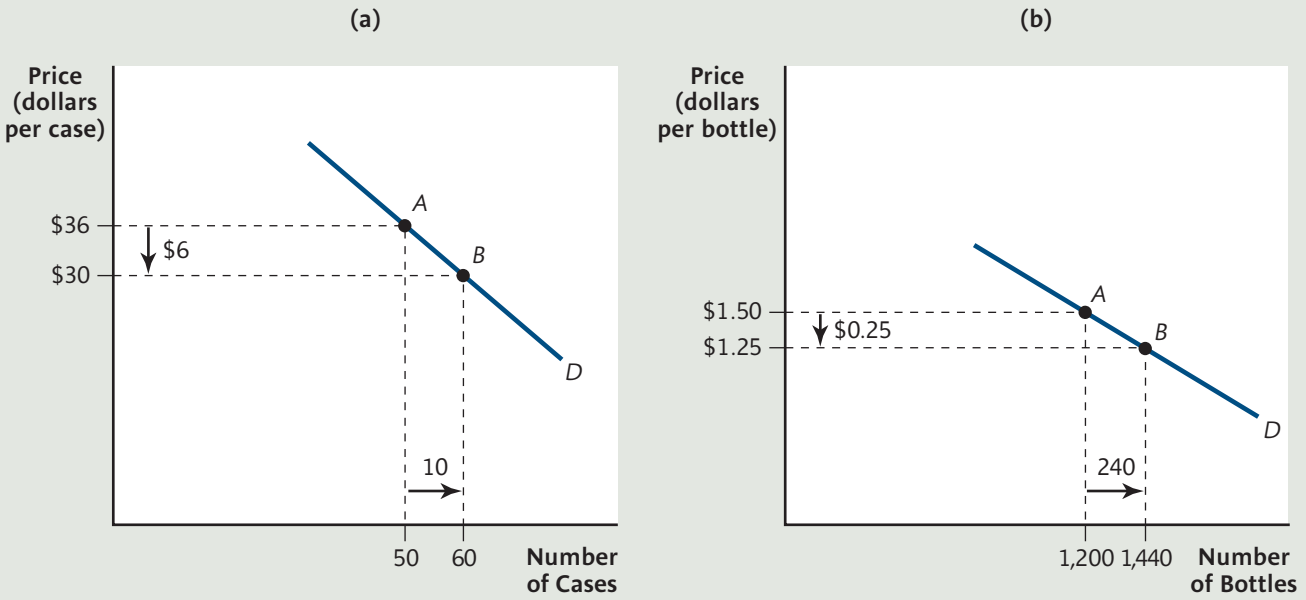
There are two problems with using slope to measure the price sensitivity of demand. First, the slope of a demand curve depends on the arbitrary units of measurement that we happen to choose.

For example, 20-ounce bottles of water are sold both individually and by the case (24 bottles per case). Suppose we use "cases," and find that the demand curve in a market is like that shown in the left panel of Figure 1. In that market, a $6 drop in price per case causes people to buy 10 more cases per day (moving us from point *A* to point *B*). Because we graph price on the vertical axis and number of cases on the horizontal, the *slope* of the demand curve would be $\Delta P/\Delta Q^D = -\$6/10 = -0.6$.

But now, for the same market, let's change our unit of measurement from "number of cases" to "number of bottles." $6 less per *case* translates to 25 cents less per *bottle*. And a 10-case increase in quantity demanded translates to 240 bottles. The right panel of Figure 1 shows the demand curve for "bottles" (with the scale of the axes adjusted so we can see larger quantities and smaller prices more easily). As we move from point *A* to *B*, the slope of the demand curve is now $\Delta P/\Delta Q^D = -0.25/240 = -.001$. Buyers respond the same way in both examples, yet—due only to an arbitrary change in units—the demand curves have very different slopes. Clearly, we can't rely on the slope as our measure of price sensitivity.

A second problem is that the slope of the demand curve doesn't tell us anything about the *significance* of a change in price or quantity—whether it is a relatively small or a

FIGURE 1 **Same Buying Behavior, Different Slopes**



In each panel, the movement from A *to* B *represents the same buying behavior in the market for bottled water. In panel (a), the unit of measurement is* cases. *When price drops by $6 per case, quantity demanded rises by 10 cases, so the slope of the demand curve is* $-\$6/10 = -0.6$. *In panel (b), the unit of measurement is* bottles. *The same price decrease ($6 per case) translates to 25 cents per bottle, and the same quantity increase (10 cases) translates to 240 bottles. Using bottles, the slope is* $-\$0.25/240 = -.001$. *Although the demand behavior is the same, the slopes are different. This is one reason why slope is a poor measure of the price sensitivity of demand.*

relatively large change. A price drop of $0.05, for example, is a tiny, hardly noticeable change for a good with a current price of $500. But it's a relatively huge change if the current price is $0.08. Our measure of price sensitivity should take this into account.

The Elasticity Approach

The elasticity approach solves both of these problems by comparing the *percentage change* in quantity demanded with the *percentage change* in price.

More specifically:

*The **price elasticity of demand*** (E_D) *for a good is the percentage change in quantity demanded divided by the percentage change in price:*

$$E_D = \frac{\%\text{ Change in Quantity Demanded}}{\%\text{ Change in Price}}$$

Price elasticity of demand The sensitivity of quantity demanded to price; the percentage change in quantity demanded caused by a 1 percent change in price.

For example, if the price of newspapers falls by 2 percent, and this causes the quantity demanded to rise by 6 percent, then $E_D = 6\%/2\% = 3.0$. We would say "the price elasticity of demand for newspapers is 3.0."

Of course, when price *falls* by 2 percent, that's a change of *negative* 2 percent, while quantity demanded changes by +6 percent. So technically speaking, elasticity should be viewed as a negative number. In this book, we'll follow a common convention of dropping the minus sign. That way, when we compare elasticities and say that one is larger, we'll be comparing absolute values.

In our example, elasticity has the value 3.0. But what, exactly, does that number mean? Here is a straightforward way to interpret the number:

The price elasticity of demand (E_D) *tells us the percentage change in quantity demanded for each 1 percent change in price.*

In our example, with $E_D = 3.0$, each 1 percent drop in price causes quantity demanded to rise by 3 percent.

Given this interpretation, it's clear that an elasticity value of 3.0 implies greater price sensitivity than an elasticity value of 2.0, or one of 0.7. More generally,

The greater the elasticity value, the more sensitive quantity demanded is to price.

Calculating Price Elasticity of Demand

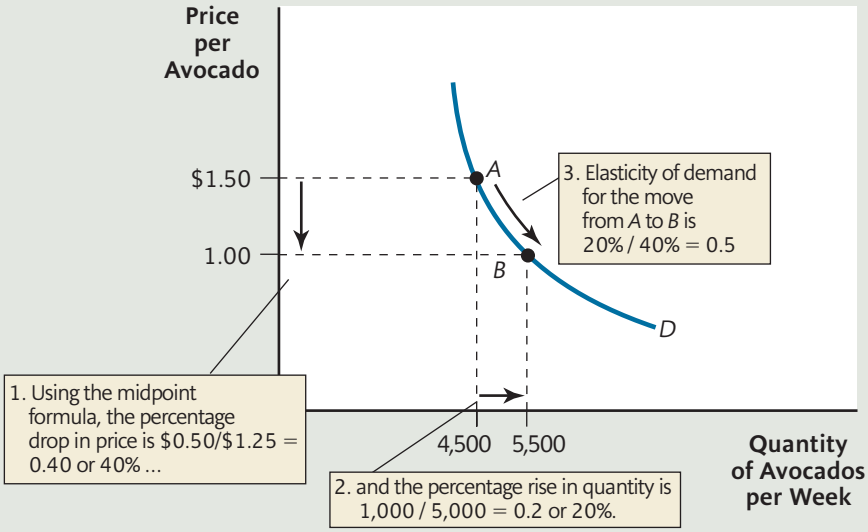
When we calculate price elasticity of demand, we imagine that *only price* is changing, while we hold constant all other influences on quantity demanded, such as buyers' incomes, the prices of other goods, and so on. Thus, we measure elasticity for a movement *along* an unchanging demand curve.

Figure 2, for example, shows a hypothetical demand curve in the market for avocados in a city. Suppose we want to measure elasticity along this demand curve between points *A* and *B*. As our first step, we'll calculate the percentage change in price.

dangerous curves

Mistakes in Observing Elasticities It's tempting to calculate an elasticity from simple observation: looking at what actually happened to buyers' purchases after some price changed. But this often leads to serious errors. Elasticity of demand tells us the effect a price change would have on quantity demanded *if* all other influences on demand remain unchanged. That is, it measures price sensitivity *along* a single demand curve. But in the real world, these other influences may change in the months or years after a price change causing the demand curve to shift. Economists and statisticians have developed tools to isolate the effects of price changes when other things are affecting demand at the same time. If you study economics further, and take a course in econometrics, you'll learn some of these techniques.

FIGURE 2 Using the Midpoint Formula for Elasticity



Let's suppose we move from point *A* to point *B*. Price falls by $0.50. Since our starting price at point *A* was $1.50, this would be a 33 percent drop in price.

But wait . . . suppose we go in the reverse direction, from point *B* to *A*. Now our starting price would be $1.00, so the $0.50 price hike would be a 50 percent rise. As you can see, the percentage change in price (33 or 50 percent) depends on the direction we are moving. And the same will be true of quantity. Therefore, our elasticity value will also depend on which direction we move.

This presents us with a problem. Ideally, we'd like our measure of price sensitivity to be the same whether we go from *A* to *B* or from *B* to *A*, since each is simply the mirror image of the other. To accomplish this goal, elasticity calculations often use a special convention to get percentage changes: Instead of dividing the change in a variable by its *starting* value, we divide the change by the *average* of its starting and ending values. This is often called the "midpoint formula," because we are dividing the change by the midpoint between the old and new values.

When determining elasticities, we calculate the percentage change in a variable using the midpoint formula: the change in the variable divided by the average of the old and new values.

For example, in Figure 2, between points *A* and *B* the average of the old and new price is ($1.50 + $1.00) / 2 = $1.25. Using this average price as our base, the percentage change in price is $0.50 / $1.25 = 0.40 or 40 percent. With the midpoint formula, the percentage change in price is the same whether we move from *A* to *B*, or from *B* to *A*.

More generally, when price changes from any value P_0 to any other value P_1, we define the percentage change in price as

$$\% \text{ Change in Price} = \frac{(P_1 - P_0)}{\left[\frac{(P_1 + P_0)}{2}\right]}$$

The term in the numerator is the change in price; the term in the denominator is the average of the two prices.

The percentage change in quantity demanded is calculated in a similar way. When quantity demanded changes from Q_0 to Q_1, the percentage change is calculated as

$$\% \text{ Change in Quantity Demanded} = \frac{(Q_1 - Q_0)}{\left[\frac{(Q_1 + Q_0)}{2}\right]}$$

Once again, we are using the average of the initial and the new quantity demanded as our base quantity.

The midpoint formula is an approximation to the actual percentage change in a variable, but it has the advantage of giving us consistent elasticity values when we reverse directions. We will use the midpoint formula only when *calculating elasticity values from data on prices and quantities*. For all other purposes, we calculate percentage changes in the normal way, using the starting value as the base.

An Example

Let's calculate the price elasticity of demand for avocados along a part of the demand curve in Figure 2. As price falls from \$1.50 to \$1.00, quantity demanded rises from 4,500 to 5,500. Using the midpoint formula (and dropping negative signs):

$$\% \text{ Change in Quantity Demanded} = \frac{5{,}500 - 4{,}500}{\left[\frac{(5{,}500 + 4{,}500)}{2}\right]} = \frac{1{,}000}{5{,}000} = 0.20 \text{ or } 20\%.$$

$$\% \text{ Change in Price} = \frac{[\$1.00 - \$1.50]}{\left[\frac{(\$1.00 + \$1.50)}{2}\right]} = \frac{\$0.50}{\$1.25} = 0.40 \text{ or } 40\%.$$

Finally, we use these numbers to calculate the price elasticity of demand:

$$E_D = \frac{\% \text{ Change in Quantity Demanded}}{\% \text{ Change in Price}} = \frac{20\%}{40\%} = 0.5.$$

Or, in simple English, a 1 percent change in price causes a ½ percent change in quantity demanded.

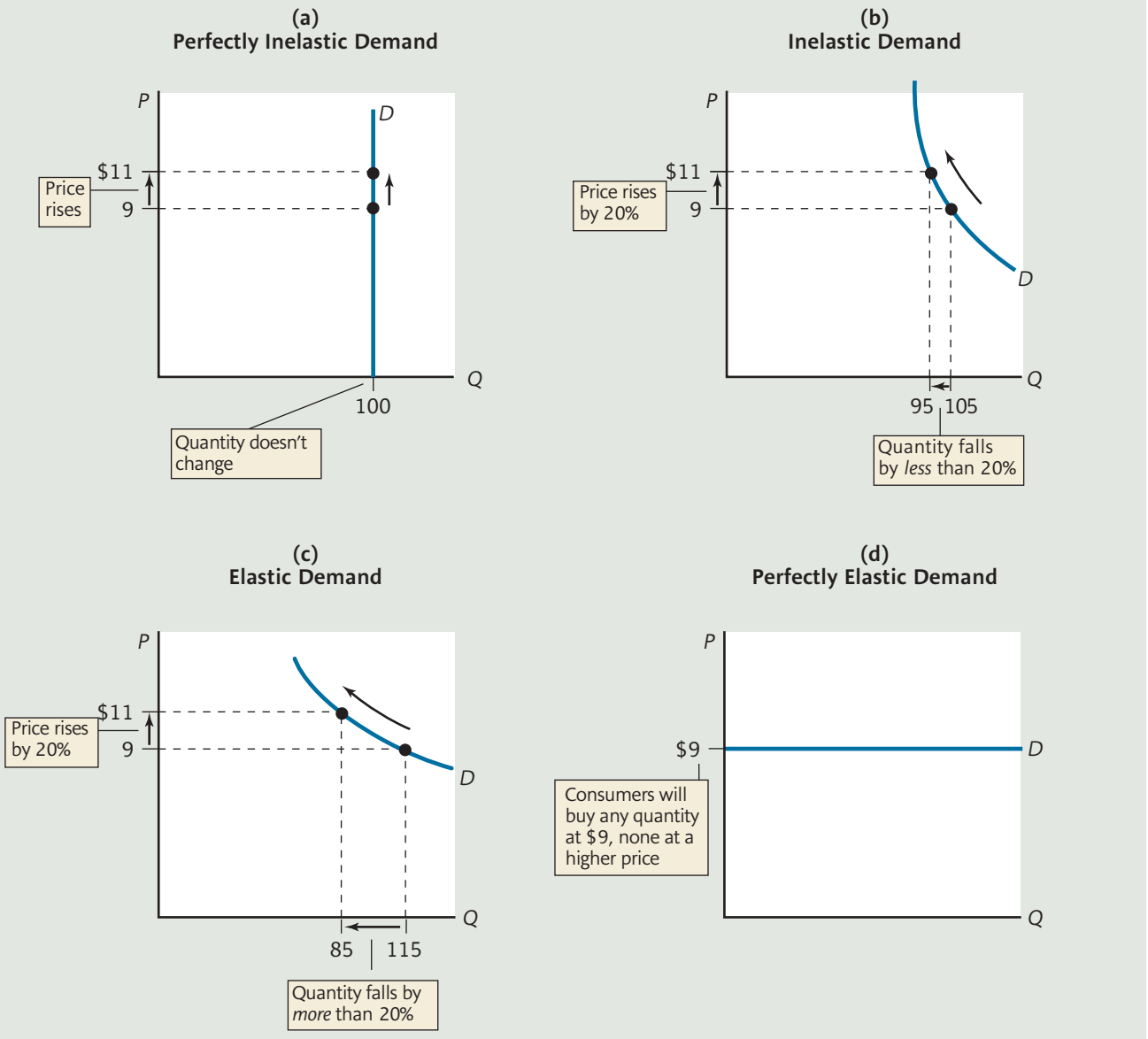
Categorizing Demand

Economists have found it useful to divide demand curves (or parts of demand curves) into categories, based on their elasticity values. These categories are illustrated in Figure 3.

Panel (a) shows an extreme theoretical case, called **perfectly inelastic demand**, where the elasticity has a value of zero. A perfectly inelastic demand curve is vertical, so a change in price causes *no* change in quantity demanded. In the figure, when price rises from \$9 to \$11 (20 percent using the midpoint formula), our formula for price elasticity of demand (E_D) gives us $E_D = 0\%/20\% = 0$.

Perfectly inelastic demand A price elasticity of demand equal to 0.

FIGURE 3 Categories of Demand Curves



Inelastic demand A price elasticity of demand between 0 and 1.

Elastic demand A price elasticity of demand greater than 1.

Panel (b) shows a case where quantity demanded has *some* sensitivity to price, but not much. Here, the same 20 percent price increase causes quantity demand to fall from 105 to 95 (a 10 percent decrease using the midpoint formula). In this case, $E_D = 10\%/20\% = 0.5$. This is an example **inelastic demand,** which occurs whenever $E_D < 1$ (quantity changes by a smaller percentage than price).

Panel (c) shows a demand curve with more price sensitivity: the 20 percent rise in price causes quantity demanded to drop by 30 percent. Our elasticity calculation is $E_D = 30\%/20\% = 1.5$. This is an example of **elastic demand,** which occurs whenever $E_D > 1$ (quantity changes by a larger percentage than price changes).

Finally, panel (d) shows another extreme case, called **perfectly elastic demand,** where the demand curve is horizontal. As long as the price stays at one particular value (where the demand curve touches the vertical axis), *any* quantity might be demanded. But even the tiniest price rise would cause quantity demanded to fall to zero. In this case, $E_D = \infty$ (elasticity is infinite) because no matter how small we make the percentage change in price (in the denominator), the percentage change in quantity (in the numerator) will always be infinitely larger.

Perfectly (infinitely) elastic demand A price elasticity of demand approaching infinity.

What about the special case when elasticity of demand is *exactly* equal to 1.0? Then demand is neither elastic nor inelastic, but lies between these categories. We call this case **unit elastic.** Take a moment to draw a demand curve that is unit elastic for a price change from \$9 to \$11, choosing your numbers for quantity carefully.

Unit elastic demand A price elasticity of demand equal to 1.

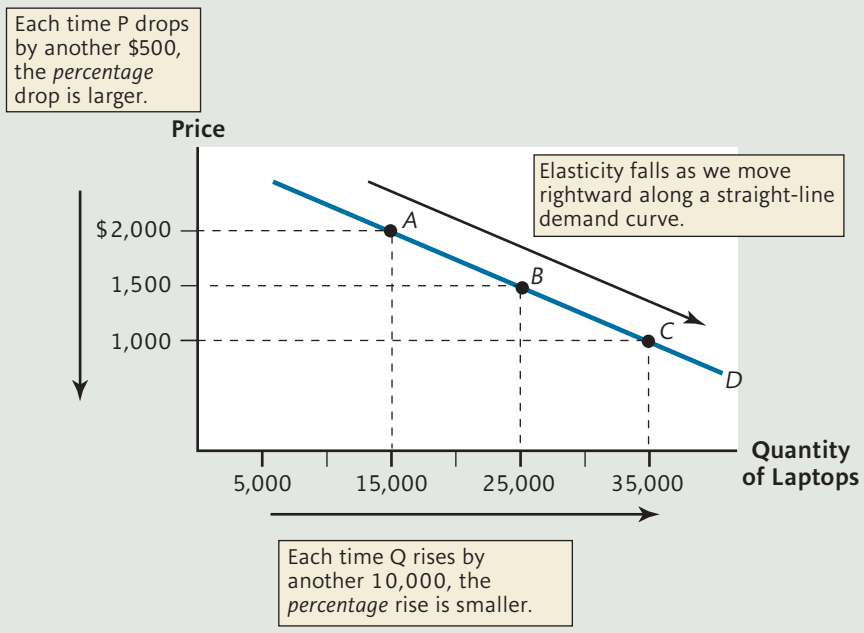
Elasticity and Straight-Line Demand Curves

Figure 4 shows a linear (straight-line) demand curve for laptop computers. Each time price drops by \$500, the quantity of laptops demanded rises by 10,000. Because this behavior remains constant all along the curve, is the price elasticity of demand also constant?

Actually, no. Elasticity is the ratio of *percentage* changes; what remains constant along a linear demand curve is the ratio of *absolute* or *unit* changes.

In fact, we can show that as we move rightward along a linear demand curve, the elasticity always decreases. For example, let's calculate the elasticity between points *A* and *B*. Price falls from \$2,000 to \$1,500, a 28.6 percent drop using the midpoint formula. Quantity rises from 15,000 to 25,000, which is a 50 percent rise

FIGURE 4 How Elasticity Changes along a Straight-Line Demand Curve



using the midpoint formula. Taking the ratio of these changes, we find that the, elasticity for a move from point A to B is 50%/28.6% = 1.75.

Now let's calculate the elasticity between points B and C, where price falls from \$1,500 to \$1,000, and quantity rises from 25,000 to 35,000. For this change (as you can verify), price falls by 40 percent while quantity demanded rises by 33.3 percent (using the midpoint formula). So the elasticity for a move from point B to C is 33.3%/40% = 0.83.

Notice what's happened: as we've moved downward and rightward along this straight-line demand curve, elasticity has fallen from 1.75 to 0.83. Demand has become *less elastic*.

There is a good reason for this. As we travel down the demand curve, the average quantity we use as the base for figuring percentage changes keeps increasing. So a constant 10,000 increase in quantity becomes a smaller and smaller *percentage* increase. The opposite also happens with price: It keeps getting smaller, so the same \$500 decrease in price becomes a growing *percentage* decrease. Thus, as we travel down a linear demand curve, with $\%\Delta Q^D$ shrinking and $\%\Delta P$ growing, the ratio $\%\Delta Q^D/\%\Delta P$ decreases.

Elasticity of demand varies along a straight-line demand curve. More specifically, demand becomes less elastic (E_D gets smaller) as we move downward and rightward.

This is a special conclusion about *linear* demand curves only. For *non*linear demand curves, moving down the curve can cause elasticity to rise, fall, or remain constant, depending on the shape of the curve.

Elasticity and Total Revenue

When the price of a good increases, the law of demand tells us that people will buy less of it. But this does not necessarily mean they will *spend* less on it. After the price rises, fewer units will be purchased in the market, but each unit will cost more. What happens to *total spending* on the good? Or, recognizing that total spending by all buyers equals the total revenue of all sellers, we can ask the same question this way: When price rises, what happens to the *total combined revenue* of all firms that sell in the market? Let's see. On the one hand, each unit sold can be sold for more, tending to increase revenue. On the other hand, fewer units will be sold, which works to decrease revenue. Which one will dominate?

The answer depends on the price elasticity of demand for the good. To see why, note that the total revenue of sellers in a market (TR) is the price per unit (P) times the quantity that people buy (Q):

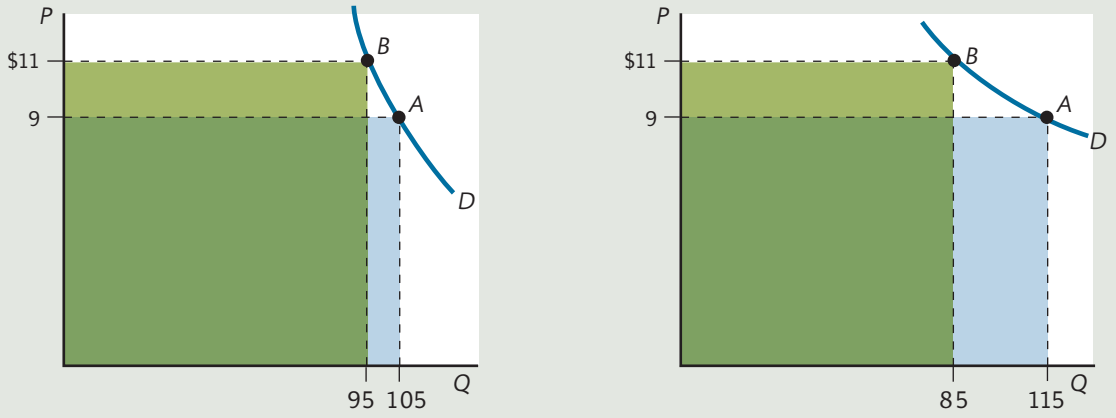
$$TR = P \times Q.$$

When we raise price, P goes up, but Q goes down. What happens to the product depends on which one changes by a larger percentage.

Suppose that demand is *inelastic* ($E_D < 1$). Then a 1 percent rise in price will cause quantity demanded to fall by *less* than 1 percent. So the greater amount sellers get on each unit outweighs the impact of the drop in quantity, and total revenue will *rise*.

The behavior of total revenue can be seen very clearly on a graph, once you learn how to interpret it. Look at the left panel of Figure 5, which duplicates the inelastic

FIGURE 5 Elasticity and Total Revenue



In panel (a), demand is inelastic, so a rise in price causes total revenue to increase. Specifically, at a price of $9 (point A), total revenue is $9 × 105 = $945. When price rises to $11 (point B), total revenue increases to $11 × 95 = $1,045. In panel (b), demand is elastic, so a rise in price causes total revenue to decrease. Specifically, at a price of $9 (point A), total revenue is $9 × 115 = $1,035. When price rises to $11 (point B), total revenue falls to $11 × 85 = $935.

demand curve introduced earlier. On this demand curve, let's start at a price of $9, and look at the rectangle with a corner at point *A*. The height of the rectangle is the price of $9, and the width is the quantity of 105, so its *area* (height × width = $P \times Q$ = $9 × 105 = $945) is the total revenue of sellers when price is $9.

More generally,

At any point on a demand curve, sellers' total revenue is the area of a rectangle with height equal to price and width equal to quantity demanded.

Now let's raise the price to $11. The total revenue rectangle becomes the larger one, with a corner at point *B*. The area of this rectangle is *TR* = $11 × 95 = $1,045. The rise in price has *increased* total revenue.

Now suppose that demand is *elastic* ($E_D > 1$). Once again, a 1 percent rise in price causes quantity demanded to fall, but this time it falls by *more* than 1 percent. So the fact that sellers get more on each unit is outweighed by the drop in the quantity they sell, and total revenue *falls*.

This is shown in the right panel of Figure 5, using the example of elastic demand from a few pages earlier. When price is $9, *TR* is the area of the rectangle with a corner at point *A*, equal to $9 × 115 = $1,035. When price rises to $11, *TR* becomes the area of the taller rectangle with corner at point *B*. This area is $11 × 85 = $935. Because demand is elastic, the rise in price *decreases* total revenue.

We can conclude that:

An increase in price raises total revenue when demand is inelastic, and shrinks total revenue when demand is elastic.

TABLE 1

Effects of Price Changes on Revenue

Where Demand Is:	A Price Increase Will:	A Price Decrease Will:
inelastic ($E_D < 1$)	increase revenue	decrease revenue
unit elastic ($E_D = 1$)	cause no change in revenue	cause no change in revenue
elastic ($E_D > 1$)	decrease revenue	increase revenue

What if price fell instead of rose? Then, in Figure 5 we'd be making the reverse move: from point *B* to point *A* on each curve. And logic tells us that if demand is inelastic, total revenue must fall. If demand is elastic, the drop in price will cause total revenue to rise.

A decrease in price shrinks total revenue when demand is inelastic, and raises total revenue when demand is elastic.

What happens if demand is unit elastic? You can probably guess. This would mean that a 1 percent change in price causes a 1 percent change in quantity, but in the opposite direction. The two effects on total revenue would cancel each other out, so total revenue would remain unchanged.

Table 1 summarizes these results about elasticity and total revenue. Don't try to memorize the table, but *do* use it to test yourself: Try to explain the logic for each entry.

Determinants of Elasticity

Table 2 lists the price elasticity of demand for a variety of goods and services. The numbers are not strictly comparable, because they were obtained using different statistical methods in different years. Still, they illustrate some interesting patterns that we can explore. Why, for example, is the demand for Tide Detergent, Pepsi, and Coke highly elastic, while the demand for gasoline and eggs is so inelastic?

Availability of Substitutes

When close substitutes are available for a good, demand for it will be more elastic. If the price of ground beef rises, with all other prices held constant, consumers can easily switch to other forms of beef, or even chicken or pork. But when the price of gasoline rises, the substitutes that are available (using mass transit, carpooling, biking, or even not going places) are not as close. Thus, it is not surprising that the demand for ground beef is more elastic than the demand for gasoline.

When close substitutes are available for a product, demand tends to be more elastic.

One factor that determines the closeness of substitutes is how narrowly or broadly we define the market we are analyzing. Demand in the market for beverages as a whole will be less elastic than demand in the market for soft drinks. And demand for soft drinks will be less elastic than the demand for Pepsi. This is because when we

TABLE 2

Some Short-Run Price Elasticities of Demand

Specific Brands		Narrow Categories		Broad Categories	
Tide Detergent	2.79	Transatlantic Air Travel	1.30	Recreation	1.09
Pepsi	2.08	Ground Beef	1.02	Clothing	0.89
Coke	1.71	Pork	0.78	Food	0.67
Kellogg's Corn Flakes	2.93 to 4.05	Milk	0.54	Imports	0.58
		Cigarettes	0.25 to 0.70	Transportation	0.56
				Alcohol and Tobacco	0.26
		Electricity (New England)	0.19		
		Beer	0.26		
		Eggs	0.26		
		Gasoline	0.26		

Sources: Bwo-Nung Huang, Chin-wei Yang, and Ming-jeng Hwang, "New Evidence on Demand for Cigarettes: A Panel Data Approach," *The International Journal of Applied Economics*, 2004, vol. 1, issue 1, pages 81–97. Akbay, C., Jones, E. "Demand Elasticities and Price-Cost Margin Ratios for Grocery Products in Different Socioeconomic Groups," *Agricultural Economics*–Czech, *52*, (5), 2006, pp. 225–235. Mark A. Bernstein and James Griffin, "Regional Differences in the Price-Elasticity of Demand for Energy," *National Renewable Energy Laboratory*, Rand Corporation, 2005. M. Espey, "Explaining the Variation in Elasticity Estimates of Gasoline Demand in the United States: A Meta-Analysis," *The Energy Journal,* 17(3), (1996), pp. 49–60. Sachin Gupta et al., "Do Household Scanner Data Provide Representative Inferences from Brand Choices? A Comparison with Store Data," *Journal of Marketing Research*, Fall 1996, pp. 383ff. F. Gasmi, J. J. Laffont, and Q. Vuong, "Econometric Analysis of Collusive Behavior in a Soft-Drink Market," *Journal of Economics and Management Strategy*, Summer 1992, pp. 277–311. J. M. Cigliano, "Price and Income Elasticities for Airline Travel," *Business Economics*, September 1980, pp. 17–21. M. R. Baye, D. W. Jansen, and Jae-Woo Lee, "Advertising Effects in Complete Demand Systems," *Applied Economics*, October 1992, pp. 1087–1096. "Estimating Interrelated Demands for Meats Using New Measures for Ground and Table Cut Beef," *American Journal of Agricultural Economics*, November 1991, pp. 1182–1194.

determine the elasticity of demand in a market, we hold constant all prices outside of the market. So in determining the elasticity for Pepsi, we ask what happens when the price of Pepsi rises but the price of Coke remains constant. Since it is so easy to switch to Coke, demand is highly elastic. But in determining the elasticity for soft drinks, we ask what happens when the price of *all* soft drinks rise together, holding constant only the prices of things that are *not* soft drinks. Demand is therefore less elastic.

Some of the entries in Table 2 confirm this influence on elasticity. For example, the demand for transportation, a very broad category, is less elastic than the demand for transatlantic air travel. But other entries seem to contradict it. (Can you find one?) Remember, though, that there are other determinants of elasticity besides the narrowness or broadness of the market.

Necessities versus Luxuries

Goods that we think of as necessary for our survival or general well-being, and for which there are no close substitutes, are often referred to as "necessities." Most people would include the broad categories "food," "housing," and "medical care" in this category. When we regard something

dangerous curves

Even "Necessities" Obey the Law of Demand Don't make the common mistake of thinking that people *must* buy a constant quantity of goods they regard as necessities, and cannot make do with any less. That would imply *perfectly inelastic* demand. Studies routinely show that demand for most goods regarded as necessities is inelastic, but not perfectly inelastic. People do find ways to cut back on them when the price rises.

as a necessity, demand for it will tend to be less elastic. This is another reason why the elasticity of demand for gasoline is so small: Many people regard gasoline as a necessity.

By contrast, goods that we can more easily do without—such as entertainment or vacation travel—are often referred to as "luxuries." Demand for these goods will tend to be more elastic, since people will cut back their purchases more when price rises.

Goods we regard as necessities tend to have less elastic demand than goods we regard as luxuries.

In Table 2, for example, among the broad categories, the demands for food and clothing are less elastic than the demand for recreation. Remember, though, that how broadly or narrowly we define the market makes an important difference. We may regard broadly defined "clothing" (not in the table) as a necessity. But "designer jeans" might be an easy-to-substitute for luxury, with a highly elastic demand.

Importance in Buyers' Budgets

When a good takes up a large part of your budget initially, a rise in price has a large impact on how much you will have left to spend on other things. All else equal, this will tend to make demand more elastic. For example, a vacation trip to Paris would take a big bite out of most peoples' budgets. If the price of the vacation rises by, say, 20 percent, many people will start to consider other alternatives, since not doing so would mean a considerable sacrifice of other purchases.

Now consider the other extreme: ordinary table salt. A family with an income of $50,000 per year would spend less than 0.005 percent of its income on this good, so the price of salt could double or triple and have no significant impact on the ability to buy other goods. We would therefore expect the demand for table salt to be inelastic.

This insight helps us to explain some seemingly anomalous results in Table 2. Demand for food is more elastic than the demand for eggs. Based on the narrowness of definition, we would expect the reverse. But eggs make up a rather small fraction of the typical family's budget, and certainly smaller than food as a whole. This tends to reduce the elasticity of demand for eggs.

When spending on a good makes up a larger proportion of families' budgets, demand tends to be more elastic.

Time Horizon

Short-run elasticity An elasticity measured just a short time after a price change.

Long-run elasticity An elasticity measured a year or more after a price change.

How much time we wait after a price change can have an important impact on the elasticity of demand. The elasticities of demand in Table 2 are all **short-run elasticities**: the quantity response is measured for just a short time (usually a year or less) after the price change. A **long-run elasticity** measures the quantity response after more time has elapsed—typically a few years or more. In study after study, we find that demand is almost always more elastic in the long run than in the short run.

Why? Because the longer we wait after a sustained price change, the more time consumers have to make adjustments in their lives that affect their quantity demanded. In general,

TABLE 3

Short-Run versus Long-Run Elasticities

Good	Area of Study	Short-Run Elasticity	Long-Run Elasticity
Gasoline	United States	0.26	0.58
	Britain	0.25	0.60
Electricity	U.S.–New England	0.19	0.33
	U.S.–Pacific Coast	0.19	0.25
Oil	United States	0.06	0.45
	France	0.07	0.57
	Japan	0.07	0.36
	Denmark	0.03	0.19
Mass Transit	Multi-Country	0.2 to 0.5	0.6 to 0.9

Sources: Molly Espey, "Explaining the Variation in Elasticity Estimates of Gasoline Demand in the United States: A Meta-Analysis," *Energy Journal*, Vol. 17, no. 3, pp. 49–60 (1996); Phil Goodwin, Joyce Dargay, and Mark Hanly, "Elasticities of Road Traffic and Fuel Consumption with Respect to Price and Income: A Review," (2003); ESRC Transport Studies Unit, University College London (www.transport.ucl.ac.uk); Mark A. Bernstein and James Griffin, "Regional Differences in the Price-Elasticity of Demand for Energy." Prepared for the National Renewable Energy Laboratory, Rand Corporation, 2005; John C. B. Cooper, "Price Elasticity of Demand for Crude Oil: Estimates from 23 Countries," *OPEC Review: Energy Economics & Related Issues*, 27:1 (March 2003); Todd Litman, "Transit Price Elasticities' and Cross-Elasticities," *Journal of Public Transportation*, Vol. 7, No. 2, 2004.

the longer we wait after a price change to measure the quantity response, the more elastic is demand. Therefore, long-run elasticities tend to be larger than short-run elasticities.

Table 3 shows some examples of short-run and long run elasticities related to energy and transportation. In each case demand is more elastic in the long-run than in the short run.

For example, the *short-run* elasticity for gasoline in many countries is typically estimated at around at 0.26, while the long-run elasticity is two or three times higher. What accounts for the difference?

Table 4, at the bottom of the page, provides some of the answers.

TABLE 4

Adjustments After a Rise in the Price of Gasoline

Short Run (less than a year)	Long Run (a year or more)
Use public transit more often	Buy a more fuel-efficient car
Arrange a car pool	Move closer to your job
Check tire pressure more often	Switch to a job closer to home
Drive more slowly on the highway	Move to a city where less driving is required
Eliminate unnecessary trips (use mail order instead of driving to stores; locate goods by phone or Internet instead of driving around; shop for food less often and buy more each time)	
If there are two cars, use the more fuel-efficient one	

It lists some of the ways people can adjust to a significant rise in the price of gasoline over the short run and the long run. Remember that the adjustments in the long-run column are *additional* adjustments people can make if given enough time. While some of them may seem extreme, thousands of families made these changes during the latter half of the 1970s, after the OPEC nations reduced oil supplies and the price of gasoline roughly quadrupled. In 2007 and 2008, when oil and gasoline prices surged again, many families began planning similar adjustments until prices started dropping in mid-2008.

Time Horizons and Demand Curves

Our analysis of short-run and long-run elasticities might have raised a question in your mind. Isn't price elasticity of demand measured *along* a demand curve? Indeed it is. But then how can we get two different elasticity measures from the same market?

The answer is: There can be more than one demand curve associated with a market. Whenever we draw a demand curve, we draw it for a *specific* time horizon. *Short-run* demand curves show quantity demanded at different prices when people only have a short period of time (a few weeks or a few months) to adjust. A *long-run* demand curve shows quantity demanded after buyers have had much longer—say, a year or more—to adjust to a price change.

The three demand curves in Figure 6 illustrate how this works in the U.S. market for gasoline. We assume that initially we are at point *A*, with a price of $2 per gallon and quantity demanded of 400 million gallons per day. As long as the price—and every other influence on demand—remains the same, we would stay at point *A*.

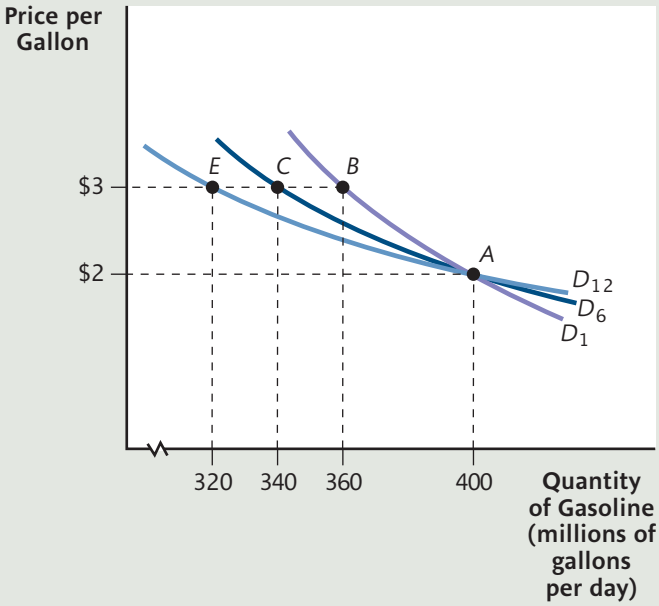
But now, suppose the price rises from $2 to $3 and stays there. To find the daily quantity demanded *one month later,* we would move along the demand curve labeled D_1. This demand curve has a very low elasticity of demand because there is not much time for gasoline buyers to adjust. For example, they might cut out some unnecessary trips, but they are unlikely to purchase a more fuel efficient car in this time frame. Using demand curve D_1, we end up at point *B*, with gasoline demand equal to 360 million gallons per day.

To find the quantity demanded *six months later,* we would move along the demand curve labeled D_6. Demand curve D_6 is more elastic than D_1, because we are allowing buyers more time to make adjustments that will reduce their purchases (e.g., for some gasoline buyers, 6 months is long enough to acquire a more fuel efficient car). So, if we wait 6 months, we'll find that we've moved from point A to point *C*, with consumers buying 340 million gallons of gas per day.

Finally, to find daily quantity demanded *12 months later,* we would move along the demand curve labeled D_{12}. This demand curve is more elastic than the other two, because we've allowed buyers even more time to adjust to the higher price. So if we measure the quantity response after waiting a full year, we'll find that we've moved from point *A* to point *E*, where quantity demanded has fallen to 320 million gallons per day.[1]

[1] You might have noticed that in Figure 6, the flatter the curve, the more elastic is demand. Yet earlier, we argued that a flatter curve is not necessarily a more elastic curve. Why does flatter mean more elastic in the figure? Because in this case, all three curves share the same starting point (point *A*) for our measurement of elasticity. With the same starting point, the flatter the curve, the greater the change in quantity (in both units and percentage). Since the percentage change in price is the same along all three curves, when we move along a flatter curve (which has a larger percentage change in quantity), demand will always be more elastic. But be careful: this simple correspondence between slope and elasticity only works when we start or end at the same point along two curves.

FIGURE 6 Different Demand Curves for Different Time Horizons



When the price of gasoline rises by $1, the decrease in quantity demanded (and the price elasticity of demand) depends on how long we wait before measuring buyers' response. If we wait just one month after the price change, we'd move along demand curve D_1, *from point* A *to point* B. *If we wait six months, we'd move along demand curve* D_6, *from point* A *to point* C. *The same rise in price causes a greater decrease in quantity demanded after six months, because buyers can make further adjustments. If we wait 12 months, we'd move from point* A *to point* E *along demand curve* D_{12}, *with quantity demanded falling even more.*

Any demand curve is drawn for a particular time horizon (a waiting period before we observe the new quantity demanded after a price change). In general, the longer the time horizon, the more elastic the demand.

As a rule of thumb, demand curves drawn for time horizons less than one year are called short-run demand curves, while those drawn for time horizons of one year or longer are called long-run demand curves.

Two Practical Examples

Knowing the price elasticity of demand for a good or service can be highly useful to economists in practical work. In this section we provide two examples.

Elasticity and Mass Transit

Earlier in this chapter, you were asked to imagine that you were a mayor trying to determine whether raising mass transit fares would increase or decrease city revenues. Now you know that the answer depends on the price elasticity of demand for mass transit.

Several studies[2] have shown that the demand for mass transit services is *inelastic*. In the short run (the first year after the price change), the elasticity of demand ranges from 0.2 to 0.5. Over the long run (five to ten years), elasticity values are 0.6 to 0.9. Notice that although elasticity is greater in the long run than the short run, demand remains inelastic even in the long run. This means that a rise in fares would likely raise mass-transit revenue for a city, even in the long run.

[2] These studies are nicely summarized in Todd Litman, "Transit Price Elasticities and Cross-Elasticities," *Journal of Public Transportation*, Vol. 7, No. 2, 2004, pp. 37–58.

Let's do an example. Suppose New York City raised subway and bus fares from \$2.00 to \$2.50, a 25 percent increase.[3] What would happen to revenue?

In the long run, elasticity is between 0.6 and 0.9, so let's choose an estimate that falls near the middle of that range: 0.7. Using this elasticity, each 1 percent increase in fares would cause a 0.7 percent decrease in ridership. So our 25 percent fare hike would decrease ridership by $25 \times 0.7 = 17.5$ percent. New Yorkers take about 2 billion trips each year, so a 17.5 percent decrease would bring trips down by 350 million, to 1.65 billion per year.

Now let's calculate revenue, before and after the price hike. At the current price of \$2, total revenue is 2 billion trips × \$2 per trip = \$4 billion. After the price hike, total revenue would be 1.65 billion trips × \$2.50 per trip = \$4.125 billion. We conclude that raising the fare would increase mass transit revenue by about \$125 million per year in the long run.

Why, then, doesn't New York City raise the fare to \$2.50? In fact, why doesn't it go further, to \$3? Or \$5? Or even \$10?

For two reasons. First, elasticity estimates come from *past* data on the response to price changes. When we ask what would happen if we raise the price, we are extrapolating from these past responses. For small price changes, the extrapolation is likely to be fairly accurate. But large price changes move us into unknown territory, and elasticity may change. Demand could be *elastic* for a very large price hike, and then total revenue would fall. This puts a limit on fares, even if a city's goal is the maximum possible revenue.

A second (and more important) reason is that generating revenue is only *one* consideration in setting mass transit fares. City governments are also concerned with providing affordable transportation to city residents, reducing traffic congestion on city streets, and limiting pollution. A fare increase, even if it would raise total revenue, would work against these other goals. This is why most cities keep the price of mass transit below the revenue-maximizing price.

Elasticity and an Oil Crisis

For the past five decades, the Persian Gulf has been a geopolitical hot spot. And the stakes for the rest of the world are high, because the region produces about one-fourth of the world's oil supply. That is why economists in government and industry are constantly asking "what if" questions about the world market for oil. One central question is this: If an event in the Persian Gulf were to disrupt oil supplies, what would happen to the price of oil on world markets? Not surprisingly, elasticity plays a crucial role in answering this question.

As you can see in Table 3, the short-run elasticity of demand for oil is about 0.06. This estimate is for the United States, but in our analysis, we'll use it as the price elasticity of demand for oil in the global market as well. Since a political or military crisis is usually a short-run phenomenon, the short-run elasticity is what we are interested in. But for this problem, we need to use elasticity in a new way. Remember that elasticity tells us the percentage decrease in quantity demanded for a 1 percent increase in price. But suppose we flip the elasticity fraction upside down, to get

$$\frac{1}{E_D} = \frac{\%\text{ Change in Price}}{\%\text{ Change in Quantity Demanded}}.$$

[3] Notice that we're calculating percentage changes the conventional way, rather than with the midpoint formula. Here, we are not trying to calculate an elasticity value. Instead, we're using one that has already been determined for us. While we could continue to use the midpoint formula, it would be cumbersome—especially when trying to determine the new quantity after the fare hike.

This number—the inverse of elasticity—tells us the percentage rise in price that would bring about each 1 percent decrease in quantity demanded. For oil, this number is 1/0.06 = 16.67. What does this number mean? It tells us that to bring about each 1 percent decrease in world oil demand, oil prices would have to rise by 16.67 percent.

Now we can make reasonable forecasts about the impact of various events on oil prices. Imagine, for example, an event that temporarily removed half of the Persian Gulf's oil from world markets. And let's assume a worst-case scenario: No other nation increases its production during the time frame being considered. What would happen to world oil prices?

Since the Gulf produces about 25 percent of the world's oil, a reduction by half would decrease world oil supplies by 12½ percent. It would then require a price increase of 12½ × 16.67 = 208 percent to restore equilibrium to the market. If oil were initially selling at $60 per barrel, we could forecast the price to rise by $60 × 2.08 = $124.80 per barrel, for a final price of about $184.80.

Of course, this analysis can only give us a rough approximation. For one thing, it assumes that the elasticity would remain at the low level of 0.06, even for a huge change in price. In fact, the elasticity value might change as the price rises—and demand could become substantially more elastic at very high prices, even in the short run. In that case (as you'll be asked to verify in one of the end-of-chapter questions), the price would not rise so high. Still, the analysis shows the crucial role elasticity plays in predicting how an external event can influence the market price.

Other Elasticities

The concept of *elasticity* is a very general one. It can be used to measure the sensitivity of virtually *any* variable to any other variable. All types of elasticity measures, however, share one thing in common: They tell us the percentage change in one variable caused by a 1 percent change in the other. Let's look briefly at three additional elasticity measures, and what each of them tells us.

Income Elasticity of Demand

You learned in Chapter 3 that household income is one of the variables that influences demand. The *income elasticity of demand* tells us how *sensitive* quantity demanded is to changes in buyers' incomes. The **income elasticity of demand** E_Y is the percentage change in quantity demanded divided by the percentage change in income, with all other influences on demand—including the price of the good—remaining constant:

Income elasticity of demand
The percentage change in quantity demanded caused by a 1 percent change in income.

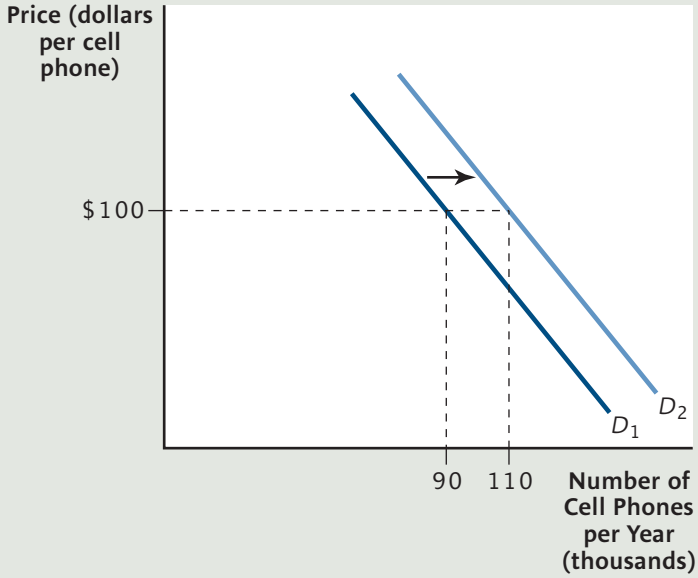
$$\text{Income Elasticity} = \frac{\%\ \text{Change in Quantity Demanded}}{\%\ \text{Change in Income}}$$

Keep in mind that while a price elasticity measures the sensitivity of demand to price as we *move along the demand curve* from one point to another, an income elasticity tells us the relative *shift* in the demand curve—the percentage increase in quantity demanded *at a given price*.

Figure 7 illustrates how we might calculate an income elasticity in the market for cell phones. In the figure, we assume that the average income of buyers in the

FIGURE 7 **Income Elasticity and Demand Curves**

Income elasticity is the percentage increase in demand (at a given price) divided by the percentage change in income. In the figure, we assume that income rises by 10%, causing quantity demanded at a price of $100 to rise from 90 to 110, or—using the midpoint rule—by 20/100 = 20%. Thus, the income elasticity of demand is 20%/10% = 2.0.



market rises by 10%. (Note that income itself is not shown in the graph.) As a result, *at a given price* ($100 in the figure), the quantity of cell phones demanded rises from 90 thousand to 110 thousand. Using our midpoint rule, we find that the percentage change in quantity is 20/100 = 20%. So, using the formula, the income elasticity of demand would be 20%/10% = 2.0.

Note that with income elasticities (unlike price elasticities), the sign of the elasticity value matters. Income elasticity will be positive when people want more of a good as their income rises. Such goods are called *normal* goods, as you learned in Chapter 3. But income elasticity can also be negative, when a rise in income *decreases* demand for a good (*inferior* goods.)

Income elasticity is positive for normal goods, but negative for inferior goods.

Inter-city bus travel is, in many markets, an inferior good. As household income rises, travelers are likely to shift away from cheaper bus travel to more expensive car, train, or airline travel. Similarly, as income rises, many households will shift from cheaper sources of calories (e.g., rice and beans) to more expensive items (steak, fresh fruit, and sushi).

Of course, when household income *falls*, the demand for inferior goods rises. During the recession of 2008–2009, the media reported significantly higher sales of goods such as pancake mix, rice, and beans—all cheap sources of calories that, in many markets, seem to be inferior goods (negative income elasticity).

An accurate knowledge of income elasticity can be crucial in predicting the growth in demand for a good as income grows over time. For example, economists know that different types of countries have different income elasticities of demand for oil. (In less-developed countries undergoing rapid industrialization, the income elasticity of demand for oil is typically twice as large as in developed countries.)

These income elasticities, along with forecasts of income growth in different developing and developed countries, are used to predict global demand for oil and forecast future oil prices.

Cross-Price Elasticity of Demand

A cross-price elasticity relates the percentage change in quantity demanded for one good to the percentage change in the price of another good. More formally, we define the **cross-price elasticity of demand** between good X and good Z as:

Cross-price elasticity of demand The percentage change in the quantity demanded of one good caused by a 1 percent change in the price of another good.

$$\frac{\%\text{ Change in Quantity Demanded of } X}{\%\text{ Change in Price of } Z}$$

In words, a cross-price elasticity of demand tells us the percentage change in quantity demanded of a good for each 1 percent increase in the price of some other good, while all other influences on demand remain unchanged.

With a cross-price elasticity (as with an income elasticity), the sign matters. A *positive* cross price elasticity means that the two goods are *substitutes:* A rise in the price of one good increases demand for the other good. For example, Coke and Pepsi are clearly substitutes, and the cross-price elasticity of Pepsi with Coke has, in one study, been estimated at 0.8.[4] This means that a 1 percent rise in the price of Coke, holding constant the price of Pepsi, causes a 0.8 percent *rise* in the quantity of Pepsi demanded.

Similarly, gasoline and mass transit are substitutes. One study (for Philadelphia) found that the cross-price elasticity of mass transit with the price of gasoline (and other trip-related automobile costs) was equal to 2.69. In simple English: A 10-percent rise in the cost of using an automobile for a trip would cause a 27 percent rise in mass transit use.

A *negative* cross-price elasticity means that the goods are *complements:* A rise in the price of one good *decreases* the demand for the other. Thus we'd expect higher gasoline prices to decrease the demand for large, fuel-inefficient cars. Indeed, when gasoline prices spiked in 2007 and 2008, the demand for SUVs and other gas-guzzling vehicles plunged.

Price Elasticity of Supply

The **price elasticity of supply** is the percentage change in the quantity of a good supplied that is caused by a 1 percent change in the price of the good, with all other influences on supply held constant.

Price elasticity of supply The percentage change in quantity supplied of a good or service caused by a 1 percent change in its price.

$$\text{Price Elasticity of Supply} = \frac{\%\text{ Change in Quantity Supplied}}{\%\text{ Change in Price}}$$

The price elasticity of supply measures the sensitivity of quantity supplied to price changes as we move *along* the supply curve. A large value for the price elasticity of supply means that quantity supplied is very sensitive to price changes. For example, an elasticity value of 5 would imply that if price increased by 1 percent, quantity supplied would rise by 5 percent.

[4] F. Gasmi, J. J. Laffont, and Q. Vuong, "Econometric Analysis of Collusive Behavior in a Soft Drink Market," *Journal of Economics and Management Strategy*, Summer, 1992, pp. 277–311.

A major determinant of supply elasticity is the ease with which suppliers can find profitable activities that are *alternatives* to producing the good in question. In general, supply will tend to be more elastic when suppliers can switch to producing alternate goods more easily.

When can we expect suppliers to have easy alternatives? First, the nature of the good itself plays a role. All else equal, the supply of envelopes should be more elastic than the supply of microprocessor chips. This is because envelope producers can more easily modify their production lines to produce alternative paper products. Microprocessor suppliers, however, would be hard-pressed to produce anything other than computer chips.

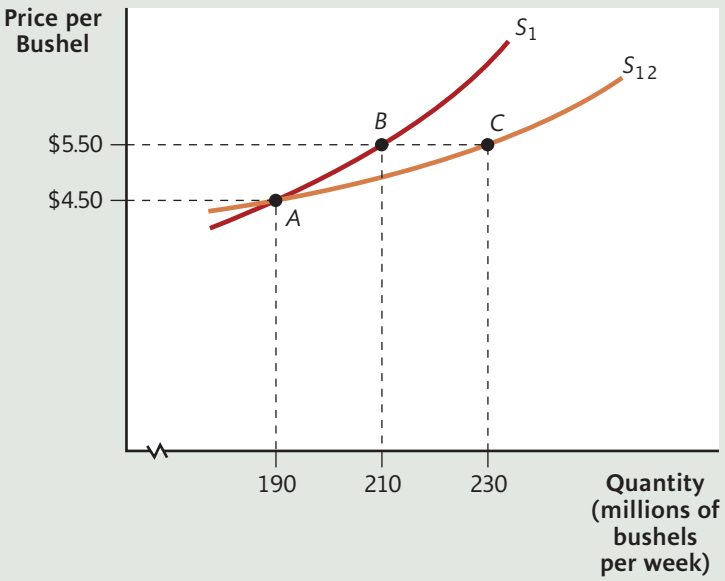
The narrowness of the market definition matters too—especially *geographic* narrowness. For example, the market for oranges in Illinois should be more supply-elastic than the market for oranges in the United States. In the Illinois market, a decrease in price would imply we are holding constant the price of oranges in *all other states*. This gives suppliers an easy alternative: They could sell their oranges in other states! Similarly, the supply of oranges to Chicago would be even more elastic than the supply of oranges to Illinois.

Finally, the *time horizon* is important. The longer we wait after a price change, the greater the supply response to a price change. As we will see when we discuss the theory of the firm, there usually is *some* response to a price change right away. Existing firms simply speed up or slow down production with their current facilities. But further responses come about as firms have time to change their plant and equipment, and new firms have time to enter or leave an industry.

Figure 8 illustrates how we can calculate the short-run and long-run elasticities of supply for U.S. corn farmers. Initially, the market is at point *A*, with a price of $4.50 per bushel and quantity supplied of 190 million bushels per week. When the price rises to $5.50 per bushel, and we measure the quantity supplied one month

FIGURE 8 Different Supply Curves for Different Time Horizons

When the price of corn rises from $4.50 to $5.50 per bushel, the increase in quantity supplied (and the price elasticity of supply) depends on how long we wait before measuring the response. If we wait just one month after the price change, we'd move along supply curve S_1, *from point* A *to point* B. *If we wait 12 months, we'd move along supply curve* S_{12}, *from point* A *to point* C. *The same rise in price causes a greater increase in quantity supplied after 12 months, because farmers can make further adjustments in quantity supplied if given more time.*



later, we move along supply curve S_1 and quantity supplied rises to 210 million bushels per week.

Using the mid-point rule, the percentage change in the price is \$1/\$5.00 = 20%, while the percentage change in quantity is 20/200 = 10%. Therefore, the short-run supply elasticity for corn is equal to $\%\Delta Q^S/\%\Delta P = 10\%/20\% = 0.50$.

What if we waited 12 months after the price hike before measuring the quantity response? Then we would move along the more-elastic supply curve S_2. Quantity is more sensitive to price in the long run, because farmers have the time to make further adjustments to increase production, such as shifting land from other crops to corn.

Supply for many products becomes more elastic the longer we allow sellers to respond to a price change. In general, long-run supply elasticities are greater than short-run supply elasticities.

Using the Theory

© DAN LAMONT/CORBIS

APPLICATIONS OF ELASTICITY

Elasticity is among the most widely used tools in microeconomics. You've already seen two examples of how economists use elasticity: predicting mass transit revenues and planning for an oil crisis. In this section, we'll discuss two extended examples in which elasticity is key to analyzing events in markets.

The War on Drugs: Should We Fight Supply or Demand?

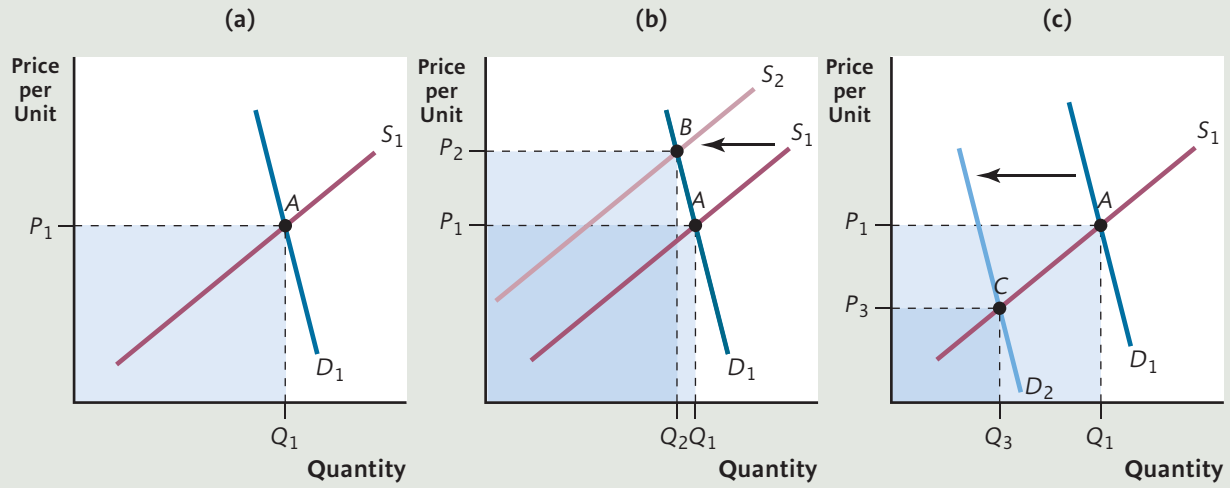
Every year, the U.S. government spends about \$10 billion intervening in the market for illegal drugs like cocaine, heroin, and marijuana. Most of this money is spent on efforts to restrict the supply of drugs. But many economists argue that society would be better off if antidrug efforts were shifted from the supply side to the demand side of the market. Why? The answer hinges on the price elasticity of demand for illegal drugs.

Look at Figure 9, which shows the market for heroin if there were no government intervention. The equilibrium would be at point *A*, with price P_1 and quantity Q_1. Total revenue of sellers—and total spending by buyers—would be the area of the shaded rectangle, $P_1 \times Q_1$.

Figure 9 shows the impact of a policy to restrict supply through any one of several methods, including vigilant customs inspections, arrest and stiff penalties for drug dealers, or efforts to reduce drug traffic from producing countries like Colombia, North Korea, and Thailand. The decrease in supply is represented by a leftward shift of the supply curve, establishing a new equilibrium at price P_2 and quantity Q_2. As you can see, supply restrictions, if they successfully reduce the equilibrium quantity of heroin, will also raise its equilibrium price.

But now let's consider the impact of this policy on the users' total *expenditure* on drugs. Although the research is difficult, a number of studies have concluded that the demand for addictive drugs such as heroin and cocaine is very price *inelastic*. This is not surprising, given their addictive properties. As you've learned, when demand is inelastic, a rise in price will *increase* the revenue of sellers, which is the same as the total expenditure of buyers. This means that a policy of restricting the supply of illegal

FIGURE 9 The War on Drugs



Panel (a) shows the market for heroin in the absence of government intervention. Total expenditures—and total receipts of drug dealers—are given by the area of the shaded rectangle. Panel (b) shows the effect of a government effort to restrict supply: Price rises, but total expenditure increases. Panel (c) shows a policy of reducing demand: Price falls, and so does total expenditure.

drugs, if successful, will also increase the total expenditure of drug users on their habit. In panel (b), total expenditure rises from the area of the shorter rectangle to the area of the taller one.

The change in total expenditure has serious consequences for our society. Many drug users support their habit through crime. If the total expenditure needed to support a drug habit rises, they may commit more crimes—and more serious ones. And don't forget that the total expenditure of drug users is also the total *revenue* of the illegal drug industry. The large revenues—and the associated larger profits to be made—attract organized as well as unorganized crime and lead to frequent and very violent turf wars.

The same logic, based on the inelastic demand for illegal drugs, has led many economists to advocate the controlled legalization of most currently illegal drugs. Others advocate a shift of emphasis in the war from decreasing supply to decreasing demand. Policies that might decrease the demand for illegal drugs and shift the demand curve leftward include stiffer penalties on drug *users*, heavier advertising against drug use, and greater availability of treatment centers for addicts. In addition, more of the effort against drug sellers could be directed at retailers rather than those higher up the chain of supply. It is the retailers who promote drugs to future users and thus increase demand.

Figure 9 illustrates the impact these policies, if successful, would have on the market for heroin. As the demand curve shifts leftward, price *falls* from P_1 to P_3, and quantity demanded falls from Q_1 to Q_3. Now, we cannot say whether the drop in quantity will be greater under a demand shift than a supply shift (it depends on the relative sizes of the shifts). But we *can* be sure that a demand-focused policy will have a very different impact on equilibrium price, moving it down instead of up. Moreover, the demand shift will decrease total expenditure on drugs—to the *inner* shaded rectangle—since both price and quantity decrease. This can contribute to a lower crime rate by drug users and make the drug industry less attractive to potential dealers and producers.

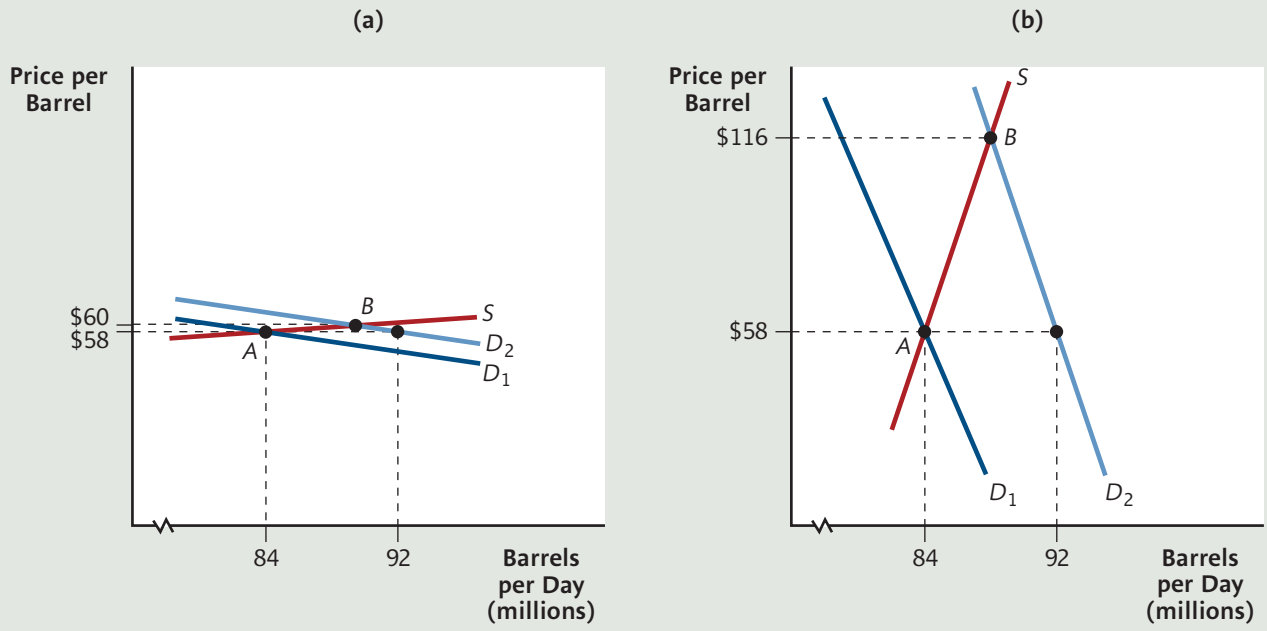
Fluctuating Commodity Prices: Causes and Consequences

In Chapter 3, we looked at the spike in oil prices that occurred in 2007 and 2008. But oil was just one among many commodities that surged in price during that period. Copper, iron ore, palm oil, wheat, rice, and many more commodities experienced similar rapid price hikes. One cause in all of these cases was an increase in demand due to rapid economic growth. For a few commodities, supply disruptions played a role as well. Income growth and supply disruptions can explain *why* prices rose. But why did they rise *so* much and *so* rapidly? It turns out that elasticities can help us find the answer.

Let's first go back to oil. Figure 10 shows two versions of the oil market. Panel (a) shows a fictional version of this market—what it would look like if both supply and demand were very elastic in the short run. Panel (b) shows a more realistic picture of the oil market—very *inelastic* supply and demand in the short run. In both panels, our initial equilibrium is point *A*—with price equal to $58 per barrel, and quantity equal to 84 million barrels per day.

Now, let's see what happens when demand increases due to global growth. In both panels, we assume the demand curve shifts rightward by 8 million barrels per day—a rough but reasonable figure for the increase in demand during the period from January 2007 to mid-2008. At the old price of $58 per barrel, there is now an

FIGURE 10 **Price Change Caused by an Increase in Demand**



If either supply or demand is very elastic, only a relatively small price increase will be needed to restore equilibrium after an increase in demand. Panel (a) illustrates the case in which both *supply and demand are very elastic. The rightward shift in the demand curve causes equilibrium price to rise from $58 to $60. Panel (b) has the same rightward shift in the demand curve, but with much less elastic supply and demand. A much larger rise in price (from $58 to $116) is needed to restore equilibrium after the increase in demand.*

excess demand of 8 million barrels per day—roughly 9%. And in both panels, the price will rise—and keep rising until the excess demand has been eliminated. This requires a *combined* change in quantity demanded and quantity supplied of about 9%. How much will the price have to rise?

That depends on *both* the price elasticity of demand *and* the price elasticity of supply. Suppose, for example, that supply and demand are both very elastic, with the same elasticity of 1.5. Then each 1% rise in price will cause quantity supplied to rise by 1.5%, and also cause quantity demanded to fall by 1.5%, eliminating about one-third (3 percentage points) of the 9% excess demand. A price rise of about three times that much—3%—would cause the excess demand to fall by the needed 9%. This is illustrated in panel (a), where—to eliminate the excess demand—price has to rise by just $2 to $60 (a roughly 3% price increase).

Now look at panel (b), which is drawn to illustrate more realistic short-run elasticities: 0.03 for supply and 0.06 for demand. Once again, the same 8 million barrel excess demand—about 9%—must be eliminated by an increase in price. But now, each 1% rise in price causes quantity supplied to rise by just .03 percent, and quantity demanded to fall by just 0.06 percent, reducing the excess demand by just 0.09 percentage points. To eliminate the excess demand now requires a price increase of 100%—a doubling of the price to $116 per barrel.[5]

As you can see, when an excess demand arises, both prices and quantities adjust to eliminate it. All else equal, the more sensitive *either* quantity supplied or quantity demanded is to price, the less price has to rise to give us the required adjustment of quantities. When demand and supply are both very *inelastic*—as is the case for oil and many other commodities—price hikes must be huge to get the needed changes in quantities.

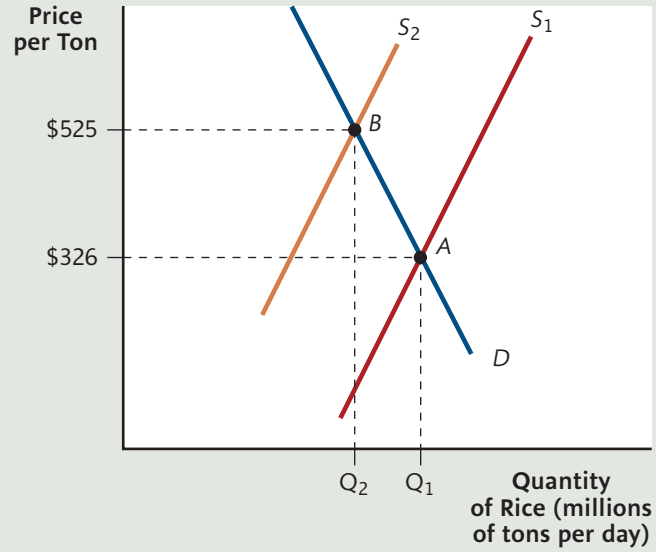
Notice, too, that it doesn't matter whether the excess demand arises from an increase in demand or a decrease is supply: In either case, the elasticities will determine how much the price must rise to close the gap. Figure 11 illustrates the impact of a decrease in *supply,* using the global market for rice. Initially, in mid-2007, with supply curve S_{2007}, the price of rice was $326 per metric ton at point *A*. Then Australia, a major rice producer, experienced severe drought in 2007 and 2008. The supply curve shifted leftward, to S_{2008}, creating an excess demand for rice at the initial price. In the figure, we assume this was the *only* change in the market for rice. With very inelastic supply and demand, the price increase had to be huge to close the gap. The result was a new equilibrium at point *B*, with the price rising (by mid-2008) to $525 per metric ton, sparking demonstrations and riots in several rice-importing countries.

In recent years, wild fluctuations in commodity prices have become increasingly common. Why?

One possible answer is that demand for basic commodities has become more *inelastic.* As incomes rise, especially in formerly poor countries like China and India, basic commodities like rice and corn take up smaller fractions of the typical family's budget. As you've learned, all else equal, the smaller the fraction of the budget, the less elastic is demand. But with less elastic demand, a supply disruption or a rise in global demand must raise price by even more in order to bring us back to equilibrium.

[5] If we use the midpoint rule for percentage changes, the new price would have to rise even more—to about $170—to give us the needed 100% increase. But when price increases are this large, either method of calculating percentage changes—midpoint or standard—is at best a very rough approximation. Among other things, elasticity values themselves are likely to change as we make such large movements along the demand and supply curves.

FIGURE 11 Price Change Caused by a Decrease in Supply



As global incomes have risen, demand for many basic commodities, such as rice, has become less elastic. So a decrease in supply requires a relatively large rise in price to restore equilibrium. In the market for rice, the decrease in supply from 2007 to 2008 caused the price to rise dramatically: from $326 to $525 per ton.

This explanation—if true—has a dark implication. Not everyone is enjoying higher incomes. Those left behind—and for whom basic staples remain a large part of their budget—remain price-sensitive. Therefore, as incomes rise *in general*, an increasing share of the price sensitivity along market demand curves is due to the response of the poorest global citizens. In essence, if quantity demanded must fall, and those with higher incomes and bigger budgets cut back less, then those with the lowest incomes must cut back even more.

SUMMARY

A useful tool for analyzing markets is *elasticity*: a measure of the sensitivity of one economic variable to another. The *price elasticity of demand* is defined as the percentage change in quantity demanded divided by the percentage change in price that caused it, without the negative sign. In general, price elasticity of demand varies along a demand curve. In the special case of a straight-line demand curve, demand becomes more and more elastic as we move upward and leftward along the curve. Along an elastic portion of any demand curve, a rise in price causes sellers' revenues (and consumers' expenditures) to fall. Along an *inelastic* portion of any demand curve, a rise in price causes sellers' revenues and consumers' expenditures to increase. Generally speaking, demand for a good tends to be more elastic the less we regard the good as a "necessity," the easier it is to find substitutes for the good, the greater the share of households' budgets that is spent on the good, and the more time we allow for quantity demanded to respond to the price change.

In addition to price elasticity of demand, there are three other commonly used elasticities. The *income elasticity of demand* is the percentage change in quantity demanded divided by the percentage change in income that causes it. The *cross-price elasticity of demand* is the percentage change in the quantity demanded of one good divided by the percentage change in the price of some other good. For income and price elasticities, the sign can be either positive or negative.

Finally, the *price elasticity of supply* is the percentage change in quantity supplied divided by the percentage change in price.

PROBLEM SET

Answers to even-numbered Questions and Problems can be found on the text website at www.cengage.com/economics/hall.

1. Using Figure 6, calculate the price elasticity of demand when gasoline rises from $2 per gallon to $3 per gallon over each of the following time horizons:
   a. 1 month (from point A to point B)
   b. 6 months (from point A to point C)
   c. 12 months (from point A to point E

2. Table 2 shows a short-run elasticity of demand for cigarettes. The same study suggested that the long-run elasticity of demand for cigarettes ranges from 1.0 to 2.5. Which is larger—short-run or long-run elasticity? Is this what we would expect? What adjustments might smokers be able to make in the long run that they cannot make in the short run that can explain this relationship between short-run and long-run elasticities?

3. Some studies suggest that "tooth extraction" is an inferior good. Which measure of elasticity (price-elasticity of demand, price-elasticity of supply, income elasticity, or cross-price elasticity) would provide evidence to support this claim? What would we look for in this elasticity measure to determine if tooth extraction were inferior?

4. In Table 2, the price elasticity of demand for Kellogg's Corn Flakes includes a wide range of estimates. It turns out one end of the range is observed for low-income households, while the other is observed for high-income households. Identify which end of the range most likely corresponds to which type of household, and explain briefly. [Hint: It has to do with one of the determinants of elasticity.]

5. The demand for bottled water in a small town is as follows:

P (per bottle)	Q_D (bottles per week)
$1.00	500
$1.50	400
$2.00	300
$2.50	200
$3.00	100

   a. Is this a straight-line demand curve? How do you know?
   b. Calculate the price elasticity of demand for bottled water for a price rise from $1.00 to $1.50. Is demand elastic or inelastic for this price change?
   c. Calculate the price elasticity of demand for a price rise from $2.50 to $3.00. Is demand elastic or inelastic for this price change?
   d. According to the chapter, demand should become less and less elastic as we move downward and rightward along a straight-line demand curve. Use your answers in *b.* and *c.* to confirm this relationship.
   e. Create another column for total revenue on bottled water at each price.
   f. According to the chapter, a rise in price should *increase* total revenue on bottled water when demand is inelastic, and *decrease* total revenue when demand is elastic. Use your answers in *b.* and *c.*, and the new total revenue column you created, to confirm this.

6. Sketch a demand curve that is unit elastic for a price change between $9 and $11. Assume that the quantity demanded is 110 when price is $9. You'll have to determine the quantity demanded when price is $11.

7. In the chapter, we calculated the likely long-run change in revenue if New York City were to raise mass transit fares from $2.00 to $2.50. Use the same procedure to calculate the *short-run* impact of this fare hike. This time, use an elasticity of 0.35, which is in the middle of the short run estimates in Table 3.

8. In the chapter, using an estimate of the short-run elasticity of demand for oil (0.06), we found that a crisis that eliminated half of Persian Gulf oil supplies would cause the world price to rise from $60 to about $125 per barrel in the short run. In the long run, the elasticity of demand for oil is considerably greater. Using the long-run elasticity value for the U.S. in Table 3, and a starting price of $60 per barrel, forecast the new long-run price after a crisis that wipes out half of Persian Gulf oil for a prolonged period.

9. The demand for rosebushes in a market is as follows:

Price (per rosebush)	Quantity Demanded (rosebushes per week)
$10	230
$11	150
$12	90
$13	40

   a. Is this a straight-line demand curve? How do you know?
   b. Calculate the price elasticity of demand for roses for a price increase from $10 to $11. Is demand elastic or inelastic for this price change?
   c. Calculate the price elasticity of demand for roses for a price increase from $12 to $13. Is demand elastic or inelastic for this price change?

10. Refer to Table 2 in this chapter and answer the following questions:
    a. Which is more elastically demanded: cigarettes or pork? Does this make sense to you? Explain briefly.
    b. If the price of milk rises by 5 percent, what will happen to the quantity demanded? (Be specific.)
11. Once again, refer to Table 2 in this chapter and answer the following questions:
    a. Is the demand for recreation more or less elastic than the demand for clothing?
    b. If 10,000 two-liter bottles of Pepsi are currently being demanded in your community each month, and the price increases from $1.90 to $2.10 per bottle, what will happen to quantity demanded? Be specific.
    c. By how much would the price of ground beef have to increase (in percentage terms) in order to reduce quantity demanded by 5 percent?
12. Three Guys Named Al, a moving company, is contemplating a price hike. Currently, they charge $30 per hour, but Al thinks they could get $40. Al disagrees, saying it will hurt the business. Al, the brains of the outfit, has calculated the price elasticity of demand for their moving services in the range from $30 to $40 and found it to be 0.5.
    a. Should they do as Al suggests and raise the price? Why or why not?
    b. Currently, Three Guys is the only moving company in town. Al reads in the paper that several new movers are planning to set up shop there within the next year. Twelve months from now, is the demand for Three Guys' services likely to be more elastic, less elastic, or the same? Why?

More Challenging

13. In February, 2003, Germany's patent office proposed a solution to reimburse copyright holders for illegal digital file sharing: charging personal computer manufacturers a fee of $13 per computer that would go into a special fund to reimburse the copyright holders. Two computer makers—Fujitsu-Siemens and Hewlett-Packard—claimed that imposing the fee would do great injury to them because they would be *unable to pass any of the fee onto consumers*. Under what assumptions about the demand curve would the computer-makers' claim be true? Is this assumption realistic?

CHAPTER 6

Consumer Choice

You are constantly making economic decisions. Some of them are rather trivial. (Have coffee at Starbucks or more cheaply at home?) Others can have a profound impact on your life. (Live with your parents a while longer or get your own place?) The economic nature of all these decisions is rather obvious, since they all involve *spending*.

But in other cases, the economic nature of your decisions may be less obvious. Did you get up early today in order to get things done, or did you sleep in? At this very moment, what have you decided *not* to do in order to make time to read this chapter? These are economic choices, too, because they require you to allocate a scarce resource—your *time*—among different alternatives.

To understand the economic choices that individuals make, we must know what they are trying to achieve (their goals) and the limitations they face in achieving them (their constraints).

Of course, we are all different from one another . . . when it comes to *specific* goals and *specific* constraints. But at the highest level of generality, we are all very much alike. All of us, for example, would like to maximize our overall level of *satisfaction*. And all of us, as we attempt to satisfy our desires, come up against the same types of constraints: too little income or wealth to buy everything we might enjoy, and too little time to enjoy it all.

We'll start our analysis of individual choice with constraints, and then move on to goals. In most of the chapter, we will focus on choices about *spending:* how people decide what to buy. This is why the theory of individual decision making is often called "consumer theory." Later, in the Using the Theory section, we'll broaden our analysis to include decisions involving scarce *time*.

The Budget Constraint

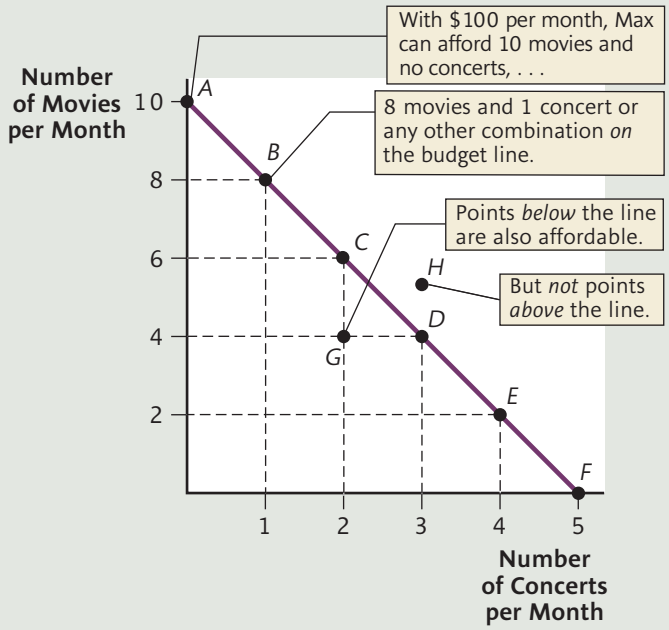
We all must face two facts of economic life: (1) We have to pay for the goods and services we buy, and (2) we have limited funds to spend. These two facts are summarized by our *budget constraint:*

A consumer's ***budget constraint*** *identifies which combinations of goods and services the consumer can afford with a limited budget, at given prices.*

Budget constraint The different combinations of goods a consumer can afford with a limited budget, at given prices.

Consider Max, a devoted fan of both movies and the local music scene, who has a total entertainment budget of $100 each month. The price of a movie is $10, while hearing a rock concert at his favorite local club costs him $20. If Max were to spend

FIGURE 1 The Budget Constraint



Max's Consumption Possibilities with Income of $100

	Concerts at $20 each		Movies at $10 each	
	Quantity	Total Expenditure on Concerts	Quantity	Total Expenditure on Movies
A	0	$ 0	10	$100
B	1	$ 20	8	$ 80
C	2	$ 40	6	$ 60
D	3	$ 60	4	$ 40
E	4	$ 80	2	$ 20
F	5	$100	0	$ 0

all of his $100 budget on concerts at $20 each, he could see at most five each month. If he were to spend it all on movies at $10 each, he could see 10 of them.

But Max could also choose to spend *part* of his budget on concerts and *part* on movies. In this case, for each number of concerts, there is some *maximum* number of movies that he could see. For example, if he goes to one concert per month, it will cost him $20 of his $100 budget, leaving $80 available for movies. Thus, if Max were to choose one concert, the *maximum* number of films he could choose would be $80/$10 = 8.

Figure 1 lists, for each number of concerts, the maximum number of movies that Max could see. Each choice in the table is affordable for Max, since each will cost him exactly $100. Combination *A,* at one extreme, represents no concerts and 10 movies. Combination *F,* the other extreme, represents 5 concerts and no movies. In each of the combinations between *A* and *F,* Max attends both concerts and movies.

The graph in Figure 1 plots the number of movies along the vertical axis and the number of concerts along the horizontal. Each of the points *A* through *F* corresponds to one of the combinations in the table. If we connect all of these points with

Budget line The graphical representation of a budget constraint, showing the maximum affordable quantity of one good for given amounts of another good.

a straight line, we have a graphical representation of Max's budget constraint, which we call Max's **budget line.**

Note that any point below or to the left of the budget line is affordable. For example, two concerts and four movies—indicated by point *G*—would cost only \$30 + \$40 = \$80. Max could certainly afford this combination. On the other hand, he *cannot* afford any combination *above* and to the right of this line. Point *H,* representing 3 concerts and 5 movies, would cost \$60 + \$50 = \$110, which is beyond Max's budget. The budget line therefore serves as a *border* between those combinations that are affordable and those that are not.

Let's look at Max's budget line more closely. The *vertical intercept* is 10, the number of movies Max could see if he attended zero concerts. Starting at the vertical intercept (point *A*), notice that each time Max increases one unit along the horizontal axis (attends one more concert), he must decrease 2 units along the vertical (see three fewer movies). Thus, the slope of the budget line is equal to −2. The slope tells us Max's *opportunity cost* of one more concert. That is, the opportunity cost of one more concert is 2 movies foregone.

Relative price The price of one good relative to the price of another.

There is an important relationship between the *prices* of two goods and the opportunity cost of having more of one or the other. If we divide one money price by another money price, we get what is called a **relative price**, the price of one good *relative* to the other. Let's use the symbol P_c for the price of a concert and P_m for the price of a movie. Since P_c = \$20 and P_m = \$10, the *relative price of a concert* is the ratio P_c/P_m = \$20/\$10 = 2. Notice that this same number, 2, is the opportunity cost of another concert in terms of movies; and, except for the minus sign, it is also the slope of the budget line. That is, *the relative price of a concert, the opportunity cost of another concert, and the slope of the budget line* have the same absolute value. This is one example of a general relationship:

> *The slope of the budget line indicates the spending tradeoff between one good and another—the amount of one good that must be sacrificed in order to buy more of another good. If* P_y *is the price of the good on the vertical axis and* P_x *is the price of the good on the horizontal axis, then the slope of the budget line is* $-P_x/P_y$.

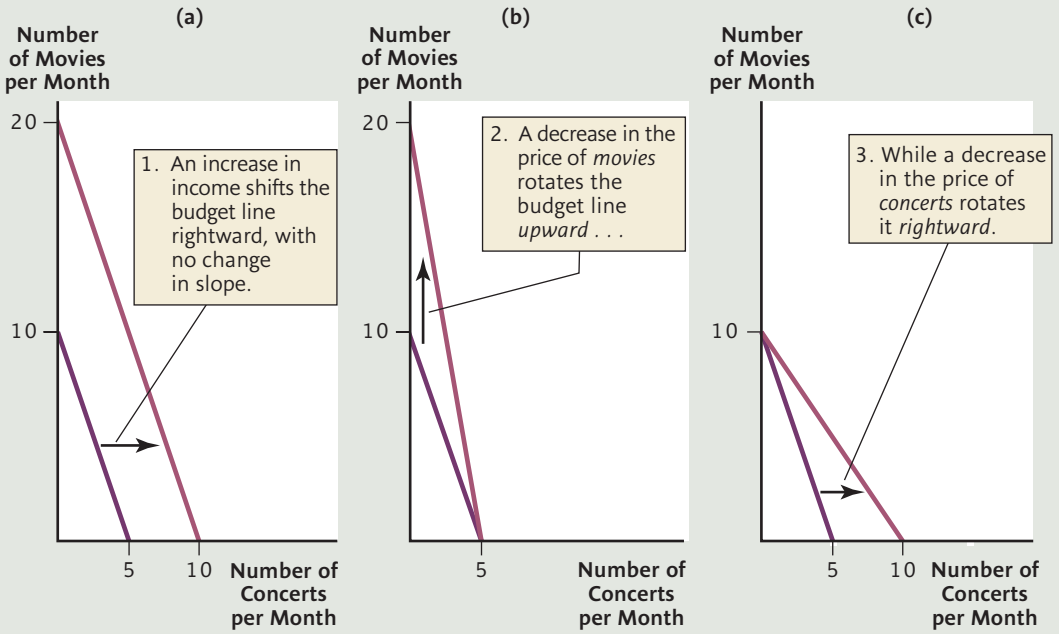
Changes in the Budget Line

To draw the budget line in Figure 1, we have assumed given prices for movies and concerts, and a given income that Max can spend on them. These "givens"—the prices of the goods and the consumer's income—are always *assumed constant* as we move along a budget line; if any one of them changes, the budget line will change as well. Let's see how.

Changes in Income

If Max's available income increases from \$100 to \$200 per month, then he can afford to see more movies, more concerts, or more of both, as shown by the change in his budget line in Figure 2(a). If Max were to devote *all* of his income to movies, he could now see 20 of them each month, instead of the 10 he was able to see before. Devoting his entire income to concerts would enable him to attend 10, rather than 5. Moreover, for any number of concerts, he will be able to see more movies than before. For example, choosing 2 concerts would allow Max to see only 6 movies. Now, with a larger budget of \$200, he can have 2 concerts and 16 movies.

FIGURE 2 Changes in the Budget Line



Notice that the old and new budget lines in Figure 2(a) are parallel; that is, they have the same slope of −2. We have changed Max's income but *not* the prices. Since the ratio P_c/P_m has not changed, the spending tradeoff between movies and concerts remains the same. Thus,

> *an increase in income will shift the budget line upward (and rightward). A decrease in income will shift the budget line downward (and leftward). These shifts are parallel: Changes in income do not affect the budget line's slope.*

dangerous curves

The Budget Line's Slope It's tempting to think that the slope of the budget line should be $-P_y/P_x$, with the price of the vertical axis good *y* in the numerator. But notice that the slope is the other way around, $-P_x/P_y$, with P_x in the numerator. That's because we're expressing the slope in terms of prices (which are *not* on the axes), rather than quantities. We've given some intuition for this way of expressing the slope in our example with Max, but if you'd like a formal proof, see the footnote.[1]

Changes in Price

Now let's go back to Max's original budget of $100 and explore what happens to the budget line when a price changes. Suppose the price of a movie falls from $10 to $5. The graph in Figure 2(b) shows Max's old and new budget lines. When the

[1] To prove that the slope of the budget line is $-P_x/P_y$, let Q_x and Q_y represent the quantities of good *x* and good *y*, respectively. Then the total amount a consumer spends on good *y* is P_yQ_y, and the total amount spent on good *x* is P_xQ_x. All along the budget line, the total amount spent on these two goods is equal to the total budget (B), so it must be that $P_xQ_x + P_yQ_y = B$. This is the equation for the budget line. Since the budget line graph has good *y* on the vertical axis, we can solve this equation for Q_y by rearranging terms: $Q_y = (B - P_xQ_x)/P_y$, which can be rewritten as $Q_y = (B/P_y) + (-P_x/P_y)\,Q_x$. This is the equation for a straight line. The first term in parentheses (B/P_y) is the vertical intercept, and the second term in parentheses ($-P_x/P_y$) is the slope. (See the mathematical appendix to Chapter 1 if you need to review slopes and intercepts.)

price of a movie falls, the budget line rotates outward; that is, the vertical intercept moves higher. The reason is this: When a movie costs $10, Max could spend his entire $100 on them and see 10; now that they cost $5, he can see a maximum of 20. The horizontal intercept—representing how many concerts Max could see with his entire income—doesn't change at all, since there has been no change in the price of a concert. Notice that the new budget line is also *steeper* than the original one, with slope equal to $-P_c/P_m = -\$20/\$5 = -4$. Now, with movies costing $5, the trade-off between movies and concerts is 4 to 1, instead of 2 to 1.

Panel (c) of Figure 2 illustrates another price change. This time, it's a fall in the price of a *concert* from $20 to $10. Once again, the budget line rotates, but now it is the horizontal intercept (concerts) that changes and the vertical intercept (movies) that remains fixed.

We could draw similar diagrams illustrating a *rise* in the price of a movie or a concert, but you should try to do this on your own. In each case, one of the budget line's intercepts will change, as well as its slope:

> *When the price of a good changes, the budget line rotates: Both its slope and one of its intercepts will change.*

The budget constraint, as illustrated by the budget line, is one side of the story of consumer choice. It indicates the tradeoff consumers *are able to* make between one good and another. But just as important is the tradeoff that consumers *want to* make between one good and another, and this depends on consumers' *preferences,* the subject of the next section.

Preferences

How can we possibly speak systematically about people's preferences? After all, people are different. They like different things. American teens delight in having Coke with dinner, while the very idea makes a French person shudder. What would satisfy a Buddhist monk would hardly satisfy the typical American.

And even among "typical Americans," there is little consensus about tastes. Some read Jane Austen, while others choose John Grisham. Some like to spend their vacations traveling, whereas others would prefer to stay home and sleep in every day. Even those who like Häagen-Dazs ice cream can't agree on which is the best flavor—the company notices consistent, regional differences in consumption. In Los Angeles, chocolate chip is the clear favorite, while on most of the East Coast, it's butter pecan—except in New York City, where coffee wins hands down.

In spite of such wide differences in preferences, we can find some important common denominators—things that seem to be true for a wide variety of people. In our theory of consumer choice, we will focus on these common denominators.

Rationality

One common denominator—and a critical assumption behind consumer theory—is that people *have* preferences. More specifically, we assume that you can look at two alternatives and state either that you prefer one to the other or that you are entirely indifferent between the two—you value them equally.

Another common denominator is that preferences are *logically consistent,* or *transitive*. If, for example, you prefer a sports car to a jeep, and a jeep to a motorcycle, then we assume that you will also prefer a sports car to a motorcycle. When a consumer can make choices, and is logically consistent, we say that she has **rational preferences.**

Rational preferences Preferences that satisfy two conditions: (1) Any two alternatives can be compared, and one is preferred or else the two are valued equally, and (2) the comparisons are logically consistent or transitive.

Notice that rationality is a matter of how you make your choices, and not what choices you make. You can be rational and like apples better than oranges, or oranges better than apples. You can be rational even if you like chocolate-covered anchovies! What matters is that you make logically consistent choices, and most of us usually do.

More Is Better

Another feature of preferences that virtually all of us share is this: We generally feel that *more is better*. Specifically, if we get more of some good or service, and nothing else is taken away from us, we will generally feel better off.

This condition seems to be satisfied for the vast majority of goods we all consume. Of course, there are exceptions. If you hate eggplant, then the more of it you have, the worse off you are. Similarly, a dieter who says, "Don't bring any ice cream into the house. I don't want to be tempted," also violates the assumption. The model of consumer choice in this chapter is designed for preferences that satisfy the "more is better" condition, and it would have to be modified to take account of exceptions like these.

So far, our characterization of consumer preferences has been rather minimal. We've assumed only that consumers are rational and that they prefer more rather than less of every good we're considering. But even this limited information allows us to say the following:

> *The consumer will always choose a point on the budget line, rather than a point below it.*

To see why this is so, look again at Figure 1. Max would never choose point *G*, representing 2 concerts and 6 movies, since there are affordable points—on the budget line—that we know make him better off. For example, point *C* has the same number of concerts, but more movies, while point *D* has the same number of movies, but more concerts. "More is better" tells us that Max will prefer *C* or *D* to *G*, so we know *G* won't be chosen. Indeed, if we look at any point below the budget line, we can always find at least one point on the budget line that is preferred, as long as more is better.

Knowing that Max will always choose a point *on* his budget line is a start. But how does he find the *best* point on the line—the one that gives him the highest level of satisfaction?

This is where your *instructor's* preferences come in. There are two theories of consumer decision making, and they share much in common. First, both assume that preferences are rational. Second, both assume that the consumer would be better off with more of any good we're considering. This means the consumer will always choose a combination of goods *on*, rather than below, his budget line. Finally, both theories come to the same general conclusions about consumer behavior. However, to *arrive* at those conclusions, each theory takes a different road.

The next section presents the "Marginal Utility" approach to consumer decision making. If, however, your instructor prefers the "Indifference Curve" approach, you can skip the next section and go straight to the appendix. Then, come back to the section titled "Income and Substitution Effects," which is where our two roads converge once again.

One warning, though. Both approaches to consumer theory are *models*. They use graphs and calculations to explain how consumers make choices. While the models are logical, they may appear unrealistic to you. And in one sense, they *are* unrealistic: Few consumers in the real world are aware of the techniques we'll discuss, yet they make choices all the time.

Economists don't imagine that, when making choices, households or consumers actually *use* these techniques. Rather, the assumption is that people mostly behave *as if* they use them. Indeed, most of the time, in most markets, household behavior has proven to be consistent with the model of consumer choice. When our goal is to describe and predict how consumers are likely to behave in markets—rather than describe what actually goes on in their minds—our theories of consumer decision making can be very useful.

Consumer Decisions: The Marginal Utility Approach

Economists assume that *any* decision maker—a consumer, the manager of a business firm, or officials in a government agency—tries to make the *best* out of any situation. Marginal utility theory treats consumers as striving to maximize their **utility**—an actual *quantitative* measure of well-being or satisfaction. Anything that makes the consumer better off is assumed to raise his utility. Anything that makes the consumer worse off will decrease his utility.

Utility A quantitative measure of pleasure or satisfaction obtained from consuming goods and services.

Utility and Marginal Utility

Figure 3 provides a graphical view of utility—in this case, the utility of a consumer named Lisa who likes ice cream cones. Look first at panel (a). On the horizontal axis, we'll measure the number of ice cream cones Lisa consumes each week. On the vertical axis, we'll measure the utility she derives from consuming each of them.

If Lisa values ice cream cones, her utility will increase as she acquires more of them, as it does in the figure. There we see that when she has one cone, she enjoys total utility of 30 "utils," and when she has two cones, her total utility grows to 50 utils, and so on. Throughout the figure, the total utility Lisa derives from consuming ice cream cones keeps rising as she gets to consume more and more of them.

But notice something interesting, and important: Although Lisa's utility increases every time she consumes more ice cream, the *additional* utility she derives from each *successive* cone gets smaller and smaller as she gets more cones. We call the *change in utility* derived from consuming an *additional unit* of a good the *marginal utility* of that additional unit.

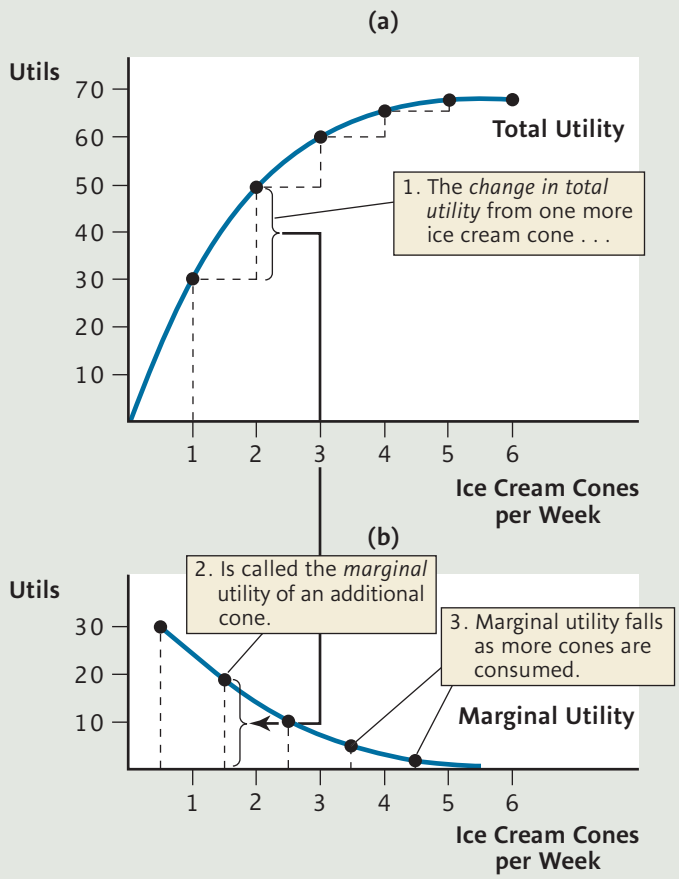
Marginal utility The change in total utility an individual obtains from consuming an additional unit of a good or service.

> ***Marginal utility** is the change in utility an individual enjoys from consuming an additional unit of a good.*

FIGURE 3 Total and Marginal Utility

Lisa's Total and Marginal Utility from Consuming Ice Cream Cones

Number of Cones	Total Utility	Marginal Utility
0	0 utils	
		30 utils
1	30 utils	
		20 utils
2	50 utils	
		10 utils
3	60 utils	
		5 utils
4	65 utils	
		3 utils
5	68 utils	
		0 utils
6	68 utils	



What we've observed about Lisa's utility can be restated this way: As she eats more and more ice cream cones in a given week, her *marginal utility* from another cone declines. We call this the **law of diminishing marginal utility**, which the great economist Alfred Marshall (1842–1924) defined this way:

> *The marginal utility of a thing to anyone diminishes with every increase in the amount of it he already has.*[2]

Law of diminishing marginal utility As consumption of a good or service increases, marginal utility decreases.

According to the law of diminishing marginal utility, when you consume your first unit of some good, like an ice cream cone, you derive some amount of utility. When you get your second cone that week, you enjoy greater satisfaction than when you only had one, but the extra satisfaction you derive from the second is likely to

[2] *Principles of Economics*, Book III, Ch. III, Appendix notes 1 & 2. Macmillan & Co., 1930. The term "marginal" is one that you'll encounter often in economics. The margin of a sheet of notebook paper is the area on the edge, just *beyond* the writing area. By analogy, a *marginal value* in economics measures what happens when we go a little bit *beyond* where we are now, by adding one more unit of something.

be smaller than the satisfaction you derived from the first. Adding the third cone to your weekly consumption will no doubt increase your utility further, but again the marginal utility you derive from that third cone is likely to be less than the marginal utility you derived from the second.

Figure 3 will again help us see what's going on. The table summarizes the information in the total utility graph. The first two columns show, respectively, the quantity of cones Lisa consumes each week and the total utility she receives each week from consuming them. The third column shows the *marginal* utility she receives from each successive cone she consumes per week. As you can see in the table, Lisa's total utility keeps increasing (marginal utility is always positive) as she consumes more cones (up to five per week), but the rate at which total utility increases gets smaller and smaller (her marginal utility diminishes) as her consumption increases.

Marginal utility is shown in panel (b) of Figure 3. Because marginal utility is the change in utility caused by a change in consumption from one level to another, we plot each marginal utility entry between the old and new consumption levels.

Notice the close relationship between the graph of total utility in panel (a) and the corresponding graph of marginal utility in panel (b). For every one-unit increment in Lisa's ice cream consumption her marginal utility is equal to the change in her total utility. Diminishing marginal utility is seen in both panels of the figure: in panel (b), by the downward sloping marginal utility curve, and in panel (a), by the positive but decreasing slope (flattening out) of the total utility curve.

One last thing about Figure 3: Because marginal utility diminishes for Lisa, by the time she has consumed a total of five cones per week, the marginal utility she derives from an additional cone has fallen all the way to zero. At this point, she is fully satiated with ice cream and gets no extra satisfaction or utility from eating any more of it in a typical week. Once this satiation point is reached, even if ice cream were free, Lisa would turn it down ("Yechhh! Not more ice cream!!"). But remember from our earlier discussion that one of the assumptions we always make about preferences is that people prefer *more* rather than less of any good we're considering. So when we use marginal utility theory, we assume that marginal utility for every good is positive. For Lisa, it would mean she hasn't yet reached five ice cream cones per week.

Combining the Budget Constraint and Preferences

The marginal utility someone gets from consuming more of a good tells us about his *preferences*. His budget constraint, by contrast, tells us only which combinations of goods he can *afford*. If we combine information about preferences (marginal utility values) with information about what is affordable (the budget constraint), we can develop a useful rule to guide us to an individual's utility-maximizing choice.

To develop this rule, let's go back to Max and his choice between movies and concerts. Table 1 shows some utility numbers for movies and concerts. Notice that, as was the case with Lisa and her ice cream cones, Max's *total* utility rises with each concert or movie he consumes. But his *marginal* utility falls as more of either good is added. For example, a second concert per month adds 800 utils to Max's total utility. But the third concert adds only 600.

TABLE 1

Total and Marginal Utility of Concerts and Movies

Number of Concerts per Month	Total Utility	Marginal Utility	Number of Movies per Month	Total Utility	Marginal Utility
0	0		0	0	
		1,000			600
1	1,000		1	600	
		800			450
2	1,800		2	1,050	
		600			350
3	2,400		3	1,400	
		400			300
4	2,800		4	1,700	
		250			250
5	3,050		5	1,950	
		200			200
6	3,250		6	2,150	
					150
			7	2,300	
					100
			8	2,400	
					75
			9	2,475	
					60
			10	2,535	

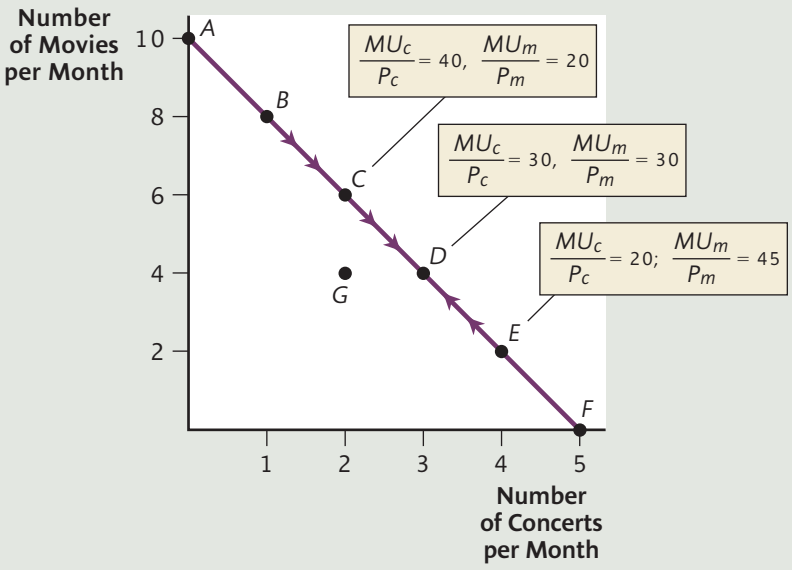
Remember that we want to find the best among the affordable combinations of these two goods. That means we'll need to consider the combinations of the two goods that are on Max's budget line, as shown in Figure 4 (reproduced from Figure 1). Figure 4 also shows some of the utility information in Table 1, but rearranged to more easily see Max's available choices. There's a lot going on in that table, so let's step through it carefully.

In column (1), the rows labeled A, B, C, etc. correspond to possible combinations on Max's budget line. For example, the row labeled C corresponds to point C on the budget line: 2 concerts and 6 movies per month. (The unlabeled rows in between would require Max to see and pay for a half of a concert, which we assume he can't do.)

Next, look at columns (2) and (5). Notice that the number of concerts runs from 0 to 5 as we travel down the rows. But the number of movies runs in the other direction, from 10 to 0. That's because—as we move along the budget line—attending more concerts means seeing fewer movies.

Now look at columns (3) and (6), which show the marginal utility from the "last" concert or movie. For example, in the row labeled C, the "last" concert Max

FIGURE 4 **Consumer Decision Making**



Budget = $100 per month

CONCERTS at $20 each				MOVIES at $10 each		
(1) Point on Budget Line	(2) Number of Concerts per Month	(3) Marginal Utility from Last Concert (MU_c)	(4) Marginal Utility per Dollar Spent on Last Concert $\left(\frac{MU_c}{P_c}\right)$	(5) Number of Movies per Month	(6) Marginal Utility from Last Movie (MU_m)	(7) Marginal Utility per Dollar Spent on Last Movie $\left(\frac{MU_m}{P_m}\right)$
A	0			10	60	6
				9	75	7.5
B	1	1000	50	8	100	10
				7	150	15
C	2	800	40	6	200	20
				5	250	25
D	**3**	**600**	**30**	**4**	**300**	**30**
				3	350	35
E	4	400	20	2	450	45
				1	600	60
F	5	250	12.5	0		

The budget line shows the maximum number of movies Max could attend for each number of concerts he attends. He would never choose an interior point like G *because there are affordable points—on the line—that make him better off. Max will choose the point on the budget line at which the marginal utilities per dollar spent on movies and concerts are equal. From the table, this occurs at point* D.

sees is the second one, with marginal utility of 800 utils. In that same row, Max sees 6 movies, and the marginal utility from the last (sixth) movie is 200. These marginal utility numbers come from Table 1.

Now, look at column (4), which shows something new: the marginal utility *per dollar* spent on concerts. To get these numbers, we divide the marginal utility of the last concert (MU_c) by the price of a concert, giving us MU_c/P_c. This tells us the gain in utility Max gets *for each dollar he spends* on the last concert. For example, at point *C,* Max gains 800 utils from his second concert during the month, so his marginal utility *per dollar* spent on that concert is 800 utils/\$20 = 40 utils per dollar. Marginal utility per dollar, like marginal utility itself, declines as Max attends more concerts. After all, marginal utility itself decreases, and the price of a concert isn't changing.

The last column gives us similar information for movies: the marginal utility per dollar spent on the last movie (MU_m/P_m). As we travel *up* this column, Max attends more movies, and both marginal utility and marginal utility per dollar decline—once again, consistent with the law of diminishing marginal utility.

Now, Max's goal is to find the affordable combination of movies and concerts—the point on his budget line—that gives him the highest possible utility. As you are about to see, this will be the point at which *the marginal utility per dollar is the same for both goods*.

To see why, imagine that Max is searching along his budget line for the utility-maximizing point, and he's currently considering point *B,* which represents 1 concert and 8 movies. Is he maximizing his utility? Let's see. Comparing the fourth and seventh entries in row *B* of the table, we see that Max's marginal utility per dollar spent on concerts is 50 utils, while his marginal utility per dollar spent on movies is only 10 utils. Since he gains more additional utility from each dollar spent on concerts than from each dollar spent on movies, he will have a net gain in utility if he shifts some of his dollars from movies to concerts. To do this, he must travel farther down his budget line.

Next suppose that, after shifting his spending from movies to concerts, Max arrives at point *C* on his budget line. What should he do then? At point *C,* Max's *MU* per dollar spent on concerts is 40 utils, while his *MU* per dollar spent on movies is 20 utils. Once again, he would gain utility by shifting from movies to concerts, traveling down his budget line once again.

Now suppose that Max arrives at point *D*. At this point, the *MU* per dollar spent on both movies and concerts is the same: 30 utils. There is no further gain from shifting spending from movies to concerts. At point *D,* Max has exploited all opportunities to make himself better off by moving down the budget line. He has maximized his utility.

But wait . . . what if Max had started at a point on his budget line *below* point *D?* Would he still end up at the same place? Yes, he would. Suppose Max finds himself at point *E,* with 4 concerts and 3 movies. Here, marginal utilities per dollar are 20 utils for concerts and 45 utils for movies. Now, Max could make himself better off by shifting spending away from concerts and toward movies. He will travel *up* the budget line, once again arriving at point *D,* where no further move will improve his well-being.

As you can see, it doesn't matter whether Max begins at a point on his budget line that's above point *D* or below it. Either way, if he keeps shifting spending toward the good with greater marginal utility per dollar, he will always end up at point *D*. And because marginal utility per dollar is the same for both goods at point *D,* there is nothing to gain by shifting spending in either direction.

What is true for Max and his choice between movies and concerts is true for *any* consumer and *any* two goods. We can generalize our result this way: For any two goods x and y, with prices P_x and P_y, whenever $MU_x/P_x > MU_y/P_y$, a consumer is made better off shifting spending away from y and toward x. When $MU_y/P_y > MU_x/P_x$, a consumer is made better off by shifting spending away from x and toward y. This leads us to an important conclusion:

A utility-maximizing consumer will choose the point on the budget line where marginal utility per dollar is the same for both goods ($MU_x/P_x = MU_y/P_y$). At that point, there is no further gain from reallocating expenditures in either direction.

We can generalize this result to the more realistic situation of choosing combinations of more than two goods: different types of food, clothing, entertainment, transportation, and so on. No matter how many goods there are to choose from, when the consumer is doing as well as possible, it must be true that $MU_x/P_x = MU_y/P_y$ for *any* pair of goods x and y. If this condition is *not* satisfied, the consumer will be better off consuming more of one and less of the other good in the pair.[3]

dangerous curves

Don't Confuse MU with MU per Dollar. It's tempting to think that a consumer should shift spending to the good with greater marginal utility, rather than greater marginal utility *per dollar*. But a simple thought experiment can convince you why this is wrong. Imagine that you like to ski and you like going out for dinner. Currently, your marginal utility for one more skiing trip is 2,000 utils, and your marginal utility for an additional dinner is 1,000 utils. Should you shift your spending from dining out to skiing? It might seem so, since skiing has the higher marginal utility.

But what if skiing costs $200 per trip, while a dinner out costs only $20? Then, while it's true that another ski trip adds twice as much utility as another dinner out, it's also true that *skiing costs 10 times as much*. You would have to sacrifice 10 dinners out for 1 ski trip, and that would make you *worse* off.

Instead, you should shift your spending in the other direction: from skiing to dining out. Money spent on additional ski trips will give you 2,000 utils/$200 = 10 utils per dollar, while money spent on additional dinners will give you 1,000 utils/$20 = 50 utils per dollar. Dining out clearly gives you "more bang for the buck" than skiing, because its marginal utility *per dollar* is greater.

What Happens When Things Change?

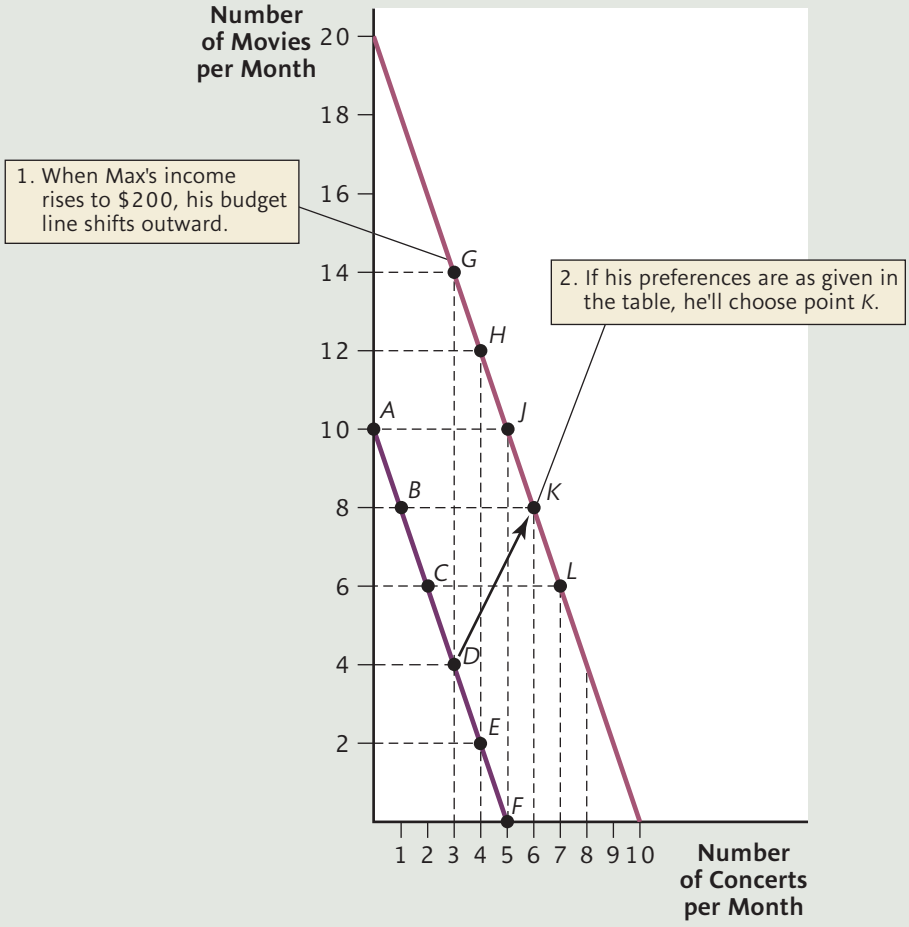
If every one of our decisions had to be made only once, life would be much easier. But that's not how life is. Just when you think you've figured out what to do, things change. In a market economy, as you've learned, prices can change for any number of reasons. (See Chapter 3.) A consumer's income can change as well. These changes cause us to rethink our spending decisions: What maximized utility before the change is unlikely to maximize it afterward.

Changes in Income

Figure 5 illustrates how an increase in income might affect Max's choice between movies and concerts. As before, we assume that movies cost $10 each, that concerts cost $20 each, and that these prices are not changing. Initially, Max has $100 in income to spend on the two goods, so his budget line is the line from point A to point F. As we've already seen, under these conditions, Max would choose point D (three concerts and four movies) to maximize utility.

[3] There is one exception to this statement: Sometimes the optimal choice is to buy *none* of some good. For example, suppose that $MU_y/P_y > MU_x/P_x$ no matter how small a quantity of good x a person consumes. Then the consumer should always reduce consumption of good x further, until its quantity is zero. Economists call this a "corner solution" because when there are only two goods being considered, the individual will locate at one of the end points of the budget line in a corner of the diagram.

FIGURE 5 **Effects of an Increase in Income**



Budget = $200 per month

	CONCERTS at $20 each			MOVIES at $10 each		
(1) Point on *New* Budget Line	(2) Number of Concerts per Month	(3) Marginal Utility from Last Concert (MU_c)	(4) Marginal Utility per Dollar Spent on Last Concert $\left(\frac{MU_c}{P_c}\right)$	(5) Number of Movies per Month	(6) Marginal Utility from Last Movie (MU_m)	(7) Marginal Utility per Dollar Spent on Last Movie $\left(\frac{MU_m}{P_m}\right)$
G	3	600	30	14	38	3.8
				13	40	4
H	4	400	20	12	45	4.5
				11	50	5
J	5	250	12.5	10	60	6
				9	75	7.5
K	**6**	**200**	**10**	**8**	**100**	**10**
				7	150	15
L	7	180	9	6	200	20

Now suppose Max's monthly income (or at least the part he budgets for entertainment) increases to $200. Then his budget line will shift upward and outward in the figure. How will he respond? As always, he will search along his budget line until he finds the point where the marginal utility per dollar spent on both goods is the same. To help us find this point, Figure 5 includes some additional marginal utility values for combinations that weren't affordable before, but are now. For example, the *sixth* concert—which Max couldn't afford in Figure 4—is now assumed to have a marginal utility of 200 utils.

With the preferences described by these marginal utility numbers, Max will search along his budget line for the best choice. This will lead him directly to point *K*, enjoying 6 concerts and 8 movies per month. For this choice, *MU/P* is 10 utils per dollar for both goods, so total utility can't be increased any further by shifting dollars from one good to the other.

Now let's take a step back from these calculations and look at the figure itself. We see that an increase in income has changed Max's best choice from point *D* on the old budget constraint to point *K* on the new one. In moving from *D* to *K*, Max chooses to buy more concerts (6 rather than 3) and more movies (8 rather than 6). As discussed in Chapter 3, if an increase in income (with prices held constant) increases the quantity of a good demanded, the good is *normal*. For Max, with the marginal utility values we've assumed in Figure 5, both concerts and movies would be normal goods.

But this is not the only possible result. If Max's preferences (and marginal utility values) had been different than those in Figure 5, he might have chosen more of one good and *less* of the other.

These possibilities are illustrated in Figure 6. Our previous result—based on the marginal utility numbers in Figure 5—had Max moving from point *D* on his initial budget line to point *K* on the new, higher one when his income rose. But with different preferences (different marginal utility values), his marginal utilities per dollar might have been equal at a point like *K′*, with nine concerts and two movies. In this case, the increase in income would cause Max's consumption of concerts to increase (from 3 to 9), but his consumption of movies to *fall* (from 4 to 2). If so, movies would be an *inferior good* for Max—one for which demand decreases when income increases—while concerts would be a *normal* good.

Finally, let's consider another possible outcome for Max: point *K″*. At this point, he attends more movies and fewer concerts compared to point *D*. If point *K″* is where Max's marginal utilities per dollar are equal after the increase in income, then *concerts* would be the inferior good, and movies would be normal.[4]

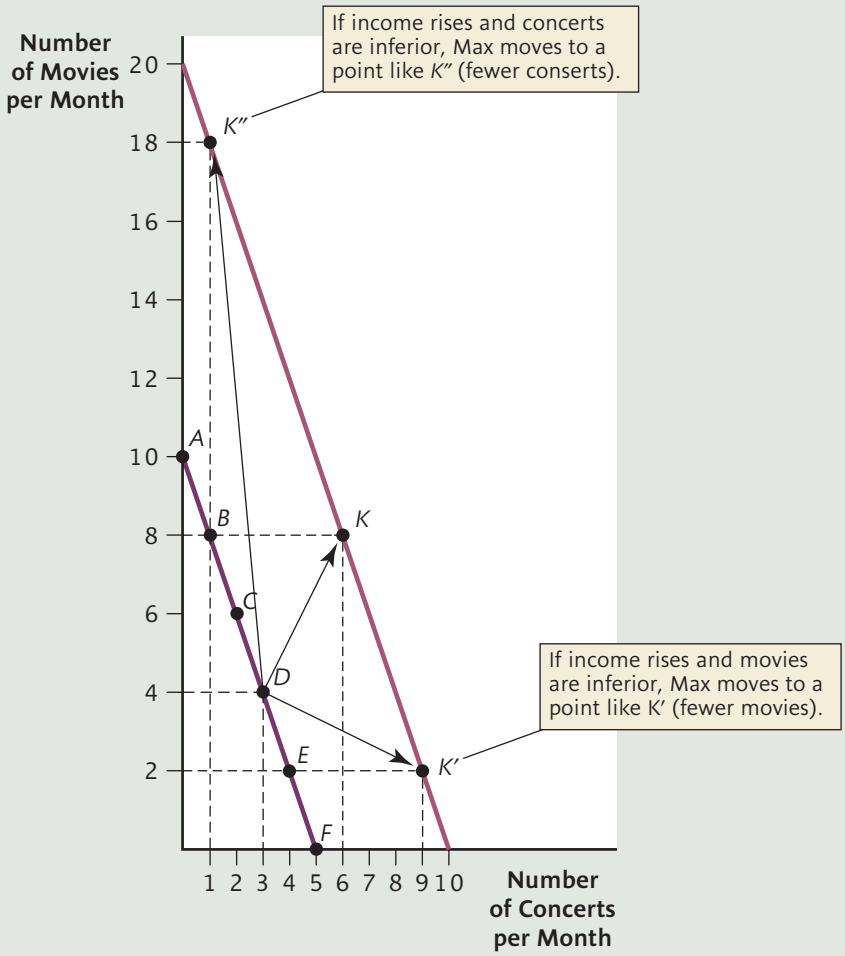
A rise in income—with no change in prices—leads to a new quantity demanded for each good. Whether a particular good is normal (quantity demanded increases) or inferior (quantity demanded decreases) depends on the individual's preferences, as represented by the marginal utilities for each good, at each point along his budget line.

dangerous curves

The Special Meaning of "Inferior" in Economics It's tempting to think that *inferior* goods are of lower quality than *normal* goods. But economists don't define normal or inferior based on the intrinsic properties of a good, but rather by the choices people make when their incomes increase. For example, Max may think that both movies and concerts are high-quality goods. When his income is low, he may see movies on most weekends because, being cheaper, they enable him to spread his budget further. But if his income increases, he might switch from movies to concerts on some nights and end up seeing fewer movies. In that case, his *behavior* tells us that movies are an inferior good for him.

[4] If you are wondering how we would use marginal utility tables like those in Figures 4 and 5 when one of the goods is inferior, wait for the end of chapter problems. In Figures 4 and 5, we have made a special assumption: That the marginal utility values for one good do *not* depend on quantities of the *other* good. But this need not be the case, as you'll see in the problems.

FIGURE 6 **Normal and Inferior Goods**



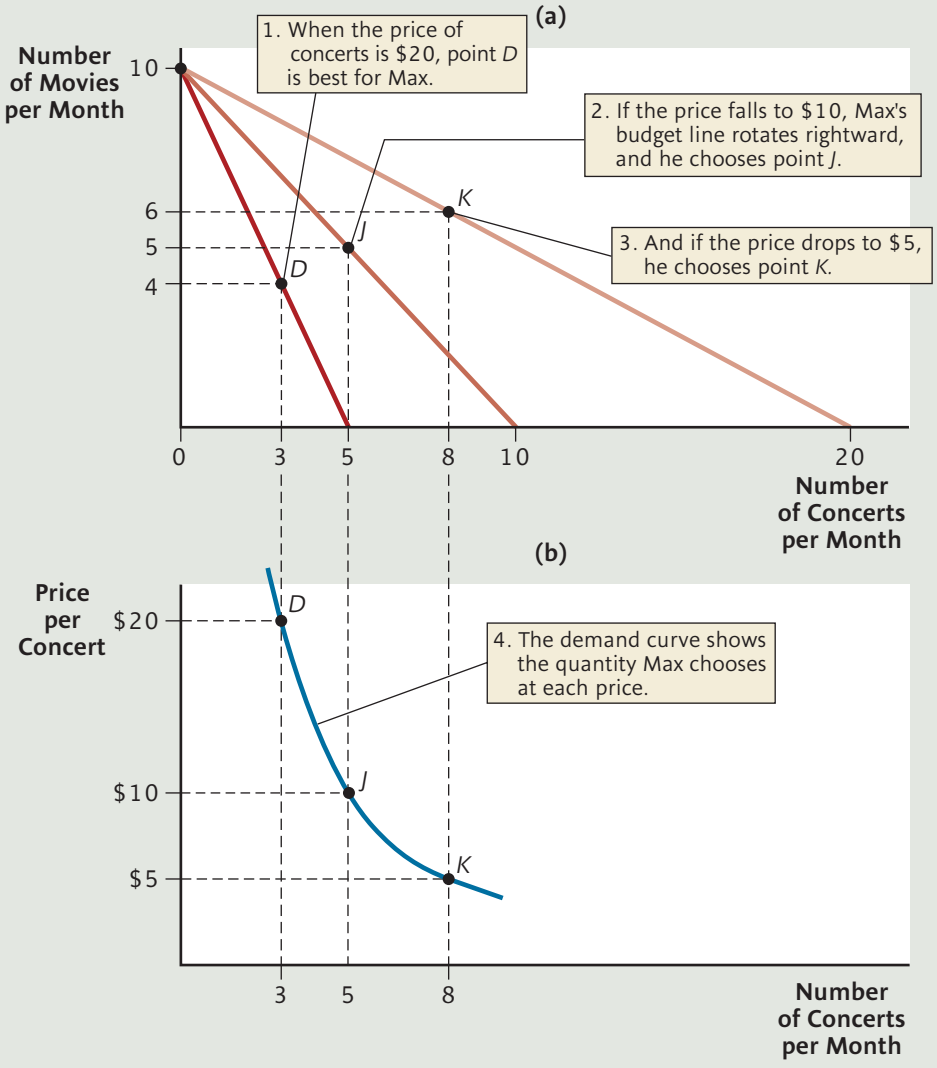
Changes in Price

Let's explore what happens to Max when the price of a concert decreases from $20 to $10, while his income and the price of a movie remain unchanged. The drop in the price of concerts rotates Max's budget line rightward, pivoting around its vertical intercept, as illustrated in the upper panel of Figure 7. What will Max do after his budget line rotates in this way? Again, he will select the combination of movies and concerts on his budget line that makes him as well off as possible. This will be the combination at which the marginal utility per dollar spent on both goods is the same.

Once again, we've taken some of Max's marginal utility values from Figure 4 and added some additional numbers to construct the table in Figure 7. This table extends what we already knew about Max's preferences to cover the new, expanded possibilities.

With the preferences represented by these marginal utility numbers, Max will search along his budget line for the best choice. This will lead him directly to point *J*, where his quantities demanded are 5 concerts and 5 movies. Note that with each concert costing only $10 now, Max can afford this combination. Moreover, it satisfies our utility-maximizing rule: Marginal utility per dollar is 25 for both goods.

FIGURE 7 Deriving the Demand Curve



Max Choices with Budget of $100 per month

	CONCERTS at $10 each			MOVIES at $10 each		
(1) Point on *New* Budget Line	(2) Number of Concerts per Month	(3) Marginal Utility from Last Concert (MU_c)	(4) Marginal Utility per Dollar Spent on Last Concert $\left(\frac{MU_c}{P_c}\right)$	(5) Number of Movies per Month	(6) Marginal Utility from Last Movie (MU_m)	(7) Marginal Utility per Dollar Spent on Last Movie $\left(\frac{MU_m}{P_m}\right)$
	3	600	60	7	150	15
	4	400	40	6	200	20
J	**5**	**250**	**25**	**5**	**250**	**25**
	6	180	18	4	300	30
	7	100	10	3	350	35

What if we dropped the price of concerts again—this time—to $5? Then Max's budget line rotates further rightward, and he will once again find the utility-maximizing point. In the figure, Max is shown choosing point *K*, attending 8 concerts and 6 movies. (The marginal utilities per dollar in the table cannot be used to find point *K*, because the table assumes the price of concerts is $10.)

The Consumer's Demand Curve

You've just seen that each time the price of concerts changes, so does the quantity of concerts Max will want to see. The lower panel of Figure 7 highlights this relationship by plotting the quantity of concerts demanded on the horizontal axis and the price of concerts on the vertical axis. For example, in both the upper and lower panels, point *D* tells us that when the price of concerts is $20, Max will see three of them. When we connect points like *D, J,* and *K* in the lower panel, we get Max's demand curve, which shows *the quantity of a good he demands at each different price*. Notice that Max's demand curve for concerts slopes downward—a fall in the price of concerts increases the quantity demanded—showing that Max's responses to price changes obey the law of demand.

But if Max's preferences—and his marginal utility values—had been different, could his response to a price change have *violated* the law of demand? The answer is yes . . . and no. Yes, it is theoretically possible. (As a challenge, try identifying points on the three budget lines that would give Max an *upward-sloping* demand curve.) But no, it does not seem to happen in practice.

To understand why and to gain other insights, the next section takes a deeper look into the effects of a price change on quantity demanded.

Income and Substitution Effects

Whether you've studied about the marginal utility approach (the previous section) or the indifference curve approach (appendix), you've learned a logical process that leads directly to an individual's demand curve. But the demand curve actually summarizes the impact of *two* separate effects of a price change on quantity demanded. As you are about to see, these two effects sometimes work together, and sometimes oppose each other.

The Substitution Effect

Suppose the price of a good falls. Then it becomes less expensive *relative to* other goods whose prices have not fallen.

For example, when the price of concerts falls, so does its relative price (relative to movies). Max can now get more entertainment from his budget by substituting concerts in place of movies, so he will demand more concerts.

This impact of a price decrease is called a substitution effect: the consumer substitutes *toward* the good whose price has decreased, and away from other goods whose prices have remained unchanged.

> *The* ***substitution effect*** *of a price change arises from a change in the relative price of a good, and it always moves quantity demanded in the opposite direction to the price change. When price decreases, the substitution effect works to increase quantity demanded; when price increases, the substitution effect works to decrease quantity demanded.*

Cheaper cell phone calls, and the substitution effect, have almost completely driven pay phones like this out of the market.

Substitution effect As the price of a good falls, the consumer substitutes that good in place of other goods whose prices have not changed.

The substitution effect is a powerful force in the marketplace. For example, while the price of cell phone calls has fallen in recent years, the price of pay phone calls has remained more or less the same. This fall in the relative price of cell phone calls has caused consumers to substitute toward them and away from using regular pay phones. As a result, pay phones have almost completely disappeared in many areas of the country.

The substitution effect is also important from a theoretical perspective: It is the main factor responsible for the law of demand. Indeed, if the substitution effect were the *only* effect of a price change, the law of demand would be more than a law; it would be a logical necessity. But as we are about to see, a price change has another effect as well.

The Income Effect

In Figure 7 (or Appendix Figure A.5), when the price of concerts drops from $20 to $10, Max's budget line rotates rightward. Max now has a wider range of options than before: He can consume more concerts, more movies, or *more of both*. The price decline of *one* good increases his total purchasing power over *both* goods.

A price cut gives the consumer a gift, which is rather like an increase in *income*. Indeed, in an important sense, it *is* an increase in *available* income: Point *D* (3 concerts and 4 movies) originally cost Max $100, but after the decrease in the price of concerts, the same combination would cost him just (3 × $10) + (4 × $10) = $70, leaving him with $30 in *available income* to spend on more movies or concerts or both. This leads to the second effect of a change in price:

Income effect As the price of a good decreases, the consumer's purchasing power increases, causing a change in quantity demanded for the good.

*The **income effect** of a price change arises from a change in purchasing power over both goods. A drop in price increases purchasing power, while a rise in price decreases purchasing power.*

How will a change in purchasing power influence the quantity of a good demanded? That depends. Recall that an increase in income will increase the demand for normal goods and decrease the demand for inferior goods. The same is true for the *income effect* of a price cut: It can work to either *increase* or *decrease* the quantity of a good demanded, depending on whether the good is normal or inferior. For example, if concerts are a normal good for Max, then the income effect of a price cut will lead him to consume more of them; if concerts are inferior, the income effect will lead him to consume fewer.

Combining Substitution and Income Effects

Now let's look again at the impact of a price change, considering the substitution and income effects together. A change in the price of a good changes both the relative price of the good (the substitution effect) and the overall purchasing power of the consumer (the income effect). The ultimate impact of the price change on quantity demanded will depend on *both* of these effects. For normal goods, these two effects work together to push quantity demanded in the same direction. But for inferior goods, the two effects oppose each other. Let's see why.

Normal Goods. When the price of a normal good falls, the substitution effect *increases* quantity demanded. The price drop will also increase the consumer's

purchasing power and—for a normal good—*increase* quantity demanded even further. The opposite occurs when price increases: The substitution effect decreases quantity demanded, and the decline in purchasing power further decreases it. Figure 8 summarizes how the substitution and income effects combine to make the price and quantity of a normal good move in opposite directions:

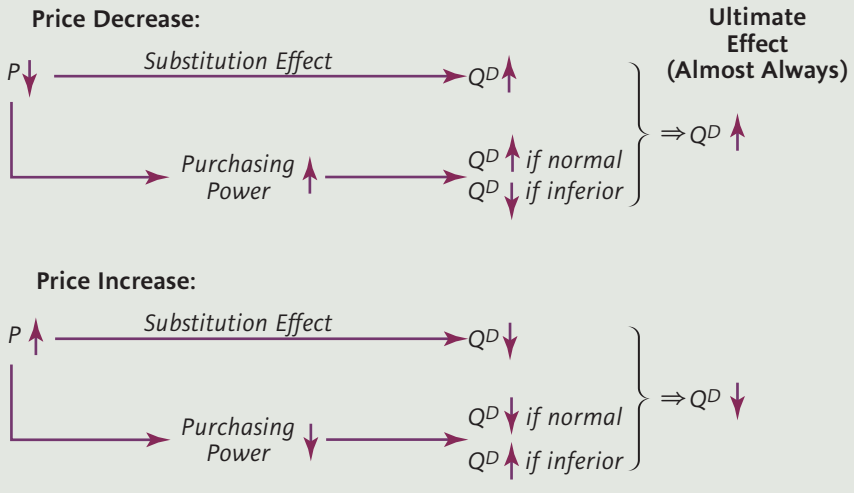
> *For normal goods, the substitution and income effects work together, causing quantity demanded to move in the opposite direction of the price. Normal goods, therefore, must always obey the law of demand.*

Inferior Goods. Now let's see how a price change affects the demand for *inferior* goods. As an example, consider intercity bus service. For many consumers, this is an inferior good: with a higher income, these consumers would choose quicker and more comfortable alternatives (such as air or train travel), and therefore demand *less* bus service. Now, if the price of bus service falls, the substitution effect would work, as always, to *increase* quantity demanded. The price cut will also, as always, increase the consumer's purchasing power. But if bus service is inferior, the rise in purchasing power will *decrease* quantity demanded. Thus, we have two opposing effects: the substitution effect, increasing quantity demanded, and the income effect, decreasing quantity demanded. In theory, either of these effects could dominate the other, so the quantity demanded could move in either direction.

In practice, however, the substitution effect virtually always dominates for inferior goods.

Why? Largely, because we consume such a wide variety of goods and services that a price cut in any one of them changes our purchasing power by only a small amount. For example, suppose you have an income of $20,000 per year, and you spend $500 per year on bus tickets. If the price of bus travel falls by, say, 20 percent,

FIGURE 8 **Income and Substitution Effects**



this would save you $100—like a gift of $100 in income. But $100 is only ½ percent of your income. Thus, a 20 percent fall in the price of bus travel would cause only a ½ percent rise in your purchasing power. Even if bus travel is, for you, an inferior good, we would expect only a tiny decrease in your quantity demanded when your purchasing power changes by such a small amount. Thus, the income effect should be very small. On the other hand, the *substitution* effect should be rather large: With bus travel now 20 percent cheaper, you will likely substitute away from other purchases and buy more bus travel.

For inferior goods, the substitution and income effects of a price change work against each other. The substitution effect moves quantity demanded in the opposite direction of the price, while the income effect moves it in the same direction as the price. But since the substitution effect virtually always dominates, consumption of inferior goods—like normal goods—will virtually always obey the law of demand.

Consumers in Markets

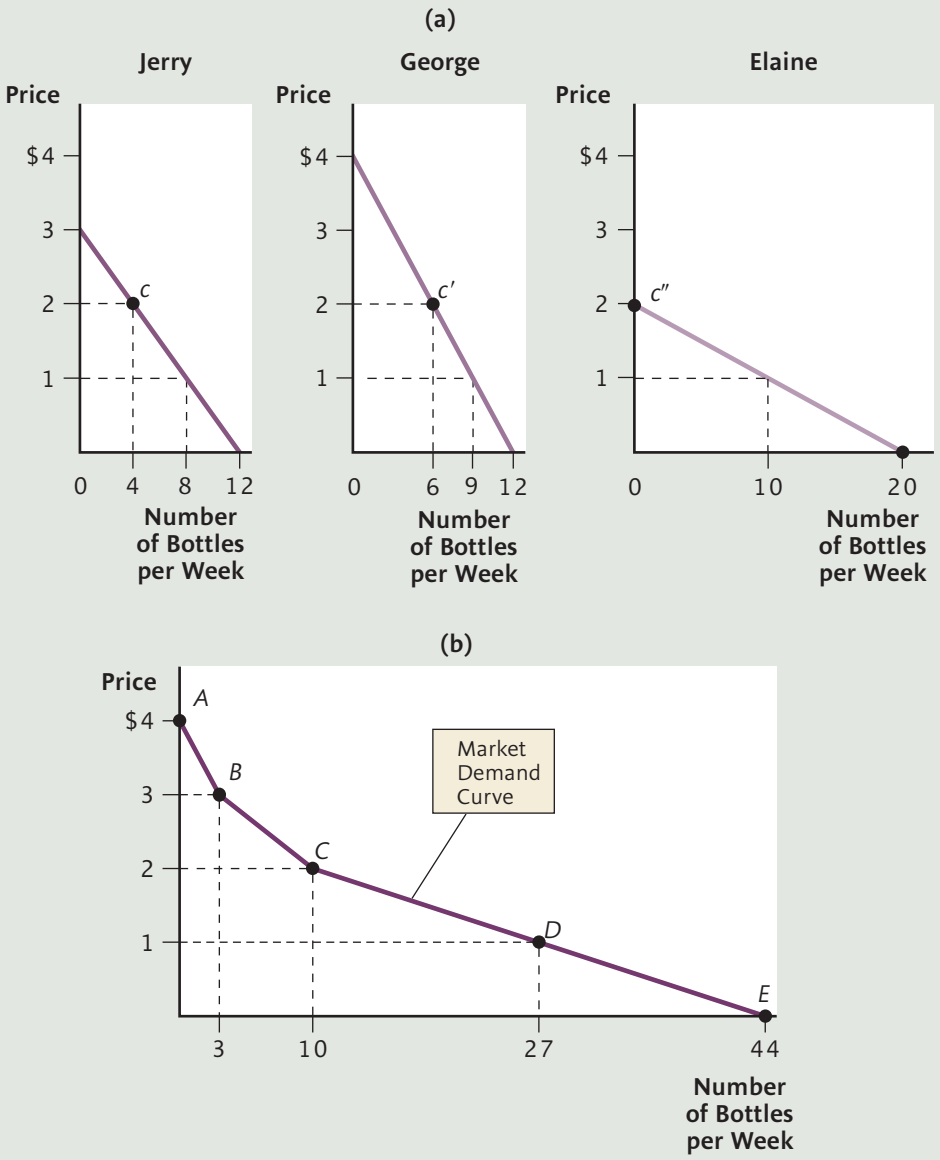
Since the market demand curve tells us the quantity of a good demanded by *all* consumers in a market, it makes sense that we can derive it by adding up the individual demand curves of every consumer in that market.

Figure 9 illustrates how this can be done in a small local market for bottled water, where, for simplicity, we assume that there are only three consumers—Jerry, George, and Elaine. The first three diagrams show their individual demand curves. If the market price were, say, $2 per bottle, Jerry would buy 4 bottles each week (point c), George would buy 6 (point c'), and Elaine would buy zero (point c''). Thus, the market quantity demanded at a price of $2 would be $4 + 6 + 0 = 10$, which is point C on the market demand curve. To obtain the entire market demand curve, we repeat this procedure at each different price, adding up the quantities demanded by each individual to obtain the total quantity demanded in the market. (Verify on your own that points A, B, D, and E have been obtained in the same way.) In effect, we obtain the market demand curve by summing horizontally across each of the individual demand curves:

The market demand curve is found by horizontally summing the individual demand curves of every consumer in the market.

Notice that as long as each individual's demand curve is downward sloping (and this will virtually always be the case), then the market demand curve will also be downward sloping. More directly, if a rise in price makes each consumer buy fewer units, then it will reduce the quantity bought by *all* consumers as well. Indeed, the market demand curve can still obey the law of demand even when *some* individuals violate it. Thus, although we are already quite confident about the law of demand at the individual level, we can be even *more* confident at the market level. This is why we always draw market demand curves with a downward slope.

FIGURE 9 From Individual to Market Demand



The individual demand curves show how much bottled water will be demanded by Jerry, George, and Elaine at different prices. As the price falls, each demands more. The market demand curve in panel (b) is obtained by adding up the total quantity demanded by all market participants at different prices.

Consumer Theory in Perspective

Our model of consumer theory—whether using marginal utility or indifference curves—may strike you as rather simple. Indeed, it was *purposely* kept simple, to bring out the "big ideas" more clearly. But can it explain and predict behavior in more complicated, real-world situations? In many cases, yes—with appropriate modification. In other cases, . . . no.

Extensions of the Model

One extension of the model—done in more advanced courses—is to incorporate choices among many goods, rather than just two. In that case, the calculations become more complicated, but the important conclusions we've reached (using either approach: marginal utility or indifference curves) remain the same.

Another extension is to recognize saving and borrowing. In our model, Max must spend his entire budget on movies and concerts, and cannot spend more. But in the real world, people can spend more than their incomes by borrowing, or spend less than their income and save funds for future use. One way to incorporate these possibilities is to define one of the goods as "future consumption." In that case, saving means sacrificing consumption of all goods now in order to have more future consumption, and borrowing means the opposite.

More difficult extensions incorporate *uncertainty* and *imperfect information*. For example, we assumed that Max knows with certainty the outcomes of his choices: the number of movies and concerts he will get, and the satisfaction he will get from them. But in many situations, you make a choice and take your chances. When you buy a car, it might be a lemon; when you pay for some types of surgery, there is a risk that it may be unsuccessful. You can reduce the chance of a bad outcome by acquiring more information. But information is often costly to obtain. When uncertainty and/or imperfect information are important aspects of consumer decision-making, additional tools are incorporated into the model. But the central idea—maximization of satisfaction based on rational preferences—remains in force.

Finally, you might think that consumer theory always regards people as relentlessly selfish, concerned only about their own consumption. In fact, when people trade in impersonal markets, this is mostly true: People *do* generally try to allocate their spending among different goods to achieve the greatest personal satisfaction. But in many areas of economic life, people act unselfishly. This, too, can be incorporated into the traditional model of consumer theory. One way is to treat the "satisfaction of others" (say, a family member or society at large) as another "good." Useful analyses of charitable giving, bequests, and voting behavior have been based on this type of modification of the model.

However, some real world situations don't seem to fit the basic ideas of consumer theory at all. In recent decades, these situations have led to the creation of an entirely new branch of the field: *behavioral economics*.

Behavioral Economics

Behavioral economics A subfield of economics focusing on decision-making patterns that deviate from those predicted by traditional consumer theory.

To understand what is different about behavioral economics, recall that traditional consumer theory assumes people make choices to maximize their satisfaction. These decisions are based on rational preferences: They compare alternative outcomes and choose the best among them in a logically consistent way. **Behavioral economics** has shown that preferences are sometimes *irrational*, and people sometimes make decisions—even highly consequential decisions—against their own interests, often in predictable ways. Businesses have long recognized and exploited these tendencies. Behavioral economists argue that government policy can exploit these same tendencies, mandating changes in how alternatives are presented to encourage people to choose more wisely.

Here are a few of the decision-making concepts stressed by behavioral economics.

Salience. Consumers often make decisions based not just on the outcome, but on the *salience* of a particular outcome—the extent to which it "jumps out at them." For example, one study showed that people will consume 77% more M&Ms when

the bowl contains 10 different colors as opposed to just seven different colors. Another study changed the price tags on certain grocery items to include sales tax, and marked them as such. Sales on those items declined, even though there was no change in the buyer's required payment at checkout. In these cases, consumers seem to be responding to what jumps out at them—the M&M colors in the bowl or the number on the price tag—rather than basing their decisions on what they will have to pay (in money or extra pounds).

Salience also plays a role in major economic decisions, such as taking out mortgage loans, borrowing on credit cards, or buying insurance. Corporations exploit salience by arranging documents or bills to keep certain facts from standing out. For example, on a credit card bill, the minimum payment may be more prominent than the total balance due, encouraging the cardholder to run a balance. Relevant information—such as how much a cardholder has actually paid in fees and interest over the past year—is not displayed at all.

Preference for Defaults. If choices are based on rational preferences, the "default choice" should not matter. In our example, Max gets the most satisfaction from three concerts and four movies per month. Consumer theory suggests he will make that choice even if he is initially given some other combination, and must "opt out" to have his preferred choice.

But studies have repeatedly shown that people tend to stick to the default choice for some decisions. In employer-sponsored retirement plans, when the default is "no contribution" and employees can easily opt-in to contribute to a retirement fund, only a small percentage do. The share that chooses to contribute remains small even when the employer matches the contributions of the employees—a substantial gift. But when the default option is changed to "contribute," so that *not* contributing requires opting *out*, the contribution rate rises dramatically.

Why do people tend to stick with the default, even when opting in or opting out is quick and easy? Perhaps they become anxious when thinking about the future or when presented with complicated, consequential decisions. Regardless of the cause, marketers and business firms often exploit this tendency. For example, websites—when obligated to get users' permission to receive advertising e-mail—will often make "Yes, please send me e-mail from related businesses" the default choice.

Decision-Making Environment. Sometimes, the *environment* in which a decision is made can exert a strong and surprising influence. This violates traditional consumer theory, where preferences are assumed to be based on a comparison of alternatives, and not on how or where the alternatives are presented. In research on preventing obesity, for example, health economists found that students seated in large groups in the cafeteria will tend to eat more than those seated in small groups. Their decisions about how much to eat are influenced by the setting. In another study, irrelevant but appealing photographs on a loan application had a profound impact on the percentage of people applying for the loan. Retailers know that the environmental context matters, and try to influence purchasing decisions by controlling the music in the store, the scent in the air, or even the thickness of the carpet beneath customers' feet.

Self-Binding. In a famous story in the Odyssey, Ulysses wants to hear the sirens sing as his ship passes near them. But their singing is so beautiful that, upon hearing it, sailors are always lured closer, ultimately smashing their ships or themselves on the rocks. So Ulysses instructs his shipmates to plug their own ears with wax and tie him to the mast. They are told not to release him or obey him—no matter what he says—until they are well beyond the sirens' voices and out of temptation.

From the perspective of rational preferences, Ulysses' strategy makes no sense. A decision-maker should always prefer a wide array of choices to a narrow one. After all, if you know what you want, expanding your choices might not always make you better off, but it should *never* make you worse off.

Yet in daily life we often see decision-makers contradict this principle, just as Ulysses did. When a dieter says, "Don't bring any ice cream into the house, I don't want to be tempted," they are saying, "If you give me an additional choice to eat ice cream later, I will be worse off." They would prefer to bind themselves to a narrower set of choices, for their own long-run good. For similar reasons, people often publicly announce their New Year's resolutions, to make it harder for them to go back on their promise later. And some even buy software programs that block their own access to games, social networking sites, or other computer distractions for a set period of time, so they can get work done.

Behavioral Economics and Policy

The insights of behavioral economics have begun to affect government policy. For example, in 2006, Congress changed the law governing pension plans, based on research by behavioral economists. For the first time, the government gave firms a financial incentive to make "automatic contribution" by employees the default choice. In such plans, employees who do *not* want to contribute to their own retirement account must take action to opt out. If past research is any guide, this simple change in policy could help millions of people achieve financial security in later life.

Another government policy based on self-binding has been instituted in some states with legal casino gambling. People with gambling problems can voluntarily put themselves on a list that bans them from any casino in the state for life. If caught violating the ban, the state will confiscate their winnings, and they will be subject to prosecution and even jail time. Thousands of former gamblers have put themselves on these lists.

In mid-2009, behavioral economics was poised to win its biggest policy victory, when the Obama administration proposed a new Consumer Financial Protection Agency to discourage some of the financial decisions that have gotten consumers into trouble. If approved by Congress, the new body would be empowered to establish industry-wide "vanilla" versions of mortgages, student loans, credit card agreements, and other financial products. These vanilla versions—which would be treated as defaults—would be simply designed and easy to understand. They would not have "teaser rates" or other enticements that sometimes encourage people to borrow or invest irresponsibly. Consequential information would be prominently displayed in simple language, rather than buried in legal fine print. Consumers would remain free to choose more complex or innovative products—say, an adjustable rate mortgage with a low teaser rate. But they would first have to "opt-out" of the government approved default versions. The economists advising the Obama administration believed that, if the program were put in place, relatively few consumers would choose to opt out.

Government policies like these, which arise from the insights of behavioral economics, are sometimes labeled "soft-paternalism." The government is being paternalistic by using the law to encourage healthier or more responsible choices. But the paternalism is "soft": the government mandates *how* choices are presented, but

ultimately leaves us free to choose. While there is some controversy over how far government should go in this direction, there is no doubt that behavioral economics has entered the mainstream of economic thought and has achieved considerable influence in government policy making.

Behavioral Economics and Traditional Theory

Traditional economic theory assumes that consumers have rational preferences. Behavioral economics analyzes decisions that violate rational preferences. Which theory is better? The answer is: Neither. The best model to use depends on the issue we are analyzing.

In most markets, most of the time, consumer responses can be understood very well with the traditional model. For example, if we are trying to understand how a gasoline tax or a rise in mass transit fares would affect the demand for automobiles, virtually all economists would reach for the traditional model of consumer choice. We would assume that consumers have rational preferences, and act to achieve the highest level of satisfaction, because—for the most part—they would do so in choosing whether to buy a new car.

But to understand why consumers often buy particular models of cars that trap them into spending more on gasoline than they realized, traditional theory would leave us short. For this problem, the insights of behavioral economics would work best.

Thus, although behavioral economics is often portrayed in the media as an "alternative" or even a "replacement" for traditional economic theory, few if any economists see it that way. Instead, behavioral economics is more commonly viewed as an addition to the existing body of economic theory—an extra limb that extends the theory's reach.

Using the Theory

© JEFF GREENBERG/ALAMY

IMPROVING EDUCATION

So far in this chapter, we've considered the problem of a consumer trying to select the best combination of goods and services. But consumer theory can be extended to consider almost *any* decision between two alternatives, including activities that cost us time rather than dollars. In this section, we apply the model of consumer choice to an important issue: the quality of education.[5]

Each year, various agencies within the U.S. Department of Education spend hundreds of millions of dollars on research to assess and implement new educational techniques. For example, suppose it is thought that computer-assisted instruction might help students learn better or more quickly. A typical research project to test this hypothesis would be a *controlled experiment* in which one group of students would be taught with the computer-assisted instruction and the other group would be taught without it. Then students in both groups would be tested. If the first group scores

[5] This section is based on ideas originally published in Richard B. McKenzie and Gordon Tullock, *The New World of Economics,* 3d ed. (Burr Ridge, IL: Irwin, 1981).

significantly higher, computer-assisted instruction will be deemed successful; if not, it will be deemed unsuccessful. To the disappointment of education researchers, most promising new techniques are found to be unsuccessful: Students seem to score about the same, no matter which techniques are tried.

Economists find these studies highly suspect, since the experiments treat students as passive responders to stimuli. Presented with a stimulus (the new technique), students are assumed to give a simple response (scoring higher on the exam). Where in this model, economists ask, are students treated as *decision makers,* who must make *choices* about allocating their scarce time?

Let's apply our model of consumer choice to a student's time allocation problem. To keep things simple, we'll assume a bleak world in which there are only two activities: studying economics and studying French. Instead of costing money, each of these activities costs *time,* and there is only so much time available. And instead of buying quantities of two goods, students "buy" points on their exams with hours spent studying.

Panel (a) of Figure 10 shows how we can represent the time allocation problem graphically. The economics test score is measured on the vertical axis and the French score on the horizontal axis. The straight line in the figure is the student's budget line, showing the tradeoff between economics and French scores. Our student can achieve any combination of scores on this budget line with her scarce time.

A few things are worth noting about the budget line in the figure. First, the more study time you devote to a subject, the better you will do on the test. But that means *less* study time for the other subject and a lower test score there. Thus, the opportunity cost of scoring better in French is scoring lower in economics, and vice versa. This is why the budget line has a negative slope: The higher the score in French, the lower the score in economics. As our student moves downward along the budget line, she is shifting hours away from studying economics and toward studying French.

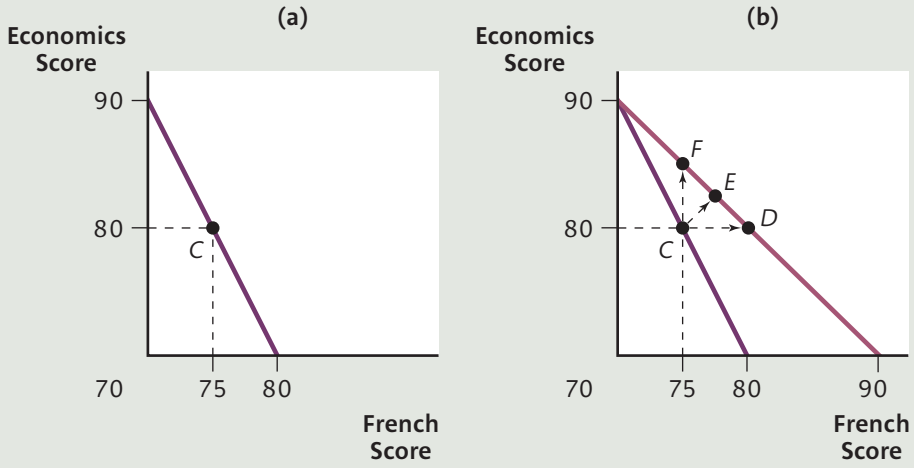
Second, notice that the vertical and horizontal axes both start at 70 rather than 0. This is to keep our example from becoming too depressing. If our student devotes *all* her study time to economics and none to French, she would score 90 in economics but still be able to score 70 (rather than zero) in French, just by attending class and paying attention. If she devotes all her time to French, she would score 80 in French and 70 in economics. (*Warning:* Do not try to use this example to convince your economics instructor you deserve at least a 70 on your next exam.)

Finally, the budget line in our example is drawn as a straight line with a slope of -2. So each additional point in French requires our student to sacrifice two points in economics. This assumption helps make the analysis more concrete. But none of our conclusions would be different if the budget line had a different slope, or even if it were curved so that the tradeoff would change as we moved along it. But let's take a moment to understand what our example implies.

As you've learned, the slope of any budget line is $-P_x/P_y$, where x is the good measured on the horizontal axis and y is the good measured on the vertical axis. We'll let P_F be the price of one point in French, and P_E be the price of one point in economics. Then, in our example, $-P_x/P_y$ translates into $-P_F/P_E$.

But what is the "price" of a test point in French or economics? Unlike the case of Max, who had to allocate his scarce *funds* between concerts and movies, our student must allocate her scarce *time* between the two "goods" she desires: test points in French and test points in economics. The *price* of a test point is therefore not a money price, but rather a *time price:* the number of study hours needed to achieve an additional point. For example, it might take an additional two hours of studying to achieve another point in French ($P_F = 2$) but just one hour to get another point in economics ($P_E = 1$). This would give us a budget line with a slope of $-P_F/P_E = -2/1 = -2$, which is the slope used in Figure 10(a).

FIGURE 10 Time Allocation



Panel (a) shows combinations of French and economics test scores that can be obtained for a given amount of study time. The slope of −2 indicates that each additional point in French requires a sacrifice of 2 points in economics. The student chooses point C. *Panel (b) shows that computer-assisted French instruction causes the budget line to rotate outward; French points are now less expensive. The student might move to point* D, *attaining a higher French score. Or she might choose* F, *using all of the time freed up in French to study economics. Or she might choose an intermediate point such as* E.

Now let's turn our attention to student decision making. Our student values both her economics score and her French score. But among all those combinations of scores on her budget line, which is the best choice? That depends on the student's preferences, whether characterized by the marginal utility approach or the indifference curve approach. Suppose that initially, this student's best choice is at point *C,* where she scores 80 in economics and 75 in French.

Now, let's introduce a new computer-assisted technique in the French class, one that is, in fact, remarkably effective: It enables students to learn more French with the same study time or to study less and learn the same amount. This is a *decrease* in the price of French points—it now takes less time to earn a point in French—so the budget line will rotate outward, as shown in panel (b) of Figure 10. On the new budget line, if our student devotes all of her time to French, she can score higher than before—90 instead of 80—so the horizontal intercept moves rightward. But since nothing has changed in her economics course, the vertical intercept remains unaffected. Notice, too, that the budget line's slope has changed to −1. Now, the opportunity cost of an additional point in French is one point in economics rather than two.

After the new technique is introduced in the French course, our *decision-making* student will locate at a point on her new budget line based once again on her preferences. Panel (b) illustrates some alternative possibilities. At point *D,* her performance in French would improve, but her economics performance would remain the same. This seems to be the kind of result education researchers have in mind when they design their experiments: If a successful technique is introduced in the French course, they expect to find the impact with a French test.

Point *F* illustrates a different choice: *Only* the economics performance improves, while the French score remains unchanged. Here, even though the technique in French is successful (it does, indeed, shift the budget line), none of its success shows up in higher French scores.

But wait: How can a new technique in the French course improve performance in economics but not at all in French? The answer is found by breaking down the impact

of the new technique into our familiar income and substitution effects. The new technique lowers the student's time price of getting additional points in French. The substitution effect (French points are relatively cheaper) will tend to improve her score in French, as she substitutes her time away from economics and toward French. But there is also an *"income"* effect: The "purchasing power" of her time has increased, since now she could use her fixed allotment of study time to "buy" higher scores in *both* courses. If performance in French is a normal good, this increase in "purchasing power" will work to increase her French score, but if it is an inferior good, it could work to *decrease* her French score. Point *F* could come about because French performance is *such* an inferior good that the negative income effect exactly cancels out the positive substitution effect. In this case, the education researchers will incorrectly judge the new technique a complete failure—it does not affect French scores at all.

Could this actually happen? Perhaps. It is easy to imagine a student deciding that 75 in French is good enough and using any study time freed up from better French instruction to improve her performance in some other course. More commonly, we expect a student to choose a point such as *E,* somewhere between points *D* and *F,* with performance improving in *both* courses. But even in this case, the higher French score measures just a *part* of the impact of the technique; the remaining effect is seen in a higher economics score.

In the real world, college students typically take several courses at once and have other competing interests for their time as well. Any time saved due to better teaching in a single course might well be "spent" on *all* of these alternatives, with only a little devoted to that single course. Thus, we cannot fully measure the impact of a new technique by looking at the score in that one course alone. This suggests why educational research is conducted as it is: A more accurate assessment would require a thorough accounting for all of a student's time, which is both expensive and difficult to achieve. Nevertheless, we remain justified in treating this research with some skepticism.

SUMMARY

Graphically, the budget constraint is represented by the *budget line*. Only combinations on or below the budget line are affordable. An increase in income shifts the budget line outward. A change in the price of a good causes the budget line to rotate. Whenever the budget line shifts or rotates, the consumer moves to a point on the *new* budget line. The consumer will always choose the point that provides the greatest level of satisfaction or *utility,* and this will depend on the consumer's unique preferences.

There are two alternative ways to represent consumer preferences, which lead to two different approaches to consumer decision making. The *marginal utility approach* is presented in the body of the chapter. In this approach, a utility-maximizing consumer chooses the combination of goods along her budget line at which the marginal utility per dollar spent is the same for all goods. When income or price changes, the consumer once again equates the marginal utility per dollar of both goods, resulting in a choice along the *new* budget line.

In the *indifference curve approach,* presented in the appendix, a consumer's preferences are represented by a collection of her *indifference curves,* called her *indifference map*. The highest level of utility or satisfaction is achieved at the point on the budget line that is also on the highest possible indifference curve. When income or price changes, the consumer moves to the point on the *new* budget line that is on the highest possible indifference curve.

Using either of the two approaches, we can trace the quantity of a good chosen at different prices for that good, and generate a downward sloping *demand* curve for that good. The downward slope reflects the interaction of the

substitution effect and the *income effect*. For a normal good, both effects contribute to the downward slope of the demand curve. For an inferior good, the two effects oppose each other. But in virtually all cases, the substitution effect dominates the income effect, so—once again—the demand curve will slope downward.

PROBLEM SET

Answers to even-numbered Questions and Problems can be found on the text website at www.cengage.com/economics/hall.

1. Parvez, a pharmacology student, has allocated $120 per month to spend on paperback novels and used CDs. Novels cost $8 each; CDs cost $6 each. Draw his budget line.
   a. Draw and label a second budget line that shows what happens when the price of a CD rises to $10.
   b. Draw and label a third budget line that shows what happens when the price of a CD rises to $10 *and* Parvez's income rises to $240.

2. [Uses the Marginal Utility Approach] Now go back to the original assumptions of problem 1 (novels cost $8, CDs cost $6, and income is $120). Suppose that Parvez is spending $120 monthly on paperback novels and used CDs. For novels, $MU/P = 5$; for CDs, $MU/P = 4$. Is he maximizing his utility? If not, should he consume (1) more novels and fewer CDs or (2) more CDs and fewer novels? Explain briefly.

3. [Uses the Marginal Utility Approach] Anita consumes both pizza and Pepsi. The following tables show the amount of utility she obtains from different amounts of these two goods:

Pizza		Pepsi	
Quantity	Utility	Quantity	Utility
4 slices	115	5 cans	63
5 slices	135	6 cans	75
6 slices	154	7 cans	86
7 slices	171	8 cans	96

   Suppose Pepsi costs $0.50 per can, pizza costs $1 per slice, and Anita has $9 to spend on food and drink. What combination of pizza and Pepsi will maximize her utility?

4. Three people have the following individual demand schedules for Count Chocula cereal that show how many boxes each would purchase monthly at different prices:

Price	Person 1	Person 2	Person 3
$5.00	0	1	2
$4.50	0	2	3
$4.00	0	3	4
$3.50	1	3	5

   a. What is the market demand schedule for this cereal? (Assume that these three people are the only buyers.) Draw the market demand curve.
   b. Why might the three people have different demand schedules?

5. Suppose that 1,000 people in a market *each* have the same monthly demand curve for bottled water, given by the equation $Q^D = 100 - 25P$, where P is the price for a 12-ounce bottle in dollars.
   a. How many bottles would be demanded in the entire market if the price is $1?
   b. How many bottles would be demanded in the entire market if the price is $2?
   c. Provide an equation for the *market* demand curve, showing how the market quantity demanded by all 1,000 consumers depends on the price.

6. What would happen to the market demand curve for polyester suits, an inferior good, if consumers' incomes rose?

7. Larsen E. Pulp, head of Pulp Fiction Publishing Co., just got some bad news: The price of paper, the company's most important input, has increased.
   a. On a supply/demand diagram, show what will happen to the price of Pulp's output (novels).
   b. Explain the resulting substitution and income effects for a typical Pulp customer. For each effect, will the customer's quantity demanded increase or decrease? Be sure to state any assumptions you are making.

8. "If a good is inferior, a rise in its price will cause people to buy more of it, thus violating the law of demand." True or false? Explain.

9. Which of the following descriptions of consumer behavior violates the assumption of *rational preferences*? Explain briefly.
   a. Joseph is confused: He doesn't know whether he'd prefer to take a job now or go to college full-time.
   b. Brenda likes mustard on her pasta, in spite of the fact that pasta is not meant to be eaten with mustard.
   c. Brewster says, "I'd rather see an action movie than a romantic comedy, and I'd rather see a romantic comedy than a foreign film. But given the choice, I think I'd rather see a foreign film than an action movie."

10. [Uses the Indifference Curve Approach] Howard spends all of his income on magazines and novels. Illustrate each of the following situations on a graph, with the quantity of magazines on the vertical axis and the quantity of novels on the horizontal axis. Use two budget lines and two indifference curves on each graph.
    a. When the price of magazines rises, Howard buys fewer magazines and more novels.
    b. When Howard's income rises, he buys more magazines *and* more novels.
    c. When Howard's income rises, he buys more magazines but *fewer* novels.

11. [Uses the Marginal Utility Approach] In Figure 5, we assumed that when Max's income rose, his marginal utility values for any given number of movies or concerts remained the same. But now suppose that when Max's income rises, and he can consume more movies and concerts, *an additional movie has less value* to Max than before. In particular, assume that Max's marginal utility values are as in the table below. Fill in the blanks for the missing values, and find Max's utility maximizing combination of concerts and movies. In Figure 5 in the chapter, locate the new combination as a point on the $200 budget line. (It will not be one of the labeled points.) With these new marginal utility values, is one of the two goods inferior? Explain.

Budget = $200 per month

Concerts at $20 each			Movies at $10 each		
(1) Number of Concerts per Month	(2) Marginal Utility from Last Concert (MU_C)	(3) Marginal Utility per Dollar Spent on Last Concert (MU_C/P_C)	(4) Number of Movies per Month	(5) Marginal Utility from Last Movie (MU_M)	(6) Marginal Utility per Dollar Spent on Last Movie (MU_M/P_M)
5	250		10	11	
			9	12	
6	180		8	12.5	
			7	13	
7	100		6	14	
			5	15	
8	50		4	16	
			3	18	
9	40		2	20	
			1	25	
10	30		0	—	—

12. [Uses the Marginal Utility Approach] In Figure 5, we assumed that when Max's income rose, his marginal utility values for any given number of movies or concerts remained the same. But now suppose that when Max's income rises, having the ability to enjoy more concerts or movies makes the *last movie* and the *last concert less valuable* to him, so all the marginal utility numbers shrink. In particular, assume that Max's marginal utility values are as in the following table. Fill in the blanks for the missing values, and find Max's utility-maximizing combination of concerts and movies. In Figure 5 in the chapter, locate the new combination as a point on the $200 budget line. (It will not be one of the labeled points.) With these new marginal utility values, is one of the two goods inferior? Explain.

Budget = $200 per month

Concerts at $20 each			Movies at $10 each		
(1) Number of Concerts per Month	(2) Marginal Utility from Last Concert (MU_C)	(3) Marginal Utility per Dollar Spent on Last Concert (MU_C/P_C)	(4) Number of Movies per Month	(5) Marginal Utility from Last Movie (MU_M)	(6) Marginal Utility per Dollar Spent on Last Movie (MU_M/P_M)
0	—		20	10	
			19	15	
1	36		18	18	
			17	20	
2	32		16	22	
			15	25	
3	28		14	30	
			13	35	
4	20		12	38	
			11	45	
5	15		0	—	

13. [Uses the Indifference Curve Approach]
    a. Draw a budget line for Cameron, who has a monthly income of $100. Assume that he buys steak and potatoes, and that steak costs $10 per pound and potatoes cost $2 per pound. Add an indifference curve for Cameron that is tangent to his budget line at the combination of 5 pounds of steak and 25 pounds of potatoes.
    b. Draw a new budget line for Cameron, if his monthly income falls to $80. Assume that potatoes are an inferior good to Cameron. Draw a new indifference curve tangent to his new budget constraint that reflects this inferiority. What will happen to Cameron's potato consumption? What will happen to his steak consumption?

14. [Uses the Indifference Curve Approach]
    a. Draw a budget line for Rafaella, who has a weekly income of $30. Assume that she buys chicken and eggs, and that chicken costs $5 per pound while eggs cost $1 each. Add an indifference curve for Rafaella that is tangent to her budget line at the combination of 4 pounds of chicken and 10 eggs.
    b. Draw a new budget line for Rafaella, if the price of chicken falls to $3 per pound. Assume that Rafaella views chicken and eggs as substitutes. What will happen to her chicken consumption? What will happen to her egg consumption?

15. [Uses the Indifference Curve Approach]
    a. Draw a budget line for Lynne, who has a weekly income of $225. Assume that she buys food and clothes, and that food costs $15 per bag while clothes cost $25 per item. Add an indifference curve for Lynne that is tangent to her budget line at the combination of 3 items of clothing and 10 bags of food.
    b. Draw a new indifference curve for Lynne, showing what will happen if her tastes change, so that she gets more satisfaction from an extra item of clothing, and less satisfaction from an extra bag of food.
    c. Returning to the original tangency, what will happen if Lynne decides to join a nudist colony?

More Challenging

16. The Smiths are a low-income family with $10,000 available annually to spend on food and shelter. Food costs $2 per unit, and shelter costs $1 per square foot per year. The Smiths are currently dividing the $10,000 equally between food and shelter. Use either the Marginal Utility Approach or Indifference Curve Approach.
    a. Draw their budget constraint on a diagram with food on the vertical axis and shelter on the horizontal axis. Label their current consumption choice. How much do they spend on food? On shelter?
    b. Suppose the price of shelter rises to $2 per square foot. Draw the new budget line. Can the Smiths continue to consume the same amounts of food and shelter as previously?
    c. In response to the increased price of shelter, the government makes available a special income supplement. The Smiths receive a cash grant of $5,000 that must be spent on food and shelter. Draw their new budget line and compare it to the line you derived in part *a*. *Could* the Smiths consume the same combination of food and shelter as in part *a*?
    d. With the cash grant and with shelter priced at $2 per square foot, *will* the family consume the same combination as in part *a*? Why, or why not?

17. When an economy is experiencing inflation, the prices of most goods and services are rising but at different rates. Imagine a simpler inflationary situation in which *all* prices, and all wages and incomes, are rising at the same rate, say 5 percent per year. What would happen to consumer choices in such a situation? (Hint: Think about what would happen to the budget line.)

18. [Uses the Indifference Curve Approach] With the quantity of popcorn on the vertical axis and the quantity of ice cream on the horizontal axis, draw indifference maps to illustrate each of the following situations. (Hint: Each will look different from the indifference maps in the appendix, because each violates one of the assumptions we made there.)
    a. Larry's marginal rate of substitution between ice cream and popcorn remains constant, no matter how much of each good he consumes.
    b. Heather loves ice cream but hates popcorn.

19. [Uses the Indifference Curve Approach] The appendix to this chapter states that when a consumer is buying the optimal combination of two goods x and y, then $MRS_{y,x} = P_x/P_y$. Draw a graph, with an indifference curve and a budget line, and with the quantity of y on the vertical axis, to illustrate the case where the consumer is buying a combination on his budget line for which $MRS_{y,x} > P_x/P_y$.

APPENDIX

The Indifference Curve Approach

This appendix presents an alternative approach to consumer decision making, which comes to the same conclusions as the approach in the body of the chapter. The appendix can be read in place of the section titled, "Consumer Decisions: The Marginal Utility Approach." We're naming it the "Indifference Curve Approach" after a graph that you will soon encounter.

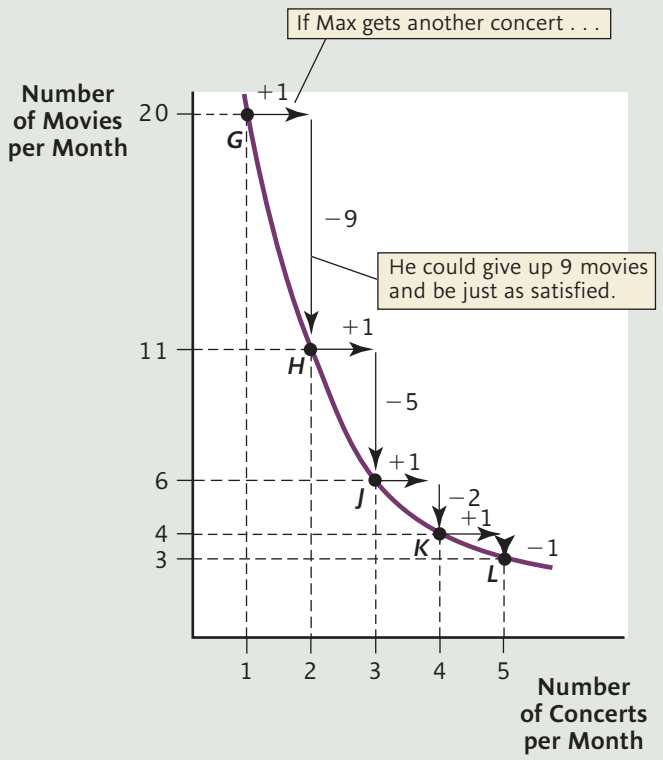
Let's start by reviewing what we've already discussed about preferences. We assume that an individual (1) can compare any two options and decide which is best, or that both are equally attractive, (2) makes choices that are logically consistent, and (3) prefers more of every good to less. The first two assumptions are summarized as *rational preferences*; the third tells us that a consumer will always choose to be *on* her budget line, rather than below it.

But now, we'll go a bit further.

An Indifference Curve

In Figure A.1, look at point *G*, which represents 20 movies and 1 concert per month. Suppose we get Max to look at this figure with us, and ask him to imagine how satisfied he would be to have the combination at point *G*. Max thinks about it for a minute, then says, "Okay, I know how satisfied I would be." Next, we say to Max, "Suppose you are at point *G* and we give you *another* concert each month, for a total of 2. That would make you even *more* satisfied, right?" Since Max likes concerts, he nods his head. But then we ask, "After giving you this additional concert, how many movies could we *take away* from you and leave you no better or worse off than you were originally, at point *G*?" Obliging fellow that he is, Max thinks about it and

FIGURE A.1 An Indifference Curve



answers, "Well, if I'm starting at point *G*, and you give me another concert, I suppose you could take away 9 movies and I'd be just as happy as I was at *G*."

Max has essentially told us that he is *indifferent between* point *G* on the one hand and point *H* on the other. We know this because starting at point *G*, adding 1 more concert and taking away 9 movies puts us at point *H*.

But let's keep going. Now we get Max to imagine that he's at point *H*, and we ask him the same question, and this time he answers, "I could trade 5 movies for 1 more concert and be equally well off." Now Max is telling us that he is indifferent between point *H* and *J*, since *J* gives him 1 more concert and 5 fewer movies than point *H*.

So far, we know Max is indifferent between point *G* and point *H*, and between point *H* and point *J*. So long as he is rational, he must be entirely indifferent among all three points—*G*, *H*, and *J*—since all three give him the same level of satisfaction. By continuing in this way, we can trace out a set of points that—as far as Max is concerned—are equally satisfying. When we connect these points with a curved line, we get one of Max's *indifference curves*.

> *An **indifference curve**[6] represents all combinations of two goods that make the consumer equally well off.*

Notice that the indifference curve in Figure A.1 slopes downward. This follows from our assumption about preferences that "more is better." Every time we give Max another concert, we make him better off. In order to find another point on his indifference curve, we must make him worse off by the same amount, *taking away* some movies.

The Marginal Rate of Substitution (MRS)

When we move along an indifference curve, from one point to another, we discover the maximum number of movies that Max would *willingly trade* for one more concert. For example, going from point *G* to point *H*, Max gives up 9 movies for 1 concert and remains indifferent. Therefore, from point *G*, if he gave up *10* movies for 1 concert, he'd be *worse off*, and he would not willingly make that trade. Thus, at point *G*, the *greatest* number of movies he'd willingly sacrifice for another concert would be 9.

[6] Bolded terms in this appendix are defined at the end of the appendix and in the glossary.

This notion of "willingness to trade," as you'll soon see, has an important role to play in our model of consumer decision making. And there's a technical term for it: the *marginal rate of substitution of movies for concerts*.[7] More generally, when the quantity of good *y* is measured on the vertical axis, and the quantity of good *x* is measured on the horizontal axis,

> *the **marginal rate of substitution of good y for good x** (**MRS**$_{y,x}$) along any segment of an indifference curve is the maximum rate at which a consumer would willingly trade units of* y *for units of* x.

For example, say we move along the indifference curve from point *G* (20 movies and 1 concert) all the way to point *L* (3 movies and 5 concerts). Since Max ends up on the same indifference curve, he'd be willing to make that move. That is, he would willingly give up 17 movies to get 4 more concerts, so his MRS would be $17 \div 4 = 4\frac{1}{4}$. We could say that Max is willing to give up "4¼ movies per concert." If we were to draw a straight line (not shown) between points *G* and *L*, the slope of that line would be $-4\frac{1}{4}$, giving us a graphical representation of the MRS for that segment of the indifference curve.

However, the value of the MRS will depend on the *size* of the move we make. Suppose we start at the same point, *G*, but make a smaller movement this time—to point *J*. For this smaller move, Max is willing to give up 14 movies to get 2 more concerts, so his MRS would be $14 \div 2 = 7$ (i.e., he is willing to give up 7 movies per concert). The MRS is now the slope (without the minus sign) of the straight line drawn between point *G* and point *J*.

MRS at a Point

In consumer theory, we are often interested in very small changes: the rate at which the consumer is willing to trade one good for a "tiny bit more" of another good. We imagine that the consumer makes a series of very tiny movements that, in total, account for the larger change we ultimately observe. Many goods (gasoline, electricity, or ground beef from a butcher) can, in fact, be consumed and traded off in arbitrarily small

[7] The term "marginal" is one that you'll encounter often in economics. The margin of a sheet of notebook paper is the area on the edge, just *beyond* the writing area. By analogy, a *marginal* value in economics measures what happens when we go a little bit *beyond* where we are now, by adding one or more unit of something.

increments. As these increments shrink, the segment of the indifference curve we are considering shrinks as well. Eventually, the segment becomes so small that—for all intents and purposes—we are looking at a single point on the curve.

Until now, we've defined the MRS using a *segment* of the indifference curve. And you've seen that the MRS depends on the size of segment we are considering. Can we use the MRS as a measure of willingness to trade when the segment shrinks to a *point*, such as point *H* in Figure A.1? Indeed we can, using the *slope* of the indifference curve itself *at point H*. To obtain that slope, we'd draw a straight line *tangent* to the indifference curve at point *H*, and use the slope of the tangent line. (To review slopes of curves and tangent lines, go back to Figure A.2 in the Mathematical Appendix to Chapter 1).

The MRS at any point on the indifference curve is equal to the (absolute value of) the slope of the curve at that point. When measured at a point, the $MRS_{y,x}$ tells us the maximum rate at which a consumer would willingly trade good Y for a tiny bit more of good X.

In our example, the "MRS at point *H*" in Figure A.1 is the rate at which Max would trade movies for concerts when we offer him just *a tiny bit more concert* (say, one-tenth or one-hundredth of a concert) beyond the one he already has. Of course, it is not possible to trade fractions of movies for fractions of concerts. The smallest trade Max could *actually* make would involve a whole concert. So for Max, as we shrink the segment we are considering, we could not realistically shrink it all the way to a single point. In that case, the slope at point *H* gives us only an *approximation* of Max's willingness to trade in the smallest increments possible: whole units. In the rest of our analysis, we'll use the slope at a point to approximate the MRS for the smallest movements that Max can realistically make.

How MRS Changes along an Indifference Curve

In Figure A.1, notice that as we move downward and rightward along the indifference curve, it gets flatter—the absolute value of its slope decreases. Another way of saying this is: As we move down an indifference curve, the *MRS* (the number of movies Max would be willing trade for another concert) gets smaller and smaller. To see why the *MRS* behaves this way, consider point *G*, high on Max's indifference curve. At this point, Max is seeing a lot of movies and relatively few concerts compared to points lower down, such as *J, K,* or *L*. With so few concerts, he'd value another one very highly. And with so many movies, each one he gives up doesn't harm him much. So, at a point like *G,* he'd be willing to trade a large number of movies for even one more concert. Using "*m*" for movies and "*c*" for concerts, his $MRS_{m,c}$ is relatively large. Since the *MRS* is the absolute value of the indifference curve's slope, the curve is relatively steep at point *G*.

But as we continue traveling down his indifference curve, from *G* to *H* to *J* and so on, movies become scarcer for Max, so each one given up hurts him a bit more. At the same time, he's attending more and more concerts, so adding another one doesn't benefit him as much as before. At a point like *K,* then, Max is more reluctant to trade movies for concerts. To get another concert, he'd willingly trade fewer movies at point *K* than at point *G*. So at point *K,* the *MRS* is relatively small and the curve is relatively flat.

The Indifference Map

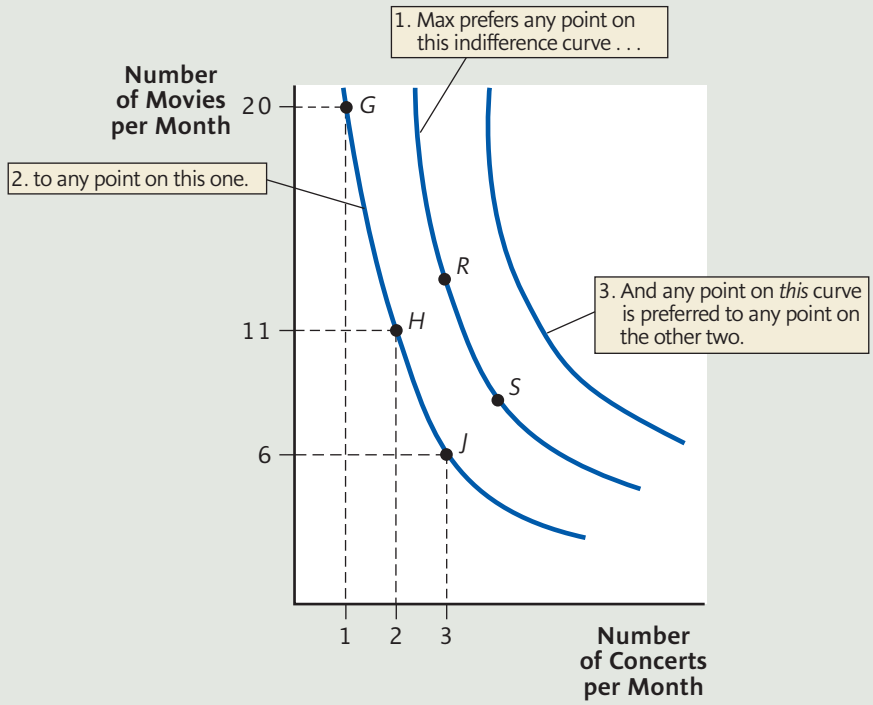
To trace out the indifference curve in Figure A.1, we began at a specific point—point *G*. Figure A.2 reproduces that same indifference curve through *G, H,* and *J*. But now consider the new point *R,* which involves more movies *and* more concerts than at point *H*. We know that point *R* is preferred to point *H* ("more is better"), so it is not on the indifference curve that goes through *H*. However, we can use the same procedure we used earlier to find a *new* indifference curve, connecting all points indifferent to point *R*. Indeed, we can repeat this procedure for any initial starting point we might choose, tracing out dozens or even hundreds of Max's indifference curves, as many as we'd like.

The result would be an **indifference map**, a set of indifference curves that describe Max's preferences, like the three curves in Figure A.2. We know that Max would always prefer any point on a higher indifference curve to any point on a lower one. For example, consider the points *H* and *S*. *S* represents more concerts but fewer movies than *H*. But Max's indifference map tells us that he *must* prefer *S* to *H*. Why? We know that he prefers *R* to *H,* since *R* has more of both goods. We also know that Max is indifferent between *R* and *S,* since they are on the same indifference curve. Since he is indifferent between *S* and *R,* but prefers *R* to *H,* then he must also prefer *S* to *H*.

The same technique could be used to show that

any point on a higher indifference curve is preferred to any point on a lower one.

FIGURE A.2 An Indifference Map



Thus, Max's indifference map tells us how he ranks all possible alternatives. An indifference map gives us a complete characterization of someone's preferences: It allows us to look at any two points and—just by seeing which indifference curves they are on—immediately know which, if either, is preferred.

Consumer Decision Making

Now we can combine everything you've learned about budget lines in the chapter, and what you've learned about indifference curves in this appendix, to determine the combination of movies and concerts that Max should choose. Figure A.3 adds Max's budget line to his indifference map. In drawing the budget line, we assume that Max has a monthly entertainment budget of $100, and that a concert costs $20 and a movie costs $10.

We assume that Max—like any consumer—wants to make himself as well off as possible. His optimal combination of movies and concerts will satisfy two criteria: (1) it will be a point *on* his budget line; and (2) it will lie on the highest indifference curve possible.

dangerous curves

Two Mistakes with Indifference Curves First, don't allow the ends of an indifference curve to "curl up," like the curve through point *B* in the following figure, so that the curve slopes upward at the ends. This violates our assumption of "more is better." To see why, notice that point *A* has more of both goods than point *B*. So as long as "more is better," *A* must be preferred to *B*. But then *A* and *B* are not indifferent, so they cannot lie on the same indifference curve. For the same reason, points *M* and *N* cannot lie on the same indifference curve. Remember that indifference curves *cannot slope upward.*

Second, don't allow two indifference curves to cross, like the two indifference curves passing through point *V*. *T* and *V* are on the same indifference curve, so the consumer must be indifferent between them. But *V* and *S* are also on the same indifference curve, so the consumer is indifferent between them, too. Since preferences must be consistent, the consumer must be indifferent between *T* and *S* as well. But this is impossible because *S* has more of both goods than *T*, a violation of "more is better." Remember that indifference curves *cannot cross.*

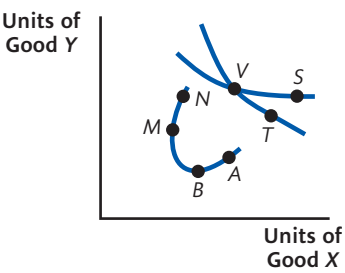
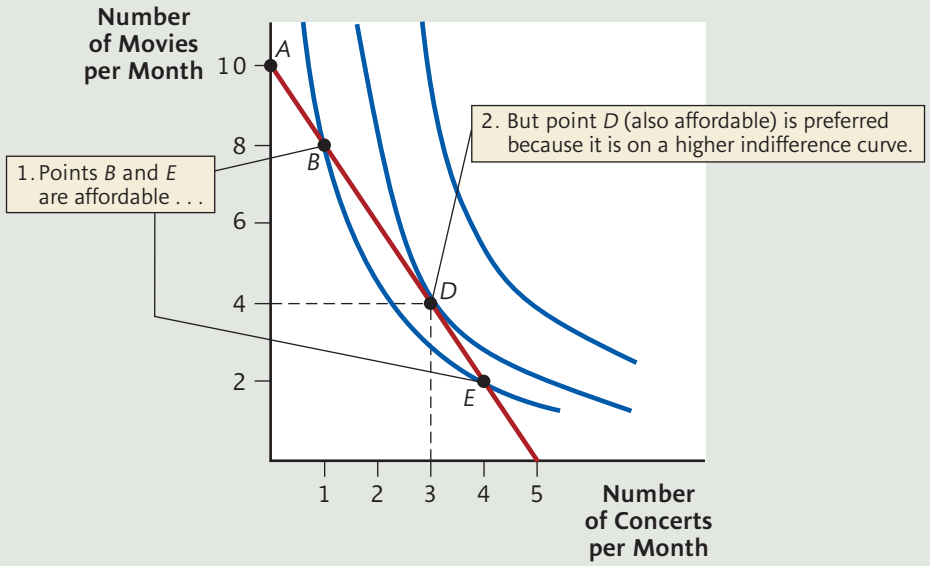


FIGURE A.3 **Consumer Decision Making with Indifference Curves**



Max can find this point by traveling down his budget line from *A*. As he does so, he will pass through a variety of indifference curves. (To see this clearly, you can pencil in additional indifference curves *between* the ones drawn in the figure.) At first, each indifference curve is higher than the one before, until he reaches the highest curve possible. This occurs at point *D,* where Max sees three concerts and four movies each month. Any further moves down the budget line will put him on lower indifference curves, so these moves would make him worse off. Point *D* is Max's optimal choice.

Notice something interesting about point *D*. First, it occurs where the indifference curve and the budget line are tangent—where they touch but don't cross. As you can see in the diagram, when an indifference curve actually crosses the budget line, we can always find some other point on the budget line that lies on a higher indifference curve.

Second, at point *D,* the slope of the indifference curve is the same as the slope of the budget line. Does this make sense? Yes, when you think about it this way: The absolute value of the indifference curve's slope—the *MRS*—tells us the rate at which Max would *willingly* trade movies for concerts. The slope of the budget line, by contrast, tells us the rate at which Max is *actually able* to trade movies for concerts. If there's any difference between the rate at which Max is *willing* to trade one good for another and the rate at which he is *able* to trade, he can always make himself better off by moving to another point on the budget line.

For example, suppose Max were at point *B* in Figure A.3. The indifference there is steeper than his budget line. In fact, the indifference curve appears to have a slope of about −4, so Max's MRS there is about 4; he'd *willingly* give up 4 movies for 1 more concert. But his budget line—as you learned earlier in the chapter—has a slope of −2. So according to his budget line, he is *able* to trade just 2 movies for each concert. If trading 4 movies for a concert would leave him indifferent, then trading just 2 movies for a concert must make him better off. We conclude that when Max's indifference curve is steeper than his budget line, he should spend more on concerts and less on movies.

Using similar reasoning, convince yourself that Max should make the opposite move—spending less on concerts and more on movies—if his indifference curve is *flatter* than his budget line, as it is at point *E*. Only when the indifference curve and the budget line have the

same slope—when they touch but do not cross—is Max as well off as possible. This is the point where the indifference curve is *tangent* to the budget line. When Max, or any other consumer, strives to be as well off as possible, he will follow this rule:

The optimal combination of goods for a consumer is the point on the budget line where an indifference curve is tangent to the budget line.

We can also express this decision-making rule in terms of the *MRS* and the prices of two goods. Recall that the slope of the budget line is $= -P_x/P_y$, so the absolute value of the budget line's slope is P_x/P_y. As you've just learned, the absolute value of the slope of an indifference curve is $MRS_{y,x}$. This allows us to state the decision-making rule as follows:

The optimal combination of two goods x and y is that combination on the budget line for which $MRS_{y,x} = P_x/P_y$.

If this condition is not met, there will be a difference between the rate at which a consumer is *willing* to trade good *y* for good *x*, and the rate at which he is *able* to trade them. This would leave the consumer with an opportunity to make himself better off.

What Happens When Things Change?

So far, as we've examined Max's search for the best combination of movies and concerts, we've assumed that Max's income, and the prices of each good, have remained unchanged. But in the real world, an individual's income, and the prices of the things they buy, *can* change. How would these changes affect his choice?

Changes in Income

Figure A.4 illustrates how an increase in income might affect Max's choice between movies and concerts. We assume that movies cost $10 each, concerts

FIGURE A.4 An Increase in Income

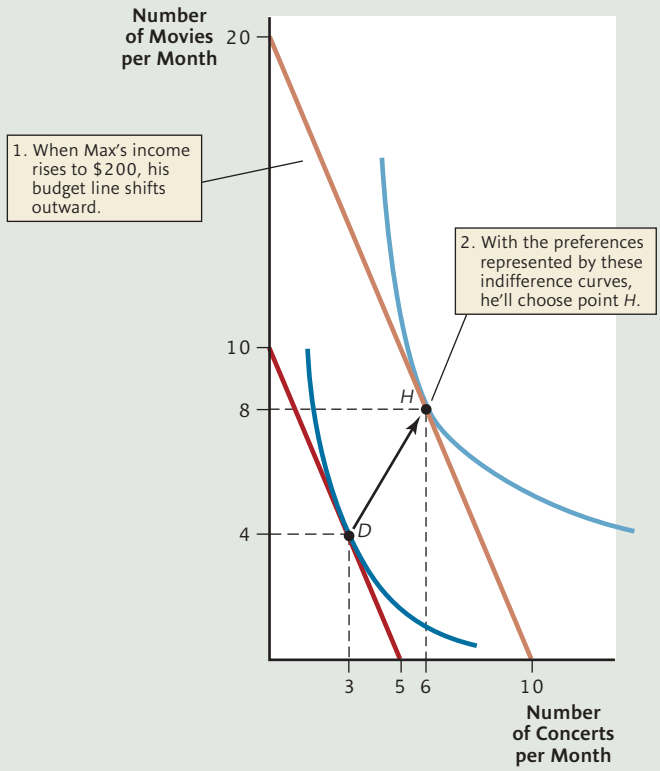
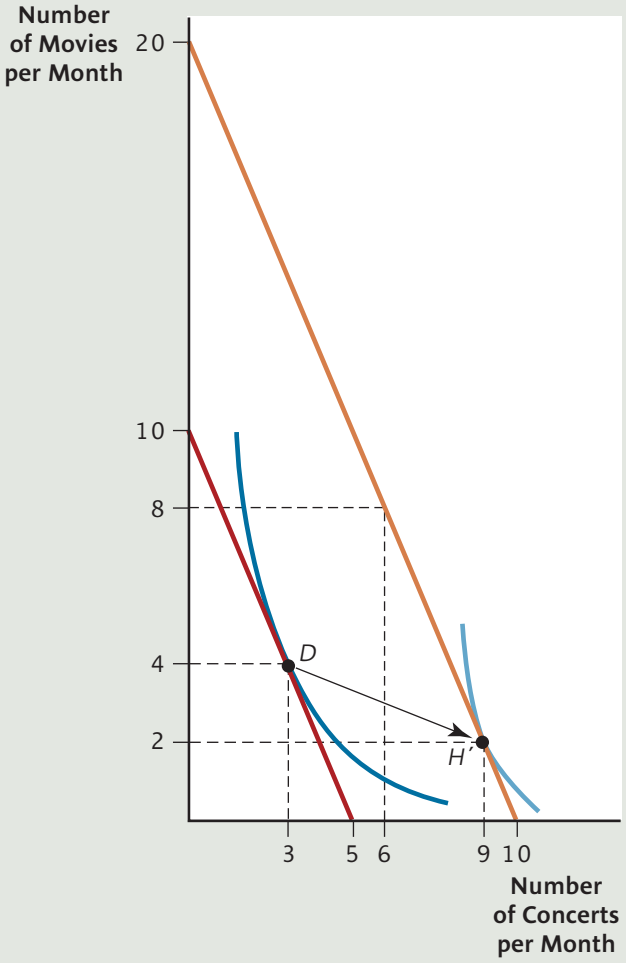


FIGURE A.5 **Income Rises and Movies are Inferior**

When income rises, whether demand for a good rises or falls depends on preferences, as represented by the consumer's indifference map. In this figure, Max's preferences make movies an inferior good. So when income rises from $100 to $200, he moves from point D *to point* H. *Concerts increase from 3 to 9, but movies decrease from 8 to 2.*



cost $20 each, and that these prices are not changing. Initially, Max has $100 to spend on the two goods, so his budget line is the lower line through point *D*. As we've already seen, under these conditions, the optimal combination for Max is point *D* (3 concerts and 4 movies).

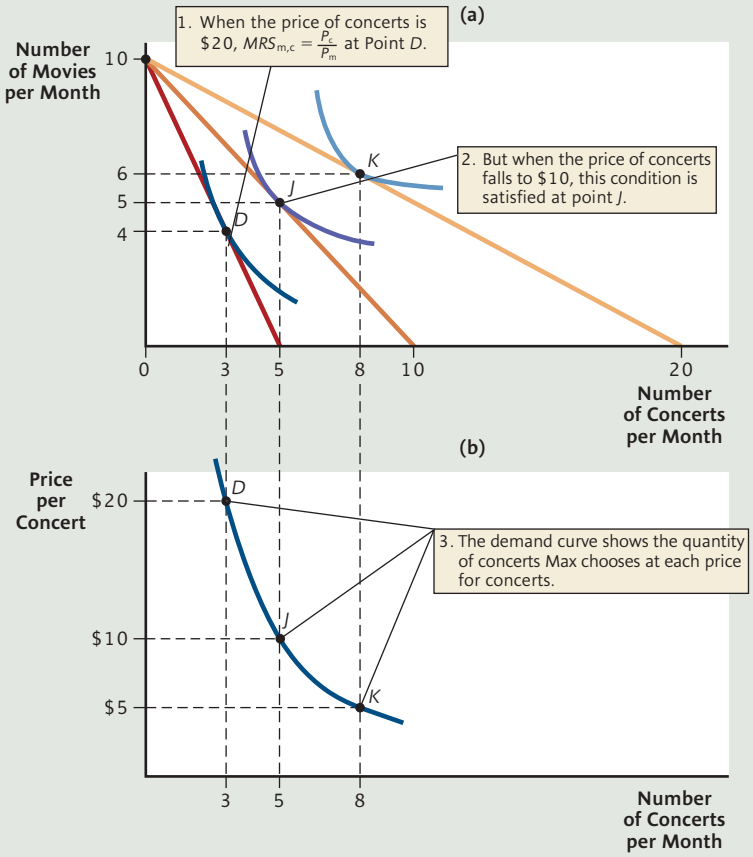
Now suppose Max's income increases to $200. Then his budget line will shift upward and rightward in the figure. How will he respond? As always, he will search along his budget line until he arrives at the highest possible indifference curve, which will be tangent to the budget line at that point.

But where will this point be? There are several possibilities, and they depend on Max's preferences, as reflected in his indifference map. In the figure, we've used an indifference map for Max that leads him to point *H*, enjoying 6 concerts and 8 movies per month. As you can see in the figure, at this point, he has reached the highest possible indifference curve that his budget allows. It's also the point at which $MRS_{m,c} = P_c/P_m = 2$.

Notice that, in moving from *D* to *H*, Max chooses to buy more concerts (6 rather than 3) and more movies (8 rather than 6). As discussed in Chapter 3, if an increase in income (with prices held constant) increases the quantity of a good demanded, the good is *normal*. For Max, with the indifference map we've assumed in Figure A.4, both concerts and movies would be normal goods.

But what if Max's preferences, and his indifference map, were as shown in Figure A.5? Here, after income rises, the tangency between his budget line and the highest indifference curve he could reach occurs at point like

FIGURE A.6 Deriving the Demand Curve with Indifference Curves



H′, with 9 concerts and 2 movies. In this case, the increase in income would cause Max's consumption of concerts to increase (from 3 to 9), but his consumption of movies to *fall* (from 6 to 2). If so, movies would be an *inferior good* for Max, one for which demand decreases when income increases, while concerts would be a normal good.

It's also possible for Max to have preferences that lead him to a different point—with more movies and *fewer* concerts than at point *D*. In this case, *concerts* would be the inferior good and movies would be normal. You can draw this case on your own.

A rise in income, with no change in prices, leads to a new quantity demanded for each good. Whether a particular good is normal (quantity demanded increases) or inferior (quantity demanded decreases) depends on the individual's preferences, as represented by his indifference map.

Changes in Price

Let's explore what happens to Max when the price of a concert decreases from $20 to $10, while his income and the price of a movie remain unchanged. The drop in the price of concerts rotates Max's budget line rightward, pivoting around its vertical intercept, as illustrated in the upper panel of Figure A.6. What will Max do after his budget line rotates in this way? Based on his indifference curves—as they appear in the figure—he'd choose point *J*. This is the new combination of movies and concerts on his budget line that makes him as well off as possible (puts him on the highest possible indifference curve that he can afford). It's also the point at which $MRS_{m,c} = P_c/P_m = 1$, since movies and concerts now have the same price.

What if we dropped the price of concerts again, this time, to $5? Then Max's budget line rotates further rightward, and he will once again find the best possible point. In the figure, Max is shown choosing point *K*, attending 8 concerts and 6 movies.

The Individual's Demand Curve

You've just seen that each time the price of concerts changes, so does the quantity of concerts Max will want to attend. The lower panel of Figure A.5 illustrates this relationship by plotting the quantity of concerts demanded on the horizontal axis and the *price* of concerts on the vertical axis. For example, in both the upper and lower panels, point *D* tells us that when the price of concerts is $20, Max will see three of them. When we connect points like *D, J,* and *K* in the lower panel, we get Max's demand curve, which shows the quantity of a good he demands at each different price. Notice that Max's demand curve for concerts slopes downward—a fall in the price of concerts increases the quantity demanded—showing that for Max, concerts obey the law of demand.

But if Max's preferences—and his indifference map—had been different, could his response to a price change have *violated* the law of demand? The answer is yes . . . and no. Yes, it is theoretically possible. (As a challenge, try penciling in a new set of indifference curves that would give Max an *upward-sloping* demand curve in the figure.) But no, it does not seem to happen in practice. To find out why, it's time to go back to the body of the chapter, to the section titled, "Income and Substitution Effects."

DEFINITIONS

Indifference curve A curve representing all combinations of two goods that make the consumer equally well off.

Indifference map A set of indifference curves that represent an individual's preferences.

Marginal rate of substitution ($MRS_{y,x}$) The maximum amount of good *y* a consumer would willingly trade for one more unit of good *x*. Also, the slope of a segment of an indifference curve.

CHAPTER 7

Production and Cost

In November 2008, the brewing company InBev purchased Anheuser-Busch for $52 billion. In January 2009, the world's largest pharmaceutical company, Pfizer, announced it would buy drug maker Wyeth for $68 billion, and a few months later another drug giant—Merck—bought rival Schering-Plough for $41 billion. A surprise merger came in July, 2009, when Amazon bought Zappos for $847 million.

Events like these are not unusual. In a typical year, more than a thousand U.S. corporations—with a total value of close to $1 trillion—are acquired by other corporations. Thousands more companies are acquired each year in Europe and Asia. The stockholders who own the firms being combined usually end up favoring these deals because they believe that a larger, combined firm will perform better in the marketplace than would two separate firms. In most cases, a major reason for this belief is the predicted impact the merger will have on *costs*.

This focus on costs is not surprising. A firm's managers strive to earn the highest possible **profit**—the difference between a firm's revenue and its costs. And controlling costs is one way to increase profit.

Profit Total revenue minus total cost.

We'll have more to say about maximizing profit in the next chapter. In this chapter we focus on cost: how to think about it, how to measure it, and how decisions at the firm cause cost to change. As a first step, we'll look more closely at the main activity of business firms, from which its costs arise: *production*.

Production

Pfizer, Merck, and InBev are all examples of **business firms**: organizations owned and operated by private individuals that specialize in production.

Business firm An organization, owned and operated by private individuals, that specializes in production.

Your first image when you hear the word *production* may be a busy, noisy factory where goods are assembled and packaged and then carted off to a warehouse for eventual sale to the public. Large manufacturers may come to mind—Ford, Boeing, or even Ben & Jerry's. All of these companies produce things, but the word *production* encompasses more than just manufacturing.

Production is the process of combining inputs *to make* goods and services.

Notice that the definition refers to goods *and* services. Some production does indeed create physical *goods,* like automobiles, aircraft, or ice cream. But production also creates *services*. Indeed, many of America's largest corporations are service providers, including Wells Fargo (banking and investment services), American Airlines

(transportation services), Verizon (telecommunications services), and Wal-Mart (retailing services).

What about the inputs that are used to produce things? These include the four resources (land, labor, capital, and entrepreneurship), as well as other things. For example, to make the book you are reading now, a business firm (Cengage Learning) used several resources, including *labor* (of the authors, editors, art designers, and others), *capital* (buildings, office furniture, computers), and *land* (under the buildings). But the company also used *other* inputs that were produced by other firms, including raw materials like paper and ink, transportation and telecommunications services, and more.

Technology and Production

When you hear the word "technology," you are likely to think of the latest electronic gadget that someone told you about. But in economics technology has a specific meaning:

Technology The methods available for combining inputs to produce a good or service.

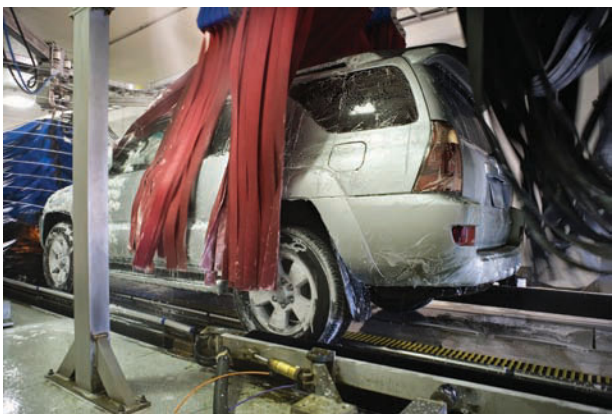
> *A firm's* ***technology*** *refers to the methods it can use to turn inputs into outputs (produced goods or services).*

If the firm's technology allows it to use *different* methods for producing the same level of output, we assume the firm will choose the cheapest method it can find. You'll learn more about how a firm makes this choice later.

In this chapter, to keep things simple, we'll spell out the production technology for a mythical firm that uses only two inputs: *capital* and *labor*. Our firm is Spotless Car Wash, whose output is a service: washing cars. The firm's capital consists of automated car-washing lines. It's labor is full-time workers who drive the cars onto the line, drive them out, towel them down at the end, and deal with customers.[1]

Short-Run versus Long-Run Decisions

When a firm changes its level of production, it will want to adjust the amounts of inputs it uses. But these adjustments depend on the *time horizon* the firm's managers are thinking about. Some inputs can be adjusted relatively quickly. Most firms, for example, can hire more labor and purchase more raw materials within a few weeks or less. But at many firms, some inputs take longer to adjust. It may take a year or longer before an automobile firm can purchase and install new assembly lines, or acquire additional factory space. And legal obligations, like leases or rental agreements, can delay efforts to downsize operations, because the firm will have to continue paying for equipment and buildings for some time, whether it uses them or not.



Thus, when we ask, "What quantities of what inputs will the firm use to produce a given level of ouput?" the answer will depend on whether we are asking about

[1] Of course, a car wash would use other inputs besides just capital and labor: water, washrags, soap, electricity, and so on. But the costs of these inputs would be minor when compared to the costs of labor and capital. To keep our example simple, we ignore these other inputs.

next month or *next year*. If it's next month, a firm may be stuck with the factory and equipment it currently has, so it can only adjust its labor and raw materials. If we're asking about next year, there is more flexibility—enough time to adjust capital equipment as well.

These considerations make it useful to divide the different time horizons firms can use into two broad categories: the *long run* and the *short run*.

*The **long run** is a time period long enough for a firm to change the quantity of* all *of its inputs.*

Long run A time horizon long enough for a firm to vary all of its inputs.

Another way to say this is that, over the long run, all the inputs the firm uses are viewed as **variable inputs**—inputs that can be adjusted up or down as the quantity of output changes.

Variable input An input whose usage can change as the level of output changes.

At Spotless Car Wash, we'll imagine it takes a year to acquire and install a new automated line, or to sell the lines it already has. Thus, for Spotless the long run is a time horizon of one year or longer. When its managers make long-run decisions, they regard all inputs (labor and capital in this case) as variable inputs.

What about shorter time periods? The company may be stuck with the current quantities of some inputs. We call these **fixed inputs**—inputs that, over the time period we're considering, cannot be adjusted as output changes. Using this terminology, we can define the short run as follows:

Fixed input An input whose quantity must remain constant, regardless of how much output is produced.

*The **short run** is a time period during which at least one of the firm's inputs is fixed.*

Short run A time horizon during which at least one of the firm's inputs cannot be varied.

For Spotless Car Wash, the short run is any time period less than a year, because that is how long it is stuck with its current quantity of automated lines. Over the short run, Spotless's labor is a variable input, but its capital is a fixed input.

Production in the Short Run

When firms make decisions over the short run, there is nothing they can do about their fixed inputs: They are stuck with whatever quantity they have. They can, however, make choices about their variable inputs. Indeed, we see examples of such short-run decisions all the time. If Levi Strauss decides to increase production of blue jeans over the next quarter, it may use additional workers, cotton cloth, and sewing machines. But it continues to make do with the same factories because there isn't time to expand them or acquire new ones. Here, workers, cloth, and sewing machines are all variable over the quarter, while the factory buildings are fixed.

At Spotless Car Wash, over the short run, labor is the only variable input, and capital is the only fixed input. The three columns in Table 1 describe Spotless's production choices in the short run. Column 1 shows the quantity of the fixed input, capital (K); column 2 the quantity of the variable input, labor (L). Note that in the short run, Spotless is stuck with one unit of capital—one automated line—but it can take on as many or as few workers as it wishes. Column 3 shows the firm's *total product* (Q).

***Total product** is the maximum quantity of output that can be produced from a given combination of inputs.*

Total product The maximum quantity of output that can be produced from a given combination of inputs.

TABLE 1

Short-Run Production at Spotless Car Wash

Quantity of Capital	Quantity of Labor	Total Product (Cars Washed per Day)
1	0	0
1	1	30
1	2	90
1	3	130
1	4	160
1	5	184
1	6	196

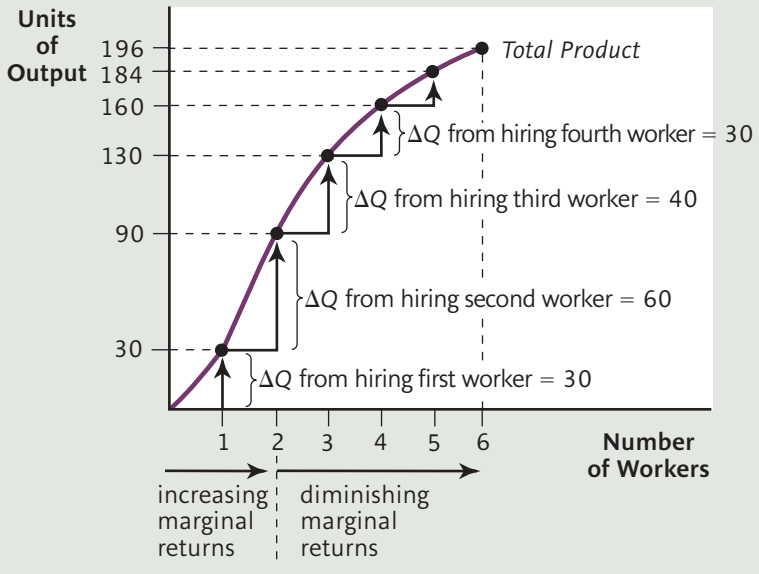
For example, the table shows us that with one automated line but no labor, total product is zero. With one line and six workers, total product is 196 cars washed per day.

The total product numbers in the table tell us the *maximum* output for each number of workers. We can also reverse this logic, and say that for each value of the total product, the labor column shows us the *lowest* number of workers that can produce it. Since labor is the only variable input, this lowest number of workers will also be the *least-cost* method of producing any level of output.

Figure 1 shows Spotless's *total product curve*. The horizontal axis represents the number of workers, while the vertical axis measures total product. (The amount of capital—which is held fixed at one automated line—is not shown on the graph.) Notice that each time the firm hires another worker, output increases, so the total product curve slopes upward. The vertical arrows in the figure show precisely *how much* output increases with each one-unit rise in employment. We call this rise in output the *marginal product of labor*.

FIGURE 1 Total and Marginal Product

The total product curve shows the total amount of output that can be produced using various numbers of workers. The marginal product of labor (MPL) is the change in total product when another worker is hired. The MPL for each change in employment is indicated by the length of the vertical arrows.



Using the Greek letter Δ ("delta") to stand for "change in," we can define marginal product this way:

*The **marginal product of labor*** **(MPL)** *is the change in total product* (ΔQ) *divided by the change in the number of workers employed* (ΔL)*:*

$$\text{MPL} = \frac{\Delta Q}{\Delta L}$$

The MPL *tells us the rise in output produced when one more worker is hired.*

Marginal product of labor The additional output produced when one more worker is hired.

For example, if employment rises from 2 to 3 workers, total product rises from 90 to 130, so the marginal product of labor for *that* change in employment is calculated as (130 − 90)/1 = 40 units of output.

Marginal Returns to Labor

Look at the vertical arrows in Figure 1, which measure the marginal product of labor, and you may notice something interesting. As more and more workers are hired, the *MPL* first increases (the vertical arrows get longer) and then decreases (the arrows get shorter). This pattern is believed to be typical at many types of firms, so it's worth exploring.

Increasing Marginal Returns to Labor

When the marginal product of labor rises as more workers are hired, there are **increasing marginal returns to labor.** Each time a worker is hired, total output rises by more than it did when the previous worker was hired. Why might this happen? Additional workers may allow production to become more specialized.

Increasing marginal returns to labor The marginal product of labor increases as more labor is hired.

For example, Figure 1 tells us that Spotless Car Wash experiences increasing returns to labor up to the hiring of the second worker. While one worker *could* operate the car wash alone, he or she would have to do everything: drive the cars on and off the line, towel them down, and deal with customers. Much time would be wasted switching from one task to another. Table 1 tells us that one worker can wash only 30 cars each day. Add a second worker, though, and now specialization is possible. One worker can collect money and drive the cars onto the line, and the other can drive them off and towel them down. Thus, with two workers, output rises all the way to 90 car washes per day; the second worker adds more to production (60 car washes) than the first (30 car washes) by making *both* workers more productive.

Diminishing Marginal Returns to Labor

When the marginal product of labor is decreasing, we say there are **diminishing marginal returns to labor.** Output still rises when another worker is added, so marginal product is positive. But the rise in output is smaller and smaller with each successive worker.

diminishing marginal returns to labor The marginal product of labor decreases as more labor is hired.

Why does this happen? For one thing, as we keep adding workers, additional gains from specialization will be harder and harder to achieve. Moreover, each worker will have less and less of the fixed inputs with which to work.

This last point is worth stressing. It applies not just to labor but to any variable input. In all kinds of production, if we keep increasing the quantity of any one input, while holding the others fixed, diminishing marginal returns will eventually set in.

For example, if a farmer keeps adding additional pounds of fertilizer to fixed amounts of land and labor, the yield may continue to increase, but eventually the *size* of the increase—the marginal product of fertilizer—will begin to come down. This tendency is so pervasive and widespread that it has been deemed a law:

Law of diminishing (marginal) returns As more and more of any input is added to a fixed amount of other inputs, its marginal product will eventually decline.

> *The **law of diminishing (marginal) returns** states that as we continue to add more of any one input (holding the other inputs constant), its marginal product will eventually decline.*

The law of diminishing returns is a physical law, not an economic one. It is based on the nature of production—on the physical relationship between inputs and outputs with a given technology.

Figure 1 tells us that at Spotless diminishing returns set in after two workers have been hired. Beyond this point, the firm is crowding more and more workers into a car wash with just one automated line. Output continues to increase, since there is usually *something* an additional worker can do to move the cars through the line more quickly, but the increase is less dramatic each time.

Thinking About Costs

The previous section dealt with *production*—the *physical* relationship between inputs and outputs. But a more critical concern for a firm is: What will it *cost* to produce any level of output? Everything you've learned about production will help you understand the behavior of costs.

Let's start by revisiting a familiar notion. In Chapter 1 you learned that economists always think of cost as *opportunity cost*—what we must give up in order to do something. This concept applies to the firm as well:

> *A firm's total cost of producing a given level of output is the opportunity cost of the owners—everything they must give up in order to produce that amount of output.*

Using the concept of opportunity cost can help us understand which costs matter—and which don't—when making business decisions.

The Irrelevance of Sunk Costs

Suppose that last year, Acme Pharmaceutical Company spent $10 million developing a new drug to treat acne that, if successful, would have generated millions of dollars in annual sales revenue. At first, the drug seemed to work as intended. But then, just before launching production, management discovered that the new drug didn't cure acne at all—but was remarkably effective in treating a rare underarm fungus. In this smaller, less lucrative market, annual sales revenue would be just $30,000. Now management must decide: Should they sell the drug as an antifungus remedy?

When confronted with a problem like this, some people will say: "Acme should *not* sell the drug. You don't sell something for $30,000 a year when it cost you $10 million to make it." Others will respond this way: "Of course Acme should sell the drug. If they don't, they'd be wasting that huge investment of $10 million." But to an economist, neither approach to answering this question is correct, because

both use the $10 million development cost to reach a conclusion—and that cost is completely *irrelevant* to the decision.

The $10 million already spent on developing the drug is an example of a *sunk cost*. More generally,

> *a* ***sunk cost*** *is one that already has been paid, or must be paid,* regardless *of any future action being considered.*

Sunk cost A cost that has been paid or must be paid, regardless of any future action being considered.

In the case of Acme, the development cost has been paid already. The firm will not get this money back, whether it chooses to sell the drug in this new smaller market or not. Because the $10 million is not part of the opportunity cost of either choice—something that would have to be sacrificed *for* that choice—it should have no bearing on the decision. For Acme, as for any business,

> *Sunk costs should not be considered when making decisions.*

What *should* be considered are the costs that *do* depend on the decision about producing the drug, namely, the cost of actually manufacturing it and marketing it for the smaller market. If these costs are less than the $30,000 Acme could earn in annual revenue, Acme should produce the drug. Otherwise, it should not.

Look again at the definition of sunk cost and you'll see that even a *future* payment can be sunk, if an *unavoidable commitment to pay it has already been made*. Suppose, for example, Acme Pharmaceuticals has signed an employment contract with a research scientist, legally binding the firm to pay her annual salary for three years even if she is laid off. Although some or all of the payments haven't yet been made, all three years of salary are sunk costs for Acme because they *must* be made no matter what Acme does. As sunk costs, they are irrelevant to Acme's decisions.

Our insight about the irrelevance of sunk costs applies beyond the business sector, to decisions in general. For example, suppose that after completing two years of medical school, you've decided that you've made a mistake: You'd rather be a lawyer. You might be tempted to stay in medical school because of the money and time you've already spent there. But you can't get your time back. And since you can't sell your two years of medical training to someone else, you can't get your money back either. Thus, the costs of your first two years in medical school have either *been* paid or the unavoidable *commitment* to pay them has already been made (say, because you've taken out a student loan). They are sunk costs and should have no influence on your career decision. Only the costs that *depend* on your decision are relevant. If you choose to stay in medical school, you'll have to spend time, effort, and expense for your *remaining* years there. If you switch to law school, you'll spend the time, effort, and expense of three years *there*. These are the costs you should consider (along with the benefits) for each choice.

Explicit versus Implicit Costs

In Chapter 2, in discussing the opportunity cost of education, you learned that there are two types of costs: *explicit* (involving actual payments) and *implicit* (no money changes hands). The same distinction applies to costs for a business firm.

Suppose you've opened a restaurant in a building that you already owned. You don't pay any rent, so there's no explicit rental cost. Does this mean that using the building is free?

TABLE 2

A Firm's Costs

Explicit Costs	Implicit Costs
Rent paid out	Opportunity cost of:
Interest on loans	Owner's land and buildings (rent foregone)
Managers' salaries	Owner's money (investment income
Hourly workers' wages	foregone)
Cost of raw materials	Owner's time (labor income foregone)

To an accountant—who focuses on actual money payments—the answer is yes. But to an economist—who thinks of opportunity cost—the answer is *absolutely not*. By using your own building for your restaurant, you are sacrificing the opportunity to rent it to someone else. This *foregone rent* is an *implicit cost,* and it is as much a cost of production as the rent you would pay if you didn't own a building yourself. In both cases, something is given up to produce your output.

Now suppose that instead of borrowing money to set up your restaurant, you used $100,000 of your own money. Therefore, you aren't paying any interest. But there is an opportunity cost: your $100,000 *could* have been put in the bank or lent to someone else, where it would be earning interest for you. If the going interest rate is 5 percent, then each year that you run your restaurant, you are giving up $5,000 in interest you could have instead. This *foregone interest* is another implicit cost of your business.

Finally, suppose you decided to manage your restaurant yourself. Have you escaped the costs of hiring a manager? Not really, because you are still bearing an opportunity cost: You *could* do something else with your time. We measure the value of your time as the income you *could* earn by devoting your labor to your next-best income-earning activity. This *foregone labor income*—the wage or salary you could be earning elsewhere—is an implicit cost of your business, and therefore part of its opportunity cost. Table 2 summarizes our discussion by listing some common categories of costs that business firms face, both explicit (on the left) and implicit (on the right).

Cost in the Short Run

Managers must answer questions about costs over different time horizons. One question might be, "How much will it cost to produce a given level of output *this year?*" Another might be, "How much will it cost us to produce a given level of output *three years from now and beyond?*" In this section, we'll explore managers' view of costs over a short-run time horizon—a time period during which *at least one* of the firm's inputs is fixed. That is, we'll be looking at costs with a *short-run* planning horizon.

Remember that no matter how much output is produced, the quantity of a fixed input *must* remain the same. Other inputs, by contrast, can be varied as output changes. Because the firm has these two different types of inputs in the short run, it will also face two different types of costs.

The costs of a firm's fixed inputs are called, not surprisingly, **fixed costs.** Like the fixed inputs themselves, fixed costs must remain the same no matter what the level of output. Typically, we treat rent and interest—whether explicit or implicit—as fixed costs, since producing more or less output in the short run will not cause these costs to change. Managers typically refer to fixed costs as their *overhead costs,* or simply, *overhead.*

The costs of obtaining the firm's variable inputs are its **variable costs.** These costs, like the usage of variable inputs themselves, will rise as output increases. Most businesses treat the wages of hourly employees and the costs of raw materials as variable costs, because quantities of labor and raw materials can usually be adjusted rather rapidly.

Measuring Short-Run Costs

In Table 3, we return to our example—Spotless Car Wash—and ask: What happens to costs as output changes in the short run? The first three columns of the table tell us the inputs Spotless will use for each output level, just as in Table 1 a few pages earlier. But there is one slight difference: In Table 3, we've reversed the order of the columns, putting total output first. We are changing our perspective slightly: Now we want to observe how a change in the quantity of *output* causes the firm's *inputs*—and therefore its *costs*—to change.

We also need to know one more thing before we can analyze Spotless's costs: what it must *pay* for its inputs. In Table 3, the price of labor is set at $120 per worker per day, and the price of each automated car-washing line at $150 per day.

How do Spotless's short-run costs change as its output changes? Get ready, because there are a surprising number of different ways to answer that question, as illustrated in the remaining columns of Table 3.

Fixed costs Costs of fixed inputs, which remain constant as output changes.

Variable costs Costs of variable inputs, which change with output.

Total Costs

Columns 4, 5, and 6 in the table show three different types of total costs. In column 4, we have Spotless's **total fixed cost** (***TFC***), the cost of all inputs that are fixed in the short run.

Total fixed cost The cost of all inputs that are fixed in the short run.

We'll assume that the cost of purchasing and installing an automated line is $912,500, and that the annual interest rate is 6%. So for one automated line, Spotless's owners sacrifice interest of .06 × $912,500 = $54,750 per year, or $150 per day. That is Spotless's total fixed cost per day. Running down the column, you can see that this cost—because it is fixed—remains the same no matter how many cars are washed each day.

Column 5 shows **total variable cost** (***TVC***), the cost of all variable inputs. For Spotless, labor is the only variable input. As output increases, more labor will be needed, so *TVC* will rise. For example, to wash 90 cars each day requires 2 workers, and each worker must be paid $120 per day, so *TVC* will be 2 × $120 = $240. But to wash 130 cars requires 3 workers, so *TVC* will rise to 3 × $120 = $360.

Total variable cost The cost of all variable inputs used in producing a particular level of output.

dangerous curves

Foregone Interest Can Be Tricky! Here are two common mistakes in calculating the implicit cost of funds you invest in your own business, such as the $100,000 in our restaurant example.

First, you might be tempted to count the entire $100,000 as a cost, rather than just the foregone interest on that sum. But a firm's costs are the *ongoing, yearly costs* for the owner. (That's what we'll eventually compare to the ongoing, annual revenue.) The $100,000 initial investment is only paid out once, not every year. If you sell the business, you will presumably get that sum back. But as long as you continue to own the business, the interest that *could* be earned on that $100,000 ($5,000 per year in our example) is an ongoing, yearly cost.

A second mistake is not realizing that, if conditions change, some of the foregone interest on your initial investment can become a *sunk cost.* For example, suppose you invest $100,000 to open a restaurant, and then the restaurant industry falls out of favor. If you sell now, you will only get $40,000. Then the ongoing cost of owning the business has just dropped to $2,000 per year—the interest foregone on $40,000 at 5 percent. That's the only part you could recover if you sold the restaurant, rather than continuing to own it. The interest you could have earned on the other $60,000 you originally invested no longer matters. It's a sunk cost because, regardless of any decision you make now or in the future, you can *not* get that $60,000—or those interest payments—back.

TABLE 3

Short-Run Costs for Spotless Car Wash

	Labor cost = $120 per day					Capital cost = $150 per day			
(1) Output (per Day)	(2) Capital	(3) Labor	(4) *TFC*	(5) *TVC*	(6) *TC*	(7) *MC*	(8) *AFC*	(9) *AVC*	(10) *ATC*
0	1	0	$150	$ 0	$150		—	—	—
						$ 4.00			
30	1	1	$150	$120	$270		$5.00	$4.00	$9.00
						$ 2.00			
90	1	2	$150	$240	$390		$1.67	$2.67	$4.33
						$ 3.00			
130	1	3	$150	$360	$510		$1.15	$2.77	$3.92
						$ 4.00			
160	1	4	$150	$480	$630		$0.94	$3.00	$3.94
						$ 5.00			
184	1	5	$150	$600	$750		$0.82	$3.26	$4.08
						$10.00			
196	1	6	$150	$720	$870		$0.77	$3.67	$4.44

Finally, column 6 shows us that

Total cost The costs of all inputs—fixed and variable.

> ***total cost*** **(TC)** *is the sum of all fixed and variable costs:*
>
> $$TC = TFC + TVC.$$

For example, at 90 units of output, *TFC* = $150 and *TVC* = $240, so *TC* = $150 + $240 = $390. Because total variable cost rises with output, total cost rises as well.

Now look at Figure 2, where we've graphed all three total cost curves for Spotless Car Wash. Both the *TC* and *TVC* curves slope upward, since these costs increase along with output. *TFC* is represented in two ways in the graph. One is the *TFC* curve, which is a horizontal line, since *TFC* has the same value at any level of output. The other is the *vertical distance* between the rising *TVC* and *TC* curves, since *TFC* is always the *difference* between *TVC* and *TC*. In the graph, this vertical distance must remain the same, at $150, no matter what the level of output.

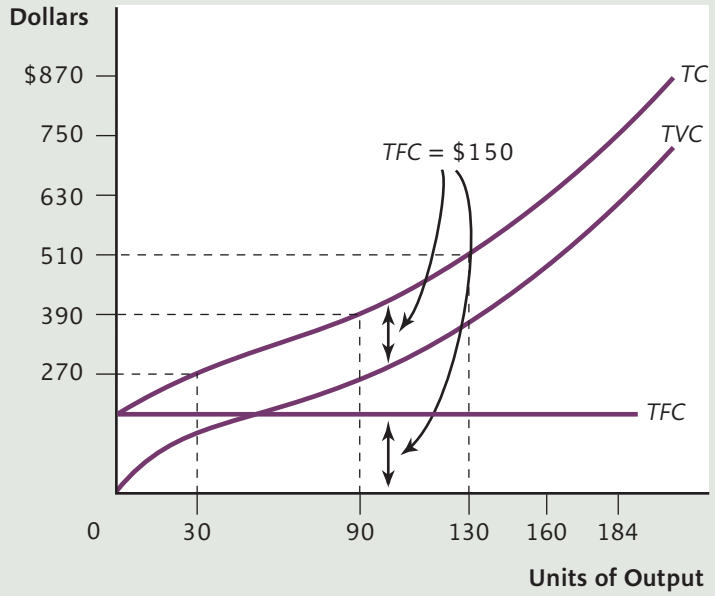
Average Costs

While total costs are important, sometimes it is more useful to track a firm's costs *per unit* of output, which we call its *average cost*. There are three different types of average cost, each obtained from one of the total cost concepts just discussed.

Average fixed cost Total fixed cost divided by the quantity of output produced.

> *The firm's* ***average fixed cost*** **(AFC)** *is its total fixed cost divided by the quantity* (Q) *of output:*
>
> $$AFC = \frac{TFC}{Q}.$$

FIGURE 2 The Firms Total Cost Curves



At any level of output, total cost (TC) *is the sum of total fixed cost* (TFC) *and total variable cost* (TVC).

No matter what kind of production or what kind of firm, *AFC* will always fall as output rises. Why? Because *TFC* remains constant, so a rise in *Q* *must* cause the ratio *TFC*/*Q* to fall.

Business managers often refer to this decline in *AFC* as "spreading their overhead" over more output. For example, a restaurant has overhead costs for its buildings, furniture, and cooking equipment. The more meals it serves, the lower will be its overhead cost per meal.

For Spotless Car Wash, look at column 8 of the table. When output is 30 units, *AFC* is $150/30 = $5.00. But at 90 units of output, *AFC* drops to $150/90 = $1.67. And *AFC* keeps declining as we continue down the column. The more output produced, the lower is fixed cost per unit of output.

Next is average *variable* cost.

Average variable cost **(AVC)** *is the cost of the variable inputs per unit of output:*

$$\text{AVC} = \frac{\text{TVC}}{\text{Q}}.$$

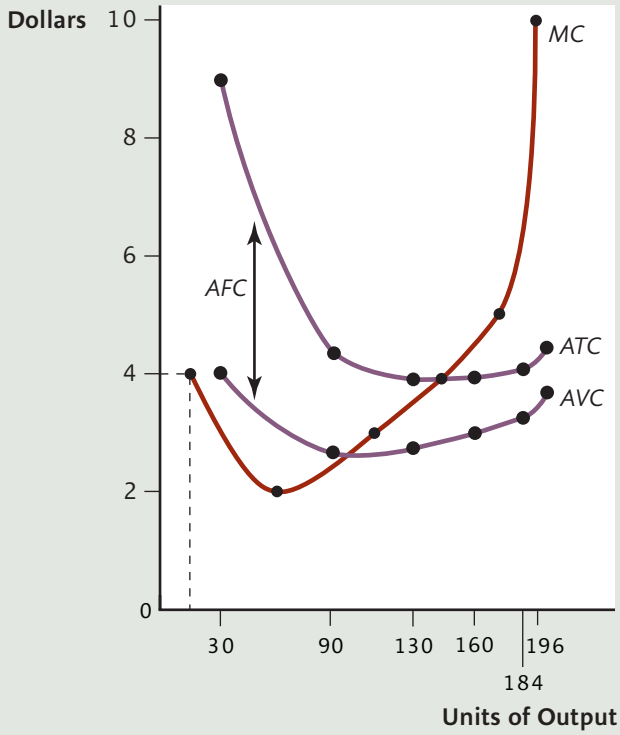
Average variable cost Total variable cost divided by the quantity of output produced.

AVC is shown in column 9 of the table. For example, at 30 units of output, *TVC* = $120, so *AVC* = *TVC*/*Q* = $120/30 = $4.00.

What happens to *AVC* as output rises? If you run your finger down the *AVC* column in Table 3, you'll see a pattern: The *AVC* numbers first decrease and then increase. Economists believe that this pattern of decreasing and then increasing average variable cost is typical at many firms. When plotted in Figure 3, this pattern causes the *AVC* curve to have a U shape. We'll discuss the reason for this characteristic U shape a bit later.

FIGURE 3 **Average and Marginal Costs**

Average variable cost (AVC) *and average total cost* (ATC) *are U-shaped, first decreasing and then increasing. Average fixed cost* (AFC), *the vertical distance between* ATC *and* AVC, *becomes smaller as output increases. The marginal cost* (MC) *curve is also U-shaped, reflecting first increasing and then diminishing marginal returns to labor.* MC *passes through the minimum points of both the* AVC *and* ATC *curves.*



The last average cost measure is average *total* cost.

Average total cost Total cost divided by the quantity of output produced.

Average total cost **(ATC)** *is the total cost per unit of output:*

$$\text{ATC} = \frac{\text{TC}}{\text{Q}}.$$

Values for *ATC* are listed in column 10 of Table 3. For example, at 90 units of output, *TC* = $390, so *ATC* = *TC*/*Q* = $390/90 = $4.33. And a quick look at column 10 shows that as output rises, *ATC* first falls and then rises. So the *ATC* curve—like the *AVC* curve—is U-shaped. However—as you can see in Figure 3—it is not identical to the *AVC* curve. At each level of output, the vertical distance between the two curves is equal to average *fixed* cost (*AFC*). Since *AFC* declines as output increases, the *ATC* curve and the *AVC* curve must get closer and closer together as we move rightward.

Marginal Cost

The total and average costs we've considered so far tell us about the firm's cost at a particular *level* of output. For many purposes, however, we are more interested in how cost *changes* when output *changes*. This information is provided by another cost concept:

Marginal cost **(MC)** *is the change in total cost* (ΔTC) *divided by the change in output* (ΔQ):

$$\text{MC} = \frac{\Delta \text{TC}}{\Delta \text{Q}}.$$

It tells us how much cost rises per unit increase in output.

Marginal cost The increase in total cost from producing one more unit of output.

For Spotless Car Wash, marginal cost is entered in column 7 of Table 3 and graphed in Figure 3. Since marginal cost tells us what happens to total cost when output *changes*, the entries in the table are placed *between* one output level and another. For example, when output rises from 0 to 30, total cost rises from $150 to $270. For this change in output, we have $\Delta TC = \$270 - \$150 = \$120$, while $\Delta Q = 30$, so $MC = \$120/30 = \$4.00$. This entry is listed *between* the output levels 0 and 30 in the table.

Explaining the Shape of the Marginal Cost Curve

Look at the graph of marginal cost in Figure 3. (For now, ignore the other two cost curves.) As in the table, each value of marginal cost is plotted *between* output levels. For example, the marginal cost of increasing output from 0 to 30 is $4, and this is plotted at output level 15—midway between 0 and 30. Similarly, when going from 30 to 90 units of output, the *MC* is plotted midway between 30 and 90.

If you look carefully at the *MC* curve in Figure 3, you'll see that *MC* first declines and then rises. Why is this? Here, we can use what we learned earlier about marginal returns to labor. At low levels of employment and output, there are increasing marginal returns to labor: $MPL = \Delta Q/\Delta L$ is rising. That is, each worker hired adds more to production than the worker before. But that means *fewer additional workers are needed to produce an additional unit of output*, so the *cost* of an additional unit of output (*MC*) must be falling. Thus, as long as *MPL* is rising, *MC* must be falling.

For Spotless, since *MPL* rises when employment increases from zero to one and again from one to two workers, *MC* must fall as the firm's output rises from zero to 30 units (produced by one worker) and then from 30 to 90 units (produced by two workers).

At higher levels of output, we have the opposite situation: Diminishing marginal returns set in and the marginal product of labor ($\Delta Q/\Delta L$) falls. Therefore, additional units of output require *more* and *more* additional labor. As a result, each additional unit of output costs more and more to produce. Thus, as long as *MPL* is falling, *MC* must be rising.

For Spotless, diminishing marginal returns to labor occur for all workers beyond the second, so *MC* rises for all increases in output beyond 90.

To sum up:

When the marginal product of labor (MPL) *rises, marginal cost* (MC) *falls. When* MPL *falls,* MC *rises. Since* MPL *ordinarily rises and then falls,* MC *will do the opposite: It will fall and then rise. Thus, the* MC *curve is U-shaped.*

The Relationship between Average and Marginal Costs

Although marginal cost and average cost are not the same, there is an important relationship between them. Look again at Figure 3 and notice that all three curves—*MC*, *AVC*, and *ATC*—first fall and then rise, but not all at the same time. The *MC* curve bottoms out before either the *AVC* or *ATC* curve. Further, the *MC* curve intersects each of the average curves *at their lowest points*. These graphical features of Figure 3 are no accident; indeed, they follow from the laws of mathematics. To understand this, let's consider a related example with which you are probably more familiar.

An Example: Average and Marginal Test Scores

Suppose you take five tests in your economics course during the term, with the results listed in Table 4. To your immense pleasure, you score 100 on your first test. Your total score—the total number of points you have received thus far during the term—is 100. Your marginal score—the *change* in your total caused by the most recent test—will also be 100, since your total rose from 0 to 100. Your average score so far is 100 as well.

Now suppose that, for the second test, you forget to study actively. Instead, you just read the text while simultaneously watching music videos and eavesdropping on your roommate's phone conversations. As a result, you get a 50, which is your marginal score. Since this score is lower than your previous average of 100, the second test will *pull your average down*. Indeed, whenever your score is lower than your previous average, it will pull down your average. In the table, we see that your average after the second test falls to 75.

Now you start to worry, so you turn off the TV while studying, and your performance improves a bit: You get a 60. Does the improvement in your score—from 50 to 60—increase your *average* score? No . . . your average will decrease once again, because your *marginal* score of 60 is *still* lower than your previous average of 75. As

TABLE 4

Average and Marginal Test Scores

Number of Tests Taken	Total Score	Marginal Score	Average Score
0	0		—
		100	
1	100		100
		50	
2	150		75
		60	
3	210		70
		70	
4	280		70
		80	
5	360		72

we know, when you score lower than your average, it pulls the average down, even if you're improving. In the table, we see that your average now falls to 70.

For your fourth exam, you study a bit harder and score a 70. This time, since your score is precisely *equal* to your previous average, the average remains unchanged at 70.

Finally, on your fifth and last test, your score improves once again, this time to 80. This time, you've scored *higher* than your previous average, pulling your average up from 70 to 72.

This example may be easy to understand because you are used to figuring out your average score in a course as you take additional exams. But the relationship between marginal and average spelled out here is universal: It is the same for grade point averages, batting averages—*and* costs.

Average and Marginal Cost

Now let's apply these insights to a firm's cost curves. We'll start with the relationship between the *MC* and *AVC* curves, because both curves reflect changes in the costs of variable inputs only. We already know that marginal cost first decreases and then increases. At low levels of output, as marginal cost decreases, it is *lower* than average variable cost, so it will pull the average down: *AVC* decreases. But then marginal cost rises (due to diminishing returns to labor). Eventually it rises above *AVC*, pulling the average up: *AVC* rises. Because *AVC* first decreases and then rises, the *AVC* curve is U-shaped.

The U-shape of the AVC *curve results from the U-shape of the* MC *curve, which in turn is based on increasing and then diminishing marginal returns to labor.*

There is a similar relationship between *MC* and *ATC*, except for one additional complication. *ATC* is the sum of *AVC* and *AFC*. *AFC always* falls as output rises. So at low levels of output, when both *AVC* and *AFC* are falling, *ATC* decreases—even more rapidly than AVC does. When *AVC* starts to rise, the rising *AVC* and falling *AFC* compete with each other. But eventually, the rise in *AVC* wins out, and *ATC* begins to rise as well. This explains why the *ATC* curve is U-shaped.

The U shape of the ATC *curve results from the behavior of both* AVC *and* AFC. *At low levels of output,* AVC *and* AFC *are both falling, so the* ATC *curve slopes downward. At higher levels of output, rising* AVC *overcomes falling* AFC*, and the* ATC *curve slopes upward.*

The relationships tell us something important about the crossing point between the *MC* curve and the average curves in Figure 3. Whenever the *MC* curve is below one of the average curves, the average curve slopes downward. Whenever the *MC* curve is above the average curve, the average curve slopes upward. Therefore, when *MC* goes from below the average to above the average—that is, where the *MC* curve *crosses* the average curve—the average must be at its very *minimum* (where it changes from a downward slope to an upward slope).

The MC *curve crosses both the* AVC *curve and the* ATC *curve at their respective minimum points.*

If you look at Table 3, you'll see that when Spotless's output rises from 30 to 90, *MC* is below *AVC*, and *AVC* falls. When output rises from 90 to 130, *MC* is above *AVC*, and *AVC* rises. As a result, in Figure 3, the *MC* curve crosses the *AVC* curve where *AVC* bottoms out. The same relationship holds for the *MC* and *ATC* curves. But because of the competing affects of *AFC* and *AVC* on *ATC*, it takes longer for the *ATC* curve to hit bottom than the *AVC* curve. That's why minimum *ATC* occurs at a higher output than does minimum *AVC*.

Time to Take a Break. By now, your mind may be swimming with concepts and terms: total, average, and marginal cost curves; fixed and variable costs; explicit and implicit costs. . . . We are covering a lot of ground here and still have a bit more to cover: production and cost in the *long run*.

As difficult as it may seem to keep these concepts straight, they will become increasingly easy to handle as you use them in the chapters to come. But it's best not to overload your brain with too much new material at one time. So if this is your first trip through this chapter, now is a good time for a break. Then, when you're fresh, come back and review the material you've read so far. When the terms and concepts start to feel familiar, you are ready to move on to the long run.

Production and Cost in the Long Run

Most of the business firms you have contact with—such as your supermarket, the stores where you buy clothes, your telephone company, and your Internet service provider—plan to be around for quite some time. They have a long-term planning horizon, as well as a short-term one. But so far, we've considered the behavior of costs only in the short run.

In the long run, costs behave differently, because the firm can adjust *all* of its inputs in any way it wants:

> *In the long run, there are no fixed inputs or fixed costs; all inputs and all costs are variable.*

How will the firm choose the inputs to use for any given output level? It will follow the *least cost rule:*

> *To produce any given level of output, the firm will choose the input mix with the lowest cost.*

Let's apply the least cost rule to Spotless Car Wash. Suppose we want to know the cost of washing 196 cars per day. In the short run, of course, Spotless does not have to worry about what input mix to use: It is stuck with one automated line, and if it wants to wash 196 cars, it must hire six workers (see Table 3). Total cost in the short run will be $6 \times \$120 + \$150 = \$870$.

In the long run, however, Spotless can vary the number of automated lines as well as the number of workers. Suppose, based on its production technology, Spotless can use four different input combinations to wash 196 cars per day. These are listed in Table 5. Combination *A* uses the least capital and the most labor—no automated lines at all and nine workers washing the cars by hand. Combination *D* uses the most capital and the least labor—three automated lines with only three workers.

TABLE 5

Four Ways to Wash 196 Cars per Day

Method	Quantity of Capital	Quantity of Labor	Cost
A	0	9	$1,080
B	1	6	$ 870
C	2	4	$ 780
D	3	3	$ 810

Since each automated line costs $150 per day and each worker costs $120 per day, it is easy to calculate the cost of each production method. Spotless will choose the one with the lowest cost: combination *C*, with two automated lines and four workers, for a total cost of $780 per day.

Retracing our steps, we have found that if Spotless wants to wash 196 cars per day, it will examine the different methods of doing so and select the one with the least cost. Once it has determined the cheapest production method, the other, more expensive methods can be ignored.[2]

dangerous curves

The Least Cost Rule When you read the *least cost rule* of production, you might think that the firm's long-run goal is to have the *lowest possible cost.* But that's not what the rule says. After all, in the long run, the lowest possible cost (zero) could be achieved by not using any inputs and producing nothing!

The least cost rule says something different: that any *given* level of output should be produced at the lowest possible cost for *that* output level.

Table 6 shows the results of going through this procedure for several different levels of output. The second column, **long-run total cost** (**LRTC**), tells us the cost of producing each quantity of output *when the least-cost input mix is chosen*. For each output level, different production methods are examined, the cheapest one is chosen, and the others are ignored.

Long-run total cost The cost of producing each quantity of output when all inputs are variable and the least-cost input mix is chosen.

Notice that the *LRTC* of zero units of output is $0. This will always be true for any firm. In the long run, all inputs can be adjusted as the firm wishes, and the cheapest way to produce zero output is to use *no* inputs at all. (For comparison, what is the *short*-run total cost of producing zero units? Why can it never be $0?)

The third column in Table 6 gives the **long-run average total cost** (**LRATC**), the cost per unit of output in the long run:

Long-run average total cost The cost per unit of producing each quantity of output in the long run, when all inputs are variable.

$$LRATC = \frac{LRTC}{Q}$$

Long-run average total cost is similar to average total cost, which was defined earlier. Both are obtained by dividing total cost by the level of output. There is one important difference, however: To calculate *ATC*, we used total cost (*TC*), which pertains to the short run, in the numerator. In calculating *LRATC*, we use *long-run* total cost (*LRTC*). Thus, *LRATC* tells us the cost per unit when the firm can vary *all* of its inputs and always chooses the cheapest input mix possible. *ATC*, however, tells us the cost per unit when the firm is stuck with some collection of fixed inputs and is able only to vary its remaining inputs, such as labor.

[2] The appendix to this chapter presents, in more detail, how firms choose the least-cost input mix when there is more than one variable input.

TABLE 6

Long-Run Costs for Spotless Car Wash

Output	*LRTC*	*LRATC*
0	$ 0	—
30	$ 200	$6.67
90	$ 390	$4.33
130	$ 510	$3.92
160	$ 608	$3.80
184	$ 720	$3.91
196	$ 780	$3.98
250	$1,300	$5.20
300	$2,400	$8.00

The Relationship between Long-Run and Short-Run Costs

If you compare Table 6 (long run) with Table 3 (short run), you will see something important: For some output levels, *LRTC* is smaller than *TC*. For example, Spotless can wash 196 cars for an *LRTC* of $780. But earlier, we saw that in the short run, the *TC* of washing these same 196 cars was $870. There is a reason for this difference.

Look back at Table 5, which lists the four different ways of washing 196 cars per day. In the short run, the firm is stuck with just one automated line, so its only option is method *B*. In the long run, however, the firm can choose any of the four methods of production, including method *C*, which is cheapest. The freedom to choose among different production methods usually enables the firm to select a cheaper input mix in the long run than it can in the short run. Thus, in the long run, the firm may be able to save money.

But not always. At some output levels, the freedom to adjust all inputs doesn't save the firm a dime. In our example, the long-run cost of washing 130 cars is $510—the same as the short-run cost (compare Tables 6 and 3). For this output level, it just so happens that the least-cost output mix uses one automated line, which is what Spotless is stuck with in the short run. So if Spotless wants to wash 130 cars, it cannot do so any more cheaply in the long run than in the short run.

More generally,

> *the long-run total cost of producing a given level of output can be less than or equal to, but not greater than, the short-run total cost (*LRTC ≤ TC).

We can also state this relationship in terms of *average* costs. That is, we can divide both sides of the inequality by *Q* and obtain $LRTC/Q \leq TC/Q$. Using our definitions, this translates to $LRATC \leq ATC$.

> *The long-run average cost of producing a given level of output can be less than or equal to, but not greater than, the short-run average total cost* (LRATC ≤ ATC).

Average Cost and Plant Size

Often, economists refer to the collection of inputs that are fixed in the short run as the firm's **plant.** For example, the plant of a computer manufacturer such as Dell might include its factory buildings and the assembly lines inside them. The plant of the Hertz car-rental company would include all of its automobiles and rental offices.

Plant The collection of fixed inputs at a firm's disposal.

For Spotless Car Wash, we've assumed that the plant is simply the company's capital equipment—the automated lines for washing cars. If Spotless were to add to its capital, then each time it acquired another automated line, it would have a different, and larger, plant. Viewed in this way, we can distinguish between the long run and the short run as follows: *In the long run, the firm can change the size of its plant; in the short run, it is stuck with its current plant.*

Now think about the *ATC* curve, which tells us the firm's average total cost in the short run. This curve is always drawn for a specific plant. That is, the *ATC* curve tells us how average cost behaves in the short run, *when the firm uses a plant of a given size*. If the firm had a different-size plant, it would be moving along a different *ATC* curve. In fact, there is a different *ATC* curve for each different plant the firm could have. In the long run, then, the firm can choose to operate on *any* of these *ATC* curves. To produce any level of output, it will always choose that *ATC* curve—among all of the *ATC* curves available—with the lowest possible average total cost. This insight tells us something about the relationship between the firm's *ATC* curves and its *LRATC* curve.

Graphing the *LRATC* Curve

Look at Figure 4, which shows several different *ATC* curves for Spotless Car Wash. There is a lot going on in this figure, so let's take it one step at a time. First, find the curve labeled ATC_1. This is our familiar *ATC* curve—the same one shown in Figure 3—which we used to find Spotless's average total cost in the short run, when it was stuck with one automated line.

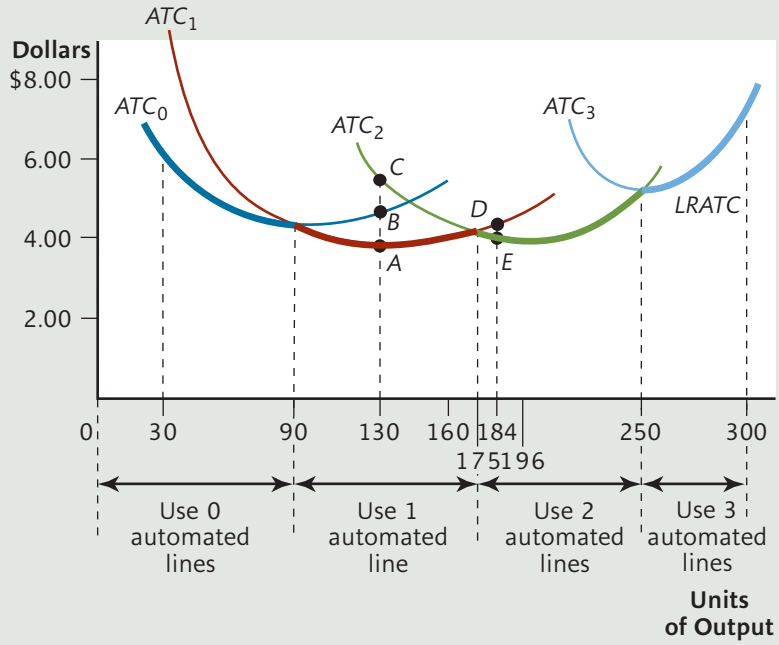
The other *ATC* curves refer to *different* plants that Spotless *might* have had instead. For example, the curve labeled ATC_0 shows how average total cost would behave if Spotless had a plant with *zero* automated lines washing all cars manually. ATC_2 shows average total cost with *two* automated lines, and so on. Since, in the long run, the firm can choose which size plant to operate, it can also choose on which of these *ATC* curves it wants to operate. And, as we know, in the long run, it will always choose the plant with the lowest possible average total cost for any output level it produces.

Let's take a specific example. Suppose that Spotless is planning to wash 130 cars per day. In the long run, what size plant should it choose? Scanning the different *ATC* curves in Figure 4, we see that the lowest possible per-unit cost—$3.92 per car—is at point *A* along ATC_1. The best plant for washing 130 cars per day, therefore, will have just one automated line. For this output level, Spotless would never choose a plant with zero lines, because it would then have to operate on ATC_0 at point *B*. Since point *B* is higher than point *A*, we know that point *B* represents a larger per-unit cost. Nor would the firm choose a plant with two lines—operating on ATC_2 at point *C*—for this would mean a still larger per-unit cost. Of all the possibilities for producing 130 units in the long run, Spotless would choose to operate at point *A* on ATC_1. So point *A* represents the *LRATC* of 130 units.

Now, suppose instead that Spotless wanted to produce 184 units of output in the long run. A plant with one automated line is no longer the best choice. Instead, the firm would choose a plant with *two* automated lines. How do we know? For an

FIGURE 4 Long-Run Average Total Cost

Average-total cost curves ATC_0, ATC_1, ATC_2, *and* ATC_3 *show average costs when the firm has zero, one, two, and three automated lines, respectively. The* LRATC *curve combines portions of all the firm's* ATC *curves. In the long run, the firm will choose the lowest-cost* ATC *curve for each level of output.*



output of 184, the firm could choose point *D* on ATC_1, or point *E* on ATC_2. Since point *E* is lower, it is the better choice. At this point, average total cost would be $1.96, so this would be the *LRATC* of 184 units.

Continuing in this way, we could find the *LRATC* for *every* output level Spotless might produce. To produce any given level of output, the firm will always operate on the *lowest ATC* curve available. As output increases, it will move along an *ATC* curve until another, lower *ATC* curve becomes available—one with lower costs. At that point, the firm will increase its plant size, so it can move to the lower *ATC* curve. In the graph, as Spotless increases its output level from 90 to 175 units of output, it will continue to use a plant with one automated line and move along ATC_1. But if it wants to produce *more* than 175 units in the long run, it will increase its plant to *two* automated lines and begin moving along ATC_2.

Thus, we can trace out Spotless's *LRATC* curve by combining just the lowest portions of all the *ATC* curves from which the firm can choose. In Figure 4, this is the thick, scallop-shaped curve.

A firm's LRATC *curve combines portions of each* ATC *curve available to the firm in the long run. For each output level, the firm will always choose to operate on the* ATC *curve with the lowest possible cost.*

Figure 4 also gives us a view of the different options facing the firm in the short run and the long run. Once Spotless builds a plant with one automated line, its options in the short run are limited: It can only move along ATC_1. If it wants to increase its output from 130 to 184 units, it must move from point *A* to point *D*. But in the long run, it can move along its *LRATC* curve—from point *A* to point *E*—by changing the size of its plant.

More generally,

in the short run, a firm can only move along its current ATC *curve. In the long run, however, it can move from one* ATC *curve to another by varying the size of its plant. As it does so, it will also be moving along its* LRATC *curve.*

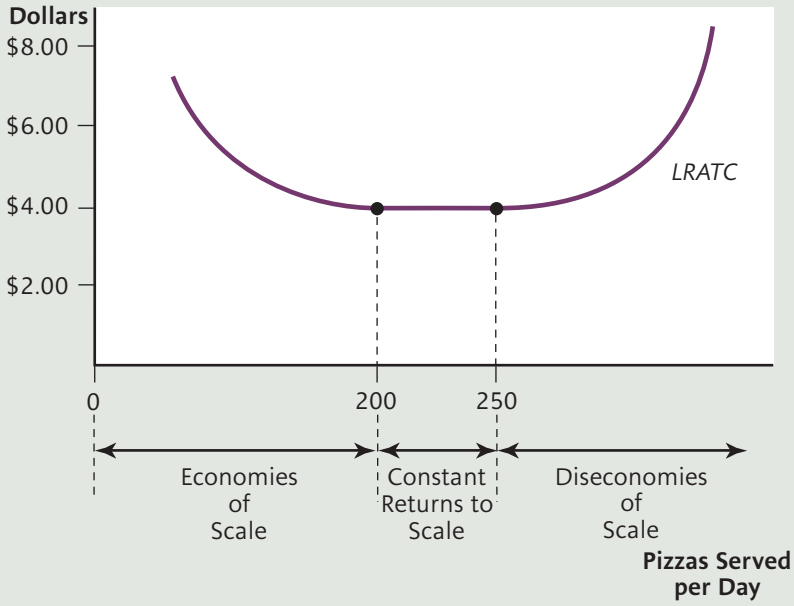
Explaining the Shape of the *LRATC* Curve

In Figure 4, the *LRATC* curve has a scalloped look because the firm can only choose among four different plants. But many firms can adjust their plant size in smaller increments. Each different plant size would be represented by a different *ATC* curve, so there would be hundreds of *ATC* curves crowded into the figure. As a result, the scallops would disappear, and the *LRATC* curve would appear as a smooth curve.

Figure 5 shows what the *LRATC* curve might look like for Mike's Pizza Restaurant. The horizontal axis measures the number of pizzas served per day. The vertical axis measures cost per pizza. Note that as we move along this curve, we are looking at *long-run* average total cost. In the long run, as output rises, not only can Mike's use more cooks, ingredients, and wait-staff, it can also adjust the size of its "plant"—its restaurant facility.

The *LRATC* curve for Mike's Pizza is U-shaped—much like the *AVC* and *ATC* curves you learned about earlier. That is, as output increases, long-run average costs first decline, then remain constant, and finally rise. Although there is no law or rule of logic that requires an *LRATC* curve to have all three of these phases, in many industries this seems to be the case. Let's see why, by considering each of the three phases in turn.

FIGURE 5 **The Shape of *LRATC***



If long-run total cost rises proportionately less than output, production reflects economies of scale, and LRATC *slopes downward. If cost rises proportionately more than output, there are diseconomies of scale, and* LRATC *slopes upward. Between those regions, cost and output rise proportionately, yielding constant returns to scale.*

Economies of Scale

Economies of scale Long-run average total cost decreases as output increases.

When an increase in output causes *LRATC* to decrease, we say that the firm is enjoying **economies of scale:** The more output produced, the lower the cost per unit. Mike's Pizza has economies of scale for all output levels up to 200.

On a purely mathematical level, economies of scale means that long-run total cost is rising by a smaller proportion than output. For example, if a doubling of output (Q) can be accomplished with less than a doubling of costs, then the ratio $LRTC/Q = LRATC$ will decline, and—voilà!—economies of scale.

> *When long-run total cost rises proportionately less than output, production is characterized by economies of scale, and the* LRATC *curve slopes downward.*

So much for the mathematics. But in the real world, *why* should total costs ever increase by a smaller proportion than output? Why should a firm experience economies of scale?

Gains from Specialization. One reason for economies of scale is gains from specialization. At very low levels of output, workers may have to perform a greater variety of tasks, slowing them down and making them less productive. But as output increases and workers are added, more possibilities for specialization are created. For example, at low levels of output, Mike's Pizza might have a very small facility with just one employee. This one worker would do everything himself: cook the pizzas, take orders, clean the tables, accept payments, order ingredients, and so on. But as output expands, Mike can run a larger operation with more workers, each specializing in one of these tasks. Since each worker is more productive, output will increase by a greater proportion than costs.

You've learned that increased specialization also plays a role in costs in the short run: it is one of the reasons why marginal cost (and therefore average costs) can decrease as output expands from low levels. But the ability of specialization to reduce costs is even greater in the long run. Remember that, in the short run, output expands by adding more and more variable inputs to unchanging amounts of fixed inputs. At some point, the fixed inputs cause diminishing returns to set in, overwhelming any further gains from specialization. In the long run, however, *all* inputs can be increased as output expands—including factory size, capital equipment, managers, and more. This opens up many more ways to re-arrange production to take full advantage of specialization.

The greatest opportunities for increased specialization occur when a firm is starting at a relatively low level of output, with a relatively small plant and small workforce. Thus, economies of scale are more likely to occur at lower levels of output.

Lumpy input An input whose quantity cannot be increased gradually as output increases, but must instead be adjusted in large jumps.

Spreading Costs of Lumpy Inputs. Another explanation for economies of scale involves the "lumpy" nature of many types of plant and equipment. **Lumpy inputs** are inputs that cannot be increased in tiny increments, but rather must be increased in large jumps. In some cases, a minimal amount of the inputs is needed to produce any output at all.

A medical practice, for example, needs the use of at least one X-ray machine in order to serve patients. And it must buy a *whole* machine, not a half or a fifth of an X-ray machine. The more patients the practice serves, the lower will be the cost of the machine per patient.

We see this phenomenon in many types of businesses: Plant and equipment must be purchased in large lumps, and a low cost per unit is achieved only at high levels of output. Other inputs besides equipment can also be lumpy in this way. A theater

must have at least one ticket taker and one projectionist, regardless of how many people come to see the show. A restaurant must pay a single license fee to the city each year, no matter how many meals it serves. In all of these cases, an increase in output allows the firm to spread the cost of lumpy inputs over greater amounts of output, lowering the cost *per unit of output*.

Spreading the costs of lumpy inputs has more impact on *LRATC* at low levels of output when these costs make up a greater proportion of the firm's total costs. At higher levels of output, the impact is smaller. For example, suppose Mike's restaurant must pay a yearly license fee of $3,650, which amounts to $10 per day. If output doubles from 10 to 20 pizzas per day, license costs per meal served will fall from $1 to $0.50. But if output doubles from 200 to 400, license costs per meal drop from $0.05 to $0.025—a hardly noticeable difference. Thus, spreading lumpy inputs across more output—like the gains from specialization—is more likely to create economies of scale at relatively low levels of output. This is another reason why the typical *LRATC* curve—as illustrated in Figure 5—slopes downward at relatively low levels of output.

Diseconomies of Scale

As output continues to increase, most firms will reach a point where bigness begins to cause problems. Large firms may require more layers of management, so communication and decision making become more time consuming and costly. Huge corporations like Ford, Microsoft, and Verizon each have several hundred high-level managers, and thousands more at lower levels.

Large firms may also have a harder time screening out misfits among new hires and monitoring those already working at the firm. This leads to more mistakes, shirking of responsibilities, and even theft from the firm. If Mike expands his facility so he can serve hundreds of pizzas per day, with dozens of employees, some of them might start sneaking pizzas home at the end of the day, others might take extra long breaks without anyone noticing, and so on. As output continues to rise and the firm has exhausted the cost-saving opportunities from increasing its scale of operations, these sorts of problems will eventually dominate, causing *LRATC* to rise.

When LRATC rises with an increase in output, we have **diseconomies of scale.** Mathematically,

Diseconomies of scale Long-run average total cost increases as output increases.

> *when long-run total cost rises more than in proportion to output, there are diseconomies of scale, and the* LRATC *curve slopes upward.*

While economies of scale are more likely at low levels of output, *dis*economies of scale are more likely at higher output levels. In Figure 6, Mike's Pizza does not experience diseconomies of scale until it is serving more than 250 pizzas per day.

Constant Returns to Scale

In Figure 5, for output levels between 200 and 250, the smoothed-out *LRATC* curve is roughly flat. Over this range of output, *LRATC* remains approximately constant as output increases. Here, output and *LRTC* rise by roughly the same proportion:

> *When both output and long-run total cost rise by the same proportion, production is characterized by* ***constant returns to scale****, and the* LRATC *curve is flat.*

Constant returns to scale Long-run average total cost is unchanged as output increases.

Why would a firm experience constant returns to scale? We have seen that as output increases, cost savings from specialization and spreading the costs of lumpy inputs will eventually be exhausted. But production may still have room to expand

before the costly problems of "bigness" kick in. The firm will then have a range of output over which average cost neither rises nor falls as production increases—constant returns to scale. Notice that constant returns to scale, if present at all, are most likely to occur at some *intermediate* range of output.

In sum, when we look at the behavior of *LRATC*, we often expect a pattern like the following: economies of scale (decreasing *LRATC*) at relatively low levels of output, constant returns to scale (constant *LRATC*) at some intermediate levels of output, and diseconomies of scale (increasing *LRATC*) at relatively high levels of output. This is why *LRATC* curves are typically U-shaped.

Of course, even U-shaped *LRATC* curves will have different appearances for firms in different industries. And as you're about to see, these differences in *LRATC* curves have much to tell us about the economy.

Cost: A Summary

This chapter has presented a number of new terms and concepts. As you first learn them, it's easy to get them confused. Table 7 provides a useful summary, which you can use both as a reference and a self-test.

TABLE 7

Types of Costs

Term	Symbol and/or Formula	Definition
Explicit cost		A cost where an actual payment is made
Implicit cost		An opportunity cost, but no actual payment is made
Sunk cost		An irrelevant cost because it cannot be affected by any current or future decision
Lumpy input cost		The cost of an input that can only be adjusted in large, indivisible amounts
Short-run costs		
Total fixed cost	***TFC***	The cost of all inputs that are fixed (cannot be adjusted) in the short run
Total variable cost	***TVC***	The cost of all inputs that are variable (can be adjusted) in the short run
Total cost	$TC = TFC + TVC$	The cost of all inputs in the short run
Average fixed cost	$AFC = TFC / Q$	The cost of all fixed inputs per unit of output
Average variable cost	$AVC = TVC/Q$	The cost of all variable inputs per unit of output
Average total cost	$ATC = TC/Q$	The cost of all inputs per unit of output
Marginal cost	$MC = \Delta TC/\Delta Q$	The change in total cost for each one-unit rise in output
Long-run costs		
Long-run total cost	***LRTC***	The cost of all inputs in the long run, using the least-cost method of producing any given output level
Long-run average	$LRATC = LRTC/Q$	Cost per unit in the long run, using the least-cost method of producing any given output level

Using the Theory

© BRAD SWONETZ/REDUX PICTURES

THE URGE TO MERGE

At the beginning of this chapter, we noted several large corporate mergers that took place during 2008 and 2009. Although there are many reasons for mergers like these, economies of scale often plays an important role.

To see why, look at Figure 6, which shows a hypothetical *LRATC* curve for a firm. Notice that this *LRATC* curve exhibits economies of scale (slopes downward) up to an output level of 20,000 units per month, and then diseconomies of scale set in (the curve slopes upward). The output level at which an *LRATC* curve like this first hits bottom is known as the **minimum efficient scale (MES)** for the firm—the lowest output level that allows it to achieve minimum cost per unit in the long run. At the MES, the gains from greater specialization and spreading the costs of lumpy inputs have been largely exhausted, but the problems of bigness haven't yet taken over. In the figure, if this firm were producing at its MES of 20,000 units, its long-run cost per unit would be $80, at point *B*. This is the lowest possible cost per unit this firm could achieve.

Minimum efficient scale The lowest output level at which the firm's *LRATC* curve hits bottom.

Now let's suppose that there are *six* firms selling in this market. We'll assume these firms are identical. Since each uses the same technology, and all pay the same prices for their inputs, they each have an *LRATC* curve just like the one in the figure.

Finally, we'll suppose that the output of the entire industry—at current prices—is 60,000 units per month, with each firm selling to one-sixth of this market. Thus, each firm operates at 10,000 units per month.

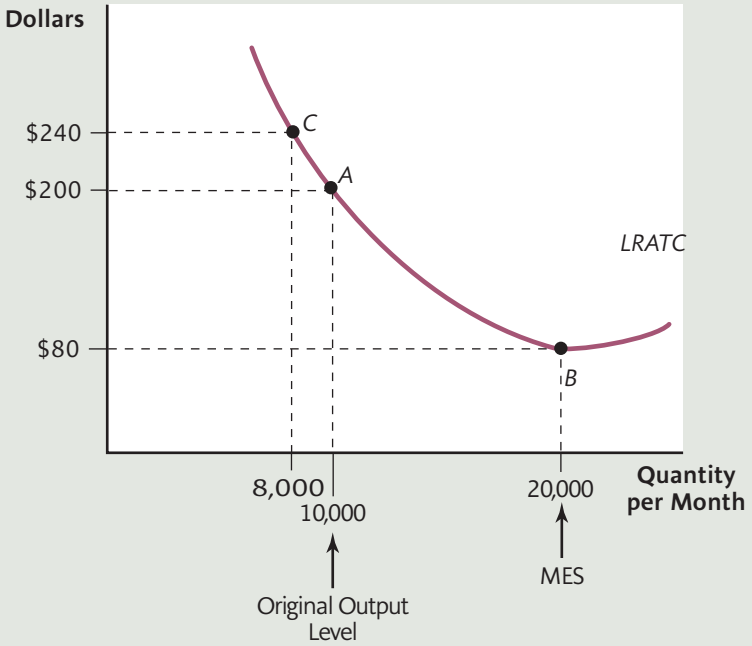
As you can see in the figure, with each firm producing 10,000 units (*less* than its MES), each is at point *A* on its *LRATC* curve. Cost per unit is $200—substantially above what cost per unit would be at the MES. There are unexploited economies of scale. We know that if this has been going on for a while, each firm must be charging *at least* $200; otherwise, it would not have been able to cover its costs and would have gone out of business.

Let's suppose (arbitrarily) that each of the six firms is charging a price of $220 per unit. (In coming chapters, you'll learn how the price is determined in different types of markets.) Since cost per unit is $200, each firm's total profit is $20 × 10,000 = $200,000 per month. All the firms are earning a profit, so you might think this situation could continue indefinitely.

But it likely won't. A market like this is ripe for mergers. Why? To keep our story as simple as possible, let's suppose for now that the total market demand for the output of these six firms is a constant 60,000 units per month, no matter what happens to the price. In this market, it won't be long before one of the firms in this industry gets a brilliant idea: to lower its price below that of its competitors, so it can take some of the market away from them. The first firm to do so can lower its price *below* $200 because, by doing so, it will increase the quantity it sells. As a

FIGURE 6 ***LRATC* for a Typical Firm in a Merger-Prone Industry**

With market quantity demanded fixed at 60,000, and six firms of equal market share, each operates at point A*, producing 10,000 units at $200 per unit. But any one firm can cut price slightly, increase market share, and operate with lower cost per unit, such as at the MES (point* B*). Other firms must match the first-mover's price; otherwise they lose market share and end up at a point like* C*, with higher cost per unit than originally. The result is a price war, with each firm ending up back at point* A*, only now—due to the lower price—they suffer losses. A series of mergers to create three large firms would enable each to operate at its MES (point* B*), with less likelihood of price wars and losses.*



result, it will slide down its *LRATC* curve and operate at a cost per unit under $200, and less than the cost per unit of its competitors.

Let's suppose that this first mover lowers the price to $190, which is just enough to increase its sales to the MES of 20,000 units. Cost per unit falls to $80 (point *B*), so profit per unit rises to $190 − $80 = $110. With 20,000 units sold, the first mover's total profit rises to $110 × 20,000 = $2,200,000. Not a bad move!

Of course, this will not make the other firms in the industry happy. Because we assume that total sales are fixed at 60,000 units per month, the gain in sales by the first mover come at the expense of its competitors, who (we assume for now) have not yet lowered their own prices. As a result of the first mover *gaining* 10,000 units in sales, each of the remaining five firms has *lost* 2,000 in sales—declining to 8,000. These firms therefore move *leftward* and *upward* along their own *LRATC* curves, to point *C*. Cost per unit for each of them is now $240.

These five slow-moving firms now have some unpleasant choices. If they *raise* their price above $240 to try to cover their costs on each unit, they will lose further sales, and slide further up their *LRATC* curve, with still higher cost. If they *lower* prices to get some of their sales back from the first mover, the result may be a price war, which could take the price down to $80, as the first mover tries to defend its market share so it can continue operating at its MES. (We'll have more to say about price wars and how they erupt in Chapter 10.)

But remember: It is impossible for *all* six firms to operate at their MES of 20,000 because we're assuming that the total demand for this product is fixed at 60,000. So if the price war leads all firms to charge a price of $80, with each firm having one-sixth of the market, then each firm, including the first mover, is once again

producing 10,000 units, at a cost of $200 per unit. Only now, with a lower price, each will suffer a loss.

We could continue this story, with all six firms deciding to raise their prices back to $220, until a price war breaks out again. But you can see that this market, with six firms, is very unstable. There are too many firms for each to satisfy the market demand of 60,000 while simultaneously operating at their MES. Price wars are likely to break out periodically, with periodic losses by all firms. If they try to make an illegal agreement not to lower their prices, in most countries (including the United States) they risk jail terms and huge fines.

Is there any way out of this mess? The title of this section gives the answer. If three of the six firms each combine with one of the others, the result will be three larger firms splitting the market among them. Each could then fully exploit its economies of scale, raising its output to the MES of 20,000, without overshooting the market's total demand of 60,000. While these firms *may* still compete for market share, the fact that each is operating at its MES means that each continues to operate at the lowest possible cost per unit merely by *maintaining* its proportional market share.

Of course, our story made several simplifications. Firms in an industry aren't really identical. And we assumed that demand was fixed at 60,000 units. In reality, we know that market demand curves have some elasticity, and that changes in price lead to changes in total market demand. We could incorporate more realistic assumptions, but they would complicate the analysis without changing its central point:

Economies of scale play an important role in determining the number of firms that survive in an industry. When there are significant, unexploited economies of scale (because the market has too many firms for each to operate near its minimum efficient scale), mergers often follow.

Economies of scale play a large role in explaining mergers in industries with high-cost lumpy inputs, such as large expenses for physical infrastructure, R&D, design, or marketing. By merging with other firms, duplicate lumpy inputs can be eliminated, and the costs of those that remain can be spread over more units of output, substantially lowering the new, larger firm's cost per unit. When InBev acquired Anheuser Busch, for example, it expected to save about $2 billion per year in costs, largely by combining and shrinking management and eliminating redundant marketing and manufacturing operations.

We'll have more to say about mergers and the behavior of large firms in future chapters. As you'll see, there are other motives for mergers that have nothing to do with economies of scale or costs. And even when a merger does bring down costs, it does not necessarily lead to lower prices for consumers. Indeed, economists become concerned when the number of firms in a market shrinks to just a few. This can mean less competition—and prices that are even higher than when the market had a greater number of higher cost firms.

SUMMARY

Business firms combine inputs to produce outputs. A firm's production *technology* determines the maximum output it can produce using different quantities of inputs. In the *short run,* at least one of the firm's inputs is fixed. In the *long run,* all inputs can be varied.

A firm's *cost of production* is the opportunity cost of its owners—everything they must give up in order to produce output. In the short run, some costs are *fixed* and independent of the level of production. Other costs—*variable costs*—can change as production changes. *Marginal cost* is the change in total cost from producing one more unit of output. The *marginal cost curve* has a U shape, reflecting the underlying marginal product of labor. A variety of average cost curves can be defined. The *average variable cost curve* and the *average total cost curve* are each U-shaped, reflecting the relationship between average and marginal cost the marginal cost curve must cross each of the average curves at their minimums.

In the long run, all costs are variable. The firm's *long-run total cost curve* indicates the cost of producing each quantity of output with the least-cost input mix. The related *long-run average total cost (LRATC) curve* is formed by combining portions of different *ATC* curves, each portion representing a different plant size. The *LRATC* curve slopes downward when there are economies of scale, slopes upward when there are diseconomies of scale, and is flat when there are constant returns to scale. Economies of scale can play a role in explaining mergers and acquisitions, especially when there are too many firms for each to operate at its *minimum efficient scale*.

PROBLEM SET

Answers to even-numbered Questions and Problems can be found on the text website at www.cengage.com/economics/hall.

1. The following table shows total output (in tax returns completed per day) of the accounting firm of Hoodwink and Finagle:

Number of Accountants	Number of Returns per Day
0	0
1	5
2	12
3	17
4	20
5	22

   Assuming the quantity of capital (computers, adding machines, desks, etc.) remains constant at all output levels:
   a. Calculate the marginal product of each accountant.
   b. Over what range of employment do you see increasing returns to labor? Diminishing returns?
   c. Explain why *MPL* might behave this way in the context of an accounting firm.

2. In mid-2009, the Obama administration announced it would cancel orders for a new fleet of presidential helicopters. About $3 billion had already been spent on developing the helicopters, which had special protective and telecommunications features. But another $8 billion would have been needed to complete the project and deliver the fleet. The administration suggested it might look for a less expensive design and start from scratch. Some media commentators criticized the decision, arguing that cancelling the project would mean wasting the $3 billion already spent.
   a. Suppose that starting from scratch on a new proposal that would be just as good as the original would cost a total of $5 billion from beginning to end. Which would be the wiser choice—sticking with the original or starting from scratch? Why?
   b. Would your answer change if the new proposal would cost a total of $10 billion? Why or why not?

3. Down On Our Luck Studios has spent $100 million producing an awful film, *A Depressing Story About a Miserable Person*. If the studio releases the film, the most cost-effective marketing plan would cost an additional $5 million, bringing the total amount spent to $105 million. Box office sales under this plan are predicted to be $12 million, which would be split evenly between the theaters and the studio. Additional studio revenue from video and DVD sales would be about $2 million. Should the studio release the film? If no, briefly explain why not. If yes, explain how it could make sense to release a film that cost $105 million but earns only $12 million.

4. The following table gives the short-run and long-run total costs for various levels of output of Consolidated National Acme, Inc.:

Q	TC_1	TC_2
0	0	350
1	300	400
2	400	435
3	465	465
4	495	505
5	540	560
6	600	635
7	700	735

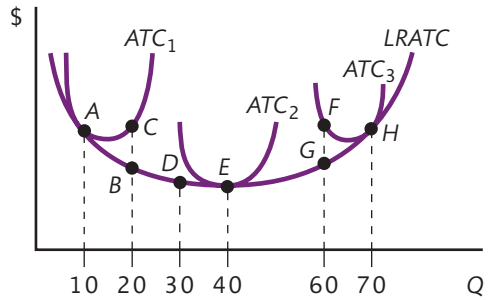
   a. Which column, TC_1 or TC_2, gives long-run total cost, and which gives short-run total cost? How do you know?
   b. For each level of output, find short-run *TFC, TVC, AFC, AVC,* and *MC.*
   c. At what output level would the firm's short-run and long-run input mixes be the same?
   d. Starting from producing two units, Consolidated's managers decide to double production to four units. So they simply double all of their inputs in the long run. Comment on their managerial skills.
   e. Over what range of output do you see economies of scale? Diseconomies of scale? Constant returns to scale?

5. In a recent year, a long, hard winter gave rise to stronger-than-normal demand for heating oil. The following summer was characterized by strong demand for gasoline by vacationers. Show what these two events might have done to the short-run *MC, AVC,* and *ATC* curves of Continental Airlines. (Hint: How would these events affect the price of oil?)

6. A study* of immunizing children in poor countries against diphtheria, pertussis, and tetanus estimated that, in the long run, a 10% increase in the number of children vaccinated increases the total cost of vaccinations by 8.4%. According to this study:
   a. Is immunization over this range characterized by economies of scale, constant returns to scale, or diseconomies of scale?
   b. [more difficult] We can define long-run marginal cost (LRMC) as the cost of increasing output by one unit when all inputs can be varied (as they can be in the long run). Based on the study, at current vaccine levels, would LRMC for vaccinations be greater than, less than, or equal to long-run average total cost (LRATC)? Why?
   c. If we want to know what it will cost to vaccinate *additional* children, and we use "cost per vaccine" as given by the current LRATC, do we overestimate, underestimate, or accurately estimate the cost per additional vaccine?

*David Bishal, Michael McQuestion, Rochika Chaudhry, and Alyssa Wigton, "The Costs of Scaling Up Vaccination in the World's Poorest Countries," Health Affairs, (Vol 25, No. 2) March/April 2006.

7. Ludmilla's House of Schnitzel is currently producing 10 schnitzels a day at point *A* on the following diagram. Ludmilla's business partner, Hans (an impatient sort), wants her to double production immediately.



   a. What point will likely illustrate Ludmilla's cost situation for the near future? Why?
   b. If Ludmilla wants to keep producing 20 schnitzels, at what point does she want to be eventually? How can she get there?
   c. Eventually, Ludmilla and company do very well, expanding until they find themselves making 70 schnitzels a day. But after a few years, Ludmilla discovers that profit was greater when she produced 20 schnitzels per day. She wants to scale back production to 20 schnitzels per day, laying off workers, selling off equipment, renting less space, and producing fewer schnitzels. Hans wants to reduce output by just cutting back on flour and milk and laying off workers. Who's right? Discuss the situation with reference to the relevant points on the diagram.
   d. Does the figure tell us what output Ludmilla should aim for? Why or why not?

8. Clean 'n' Shine is a competitor to Spotless Car Wash. Like Spotless, it must pay $150 per day for each automated line it uses. But Clean 'n' Shine has been able to tap into a lower-cost pool of labor, paying its workers only $100 per day. Clean 'n' Shine's production technology is given in the following table. To determine its short-run cost structure, fill in the blanks in the table.

Short-Run Costs for Clean 'n' Shine Car Wash

(1) Output (per Day)	(2) Capital	(3) Labor	(4) *TFC*	(5) *TVC*	(6) *TC*	(7) *MC*	(8) *AFC*	(9) *AVC*	(10) *ATC*
0	1	0	$___	$___	$___		—	—	—
						$___			
30	1	1	$___	$___	$___		$___	$___	$___
						$___			
70	1	2	$___	$___	$___		$___	$___	$___
						$___			
120	1	3	$___	$___	$___		$___	$___	$___
						$___			
160	1	4	$___	$___	$___		$___	$___	$___
						$___			
190	1	5	$___	$___	$___		$___	$___	$___
						$___			
210	1	6	$___	$___	$___		$___	$___	$___

a. Over what range of output does Clean 'n' Shine experience increasing marginal returns to labor? Over what range does it experience diminishing marginal returns to labor?
b. As output increases, do average fixed costs behave as described in the text? Explain.
c. As output increases, do marginal cost, average variable cost, and average total cost behave as described in the text? Explain.
d. Looking at the numbers in the table, but without drawing any curves, is the relationship between *MC* and *AVC* as described in the text? What about the relationship between *MC* and *ATC*?

9. In Table 3, when output rises from 90 to 130 units, marginal cost is $3.00. For this change in output, marginal cost is greater than the previous *AVC* ($2.67) but less than the previous *ATC* ($4.33). According to the relationship between marginals and averages you learned in this chapter:
   a. What should happen to *AVC* due to this change in output? Does it happen?
   b. What should happen to *ATC* due to this change in output? Does it happen?
10. A soft drink manufacturer that uses just labor (variable) and capital (fixed) paid a consulting firm thousands of dollars to calculate short-run costs at various output levels. But after the cost table (see below) was handed over to the president of the soft drink company, he spilled Dr Pepper on it, making some of the entries illegible. The consulting firm, playing tough, is demanding another payment to provide a duplicate table.

Output per Day	Units of Capital	Number of Workers	*TFC*	*TVC*	*TC*	*MC*	*AFC*	*AVC*	*ATC*
0	10	0	$1,000	?	?		?	?	?
							?		
20,000	10	100	?	$9,000	?		?	?	?
							?		
40,000	10	?	?	?	?		?	?	$0.325
							?		
60,000	10	225	?	?	?		?	?	?
							?		
80,000	10	?	?	?	$30,000		?	?	?

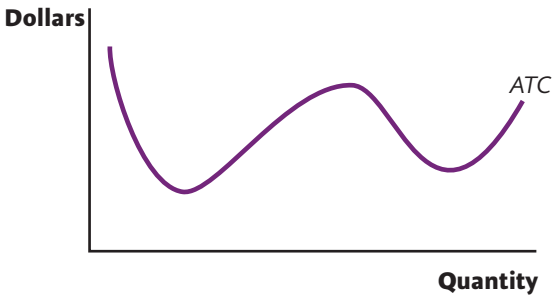
a. Should the soft drink president pay up? Or can he fill in the rest of the entries on his own? Fill in as many entries as you can to determine your answer. (Hint: First, determine the price of labor.)
b. Do *MC*, *AVC*, and *ATC* have the relationship to each other that you learned in this chapter? Explain.

11. "If a firm has diminishing returns to labor over some range of output, it cannot have economies of scale over that range." True or false? Explain briefly.

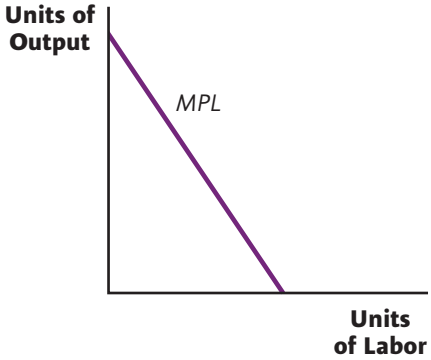
More Challenging

12. Draw the long-run total cost and long-run average cost curves for a firm that experiences:
    a. Constant returns to scale over all output levels.
    b. Diseconomies of scale over low levels of output, constant returns to scale over intermediate levels of output, and economies of scale over high output levels. Does this pattern of costs make sense? Why or why not?

13. A firm has the strange *ATC* curve drawn in the following figure. Sketch in the marginal cost curve this firm must have. (Hint: Use what you know about the marginal-average relationship. Note that this *MC* curve does *not* have a standard shape.)



14. The following curve shows the *marginal* product of labor for a firm at different levels of output.
    a. Show what the corresponding total product curve would look like.
    b. Do the total and marginal product curves for this firm ever exhibit diminishing marginal returns to labor? Increasing marginal returns to labor?



APPENDIX

Isoquant Analysis: Finding the Least-Cost Input Mix

When a firm can vary more than one input, it can usually choose to produce any given output level in different ways. For example, in the long run, a car wash can vary both its labor and the number of automated car washing lines it uses. If it acquires more automated lines, it can wash the same number of cars with less labor. Even in the short run, a firm can often vary more than one input. A farmer might be able to achieve the same total yield using less fertilizer if he hires more labor to care for and harvest the crop.

When a firm can choose among different methods of production, it will choose to produce any given level of output in the cheapest way possible. This appendix presents a graphical technique to show how a firm with two variable inputs finds this least-cost production method. The technique can be used over any time horizon—short run or long run—as long as there are two inputs whose quantities can be varied within that horizon. Finally, at the end of this appendix, the technique will be generalized to the case of *more than two* variable inputs.

Isoquants

Imagine that you own an artichoke farm, and you are free to vary two inputs: labor and land. Your output is measured in "boxes of artichokes per month." Your farm's production technology determines the maximum possible number of boxes you could produce in a given month using different combinations of labor and land. Alternatively, it tells us all the different input mixes that could be used to produce any given quantity of output.

Table A.1 lists some of the information we could obtain based on the technology of production on your farm. Notice that, to produce each of the three output levels included in the table, there are many different combinations of inputs you could use. For example, the table tells us that your farm could produce 4,000 boxes of artichokes using 2 hectares of land and 18 workers, or 3 hectares and 11 workers, or 5 hectares and 5 workers, and so on.

(Note: If it seems to you that no artichoke farm would ever use some of these combinations—such as 14 hectares of land and 2 workers—you are right. But that is because of the relative *costs* of labor and land, which we haven't discussed yet. Table A.1 simply tells us what is *possible* for the firm, not what is economically sensible.)

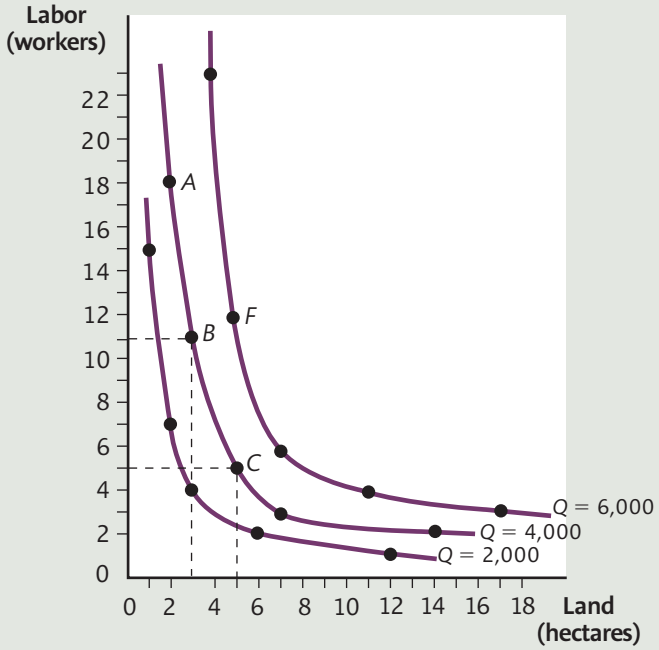
The information in the table can also be illustrated with a graph. In Figure A.1, the quantity of land is plotted along the horizontal axis, and the number of workers on the vertical axis. Each combination of the two inputs is represented by a point. For example, the combination *3 hectares, 11 workers* is represented by the point labeled *B*,

TABLE A.1

Production Technology for an Artichoke Farm

2,000 Boxes of Artichokes per Month		4,000 Boxes of Artichokes per Month		6,000 Boxes of Artichokes per Month	
Hectares of Land	*Number of Workers*	*Hectares of Land*	*Number of Workers*	*Hectares of Land*	*Number of Workers*
1	15	2	18	4	23
2	7	3	11	5	12
3	4	5	5	7	6
6	2	7	3	11	4
12	1	14	2	17	3

FIGURE A.1 An Isoquant Map



Each of the curves in the figure is an isoquant, showing all combinations of labor and land that can produce a given output level. The middle curve, for example, shows that 4,000 units of output can be produced with 11 workers and 3 hectares of land (point B), with 5 workers and 5 hectares of land (point C), as well as other combinations of labor and land. Each isoquant is drawn for a different level of output. The higher the isoquant line, the greater the level of output.

while the combination *5 hectares, 12 workers* is represented by point *F*.

Now let's focus on a single output level: 4,000 boxes per month. The middle columns of Table A.1 shows 5 of the different input combinations that can produce this output level, each represented by a point in Figure A.1. When we connect all 5 points with a smooth line we get the curve labeled "Q = 4,000" in Figure A.1. This curve is called an **isoquant**[3] ("iso" means "same," and "quant" stands for "quantity of output").

Every point on an isoquant represents an input mix that produces the same quantity of output.[4]

Figure A.1 also shows two additional isoquants. The higher one is drawn for the output level Q = 6,000, and the lower one for the output level Q = 2,000. When these curves are shown together on a graph, we have an **isoquant map** for the firm.

Things to Know About Isoquants

As we move along any isoquant, the quantity of output remains the same, but the combination of inputs changes. More specifically, as we move along an isoquant, we are *substituting one input for another.* For example, as we move from point *B* to point *C* along the isoquant labeled Q = 4,000, the quantity of land rises from 3 to 5 hectares, while the number of workers falls from 11 to 5. You are substituting land for labor, while maintaining the same level of output. Since each of the two inputs contributes to production, every time you increase one input, you must decrease the other in order to maintain the same level of output.

An increase in one input requires a decrease in the other input to keep total production unchanged. This is why isoquants always slope downward.

[3] Bolded terms in this appendix are defined in the glossary.

[4] If you've read the appendix to Chapter 5, you will recognize that isoquants are similar to indifference curves. But while an indifference curve represents different combinations of two *goods* that give same level of *consumer satisfaction,* an isoquant represents different combinations of two *inputs* that give the same level of *firm output*.

What happens as we move from isoquant to isoquant? Whenever we move to a higher *isoquant* (moving northeasterly in Figure A.1), the quantity of output increases. Moving directly northward means you are using more labor with the same amount of land, and moving directly eastward means you are using more land with the same amount of labor. When you move both north and east simultaneously (as in the move from point *B* to point *F*), you are using more of *both* inputs. For all of these movements, output increases. For the same reason, if we move southwestward, output decreases.

Higher isoquants represent greater levels of output than lower isoquants.

Finally, notice something else about Figure A.1: As we move rightward along any given isoquant, it becomes flatter. To understand why, we must take a closer look at an isoquant's slope.

The Marginal Rate of Technical Substitution

The (absolute value of the) slope of an isoquant is called the **marginal rate of technical substitution (MRTS).** As the name suggests, it measures the rate at which a firm can substitute one input for another while keeping output constant. In our example, if we use "L" for labor and "N" for land, the $MRTS_{L,N}$ tells us how many *fewer* workers you can employ each time you use *one more hectare of land*, and still maintain the same level of output.

For example, if you move from point *A* to point *B* along isoquant $Q = 4{,}000$, you use 1 more hectare and 7 fewer workers, so the $MRTS_{L,N} = 7/1 = 7$ for that move. Going from point *B* to point *C*, you use 2 more hectares of land, and 6 fewer workers, so the $MRTS_{L,N} = 6/2 = 3$.

Using this new term, the changing slope of an isoquant can be expressed this way:

as we move rightward along any given isoquant, the marginal rate of technical substitution decreases.

But why does the $MRTS_{L,N}$ decrease? To answer this question, it helps to understand the relationship between the *MRTS* and the *marginal products* of land and labor. You've already learned that the marginal product of labor (*MPL*) is a firm's additional output when one more worker is hired and all other inputs remain constant. The marginal product of land (*MPN*, using "N" for land) is defined in a similar way: It's the additional output a firm can produce with one additional unit of land (one more hectare, in our example), holding all other inputs constant.

Suppose that, starting from a given input mix, you discover that your *MPN* is 21 boxes of artichokes, and your *MPL* is 7 boxes. Then conduct the following mental experiment: Add one hectare of land, with no change in labor, and your output *increases* by 21 boxes. Then, give up 3 workers, with no change in land, and your output *decreases* by $3 \times 7 = 21$ boxes. In this case, adding 1 hectare of land, and hiring 3 fewer workers leaves your output unchanged. The slope of the isoquant for a move like this would be $\Delta L/\Delta N = -3/1 = -3$.

More generally, each time we change the amount of labor (L), the firm's output will change by $\Delta L \times MPL$. Each time we change the firm's land (N), the change in output will be $\Delta N \times MPN$. If we want the net result to be zero change in output, we must have

$$\Delta L \times MPL + \Delta N \times MPN = 0 \text{ or}$$
$$\Delta L \times MPL = -\Delta N \times MPN.$$

Rearranging this equation gives us:

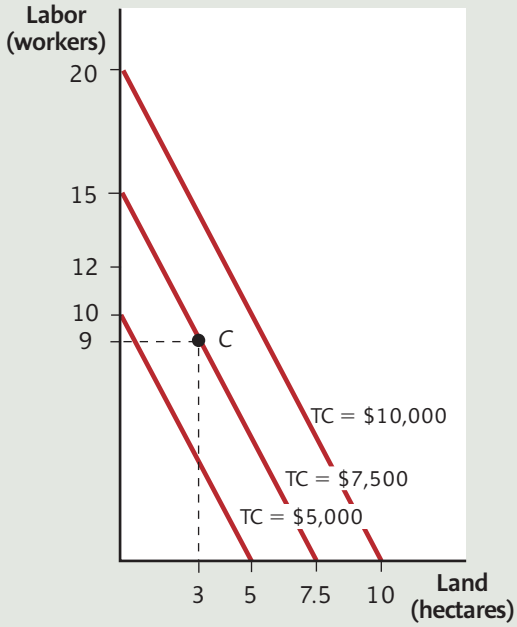
$$\Delta L/\Delta N = -MPN/MPL.$$

The left-hand side is the ratio of the change in labor to the change in land needed to keep output unchanged, that is, *the slope of the isoquant*. The right-hand side tells us that this slope is equal to the ratio of the marginal products of land and labor, except for the sign, which is negative. That is,

At each point along an isoquant with land measured horizontally, and labor measured vertically, the (absolute value of the) slope of the isoquant, which we call the $MRTS_{L,N}$*, is the ratio of the marginal products,* MPN/MPL.

Now, what does this have to do with the shape of the isoquant? As we move rightward and downward along an isoquant, the firm is acquiring more and more land, and using less and less labor. The marginal product of land will decrease—since land is becoming more plentiful—and the marginal product of labor will increase—since labor is becoming more and more scarce. Taken together, these changes tell us that the ratio *MPN*/*MPL* must fall and so must the slope of the isoquant.

FIGURE A.2 Isocost Lines



Each of the lines in the figure is an isocost line, showing all combinations of labor and land that have the same total cost. The middle line, for example, shows that total cost will be $7,500 if 9 workers and 3 hectares of land are used (point C). All other combinations of land and labor on the middle line have the same total cost of $7,500. Each isocost line is drawn for a different value of total cost. The higher the isocost line, the greater is total cost.

An isoquant becomes flatter as we move rightward because the MPN *decreases, while the* MPL *increases, so the ratio*—MPN/MPL—*decreases.*

Isocost Lines

An isoquant map shows us the different input mixes capable of producing different amounts of output. But how should the firm *choose* among all of these input mixes? In order to answer that question, we must know something about input *prices*. After all, if you own an artichoke farm, you must *pay* for your land and labor.

To keep the math simple, let's use round numbers. We'll suppose that the price of labor—the wage—is $500 per month ($P_L$ = $500), and the price of land—what you must pay in rent to its owner, or your implicit cost if you own the land yourself—is $1,000 per hectare per month (P_N = $1,000). An **isocost line** ("same cost" line) tells us all combinations of the two inputs that would require the same total outlay for the firm. It is very much like the *budget line* you learned about in Chapter 5, which showed all combinations of two *goods* that resulted in the same cost for the consumer. The difference is that an isocost line represents total cost to a *firm* rather than a consumer, and is based on paying for *inputs* rather than goods.

Figure A.2 shows three isocost lines for your artichoke farm. The middle line (labeled *TC* = $7,500) tells us all combinations of land and labor that would cost $7,500 per month. For example, point *G* represents the combination *3 hectares, 9 workers,* for a total cost of 3 × $1,000 + 9 × $500 = $7,500.

Things to Know about Isocost Lines

Notice that all three isocost lines in Figure A.2 *slope downward*. Why is this? As you move rightward in the figure, you are using more land. If you continued to use an unchanged amount of labor, your cost would therefore increase. But an isocost line shows us input combinations with the *same* cost. Thus, to keep your cost unchanged as you use more land (move rightward), you must also employ *fewer* workers (move downward).

If you use more of one input, you must use less of the other input in order to keep your total cost unchanged. This is why isocost lines always slope downward.

Notice, though, that the *slope* of the isocost line remains *constant* as we move along it. That is, isocost lines are *straight lines*. Why? Let's find an expression for the slope of the isocost line. Each time you change the number of workers by ΔL, your total cost will change by $P_L \times \Delta L$. Each time you change the amount of land you use by ΔN, your total cost will change by $P_N \times \Delta N$. In order for your total cost to remain the same as you change the amounts of both land and labor, the changes must satisfy the equation:

$$P_L \times \Delta L + P_N \times \Delta N = 0,$$

or

$$P_L \times \Delta L = -P_N \times \Delta N$$

which can be rearranged to

$$\Delta L/\Delta N = -P_N/P_L.$$

The term on the left is the change in labor divided by the change in land that leaves total cost unchanged—the slope of the isocost line. The term on the right is the (negative of the) ratio of the inputs' prices. In our example, with P_N = \$1,000 and P_L = \$500, the slope of the isocost line is −\$1,000/\$500 = −2.

Now you can see why the isocost line is a straight line: As long as the firm can continue to buy its inputs at unchanged prices, the ratio $-P_N/P_L$ will remain constant. Therefore, the slope of the isocost line will remain constant as well.

The slope of an isocost line with land (N) *on the horizontal axis and labor* (L) *on the vertical axis is* $-P_N/P_L$. *This slope remains constant as we move along the line.*

Finally, there is one more thing to note about isocost lines. As you move in a northeasterly direction in Figure A.2, to higher isocost lines, you are paying for greater amounts of land and labor, so your total cost must rise. For the same reason, as you move in a southwesterly direction, you are paying for smaller amounts of land and labor, so your total costs fall.

Higher isocost lines represent greater total costs for the firm than lower isocost lines.

In Figure A.2, the highest line represents all inputs combinations with a total cost of \$10,000, and the lowest line represents all combinations with a total cost of \$5,000.

The Least-Cost Input Combination

Now we are ready to combine what we know about a firm's production—represented by its isoquants—with our knowledge of the firm's costs—represented by its isocost lines. Together, these will allow us to find the least-cost input combination for producing any level of output a firm might choose to produce.

Suppose you want to know what is the best way to produce 4,000 boxes of artichokes per month. Figure A.3 reproduces the isoquant labeled Q = 4,000 from Figure A.1, along with the three isocost lines from Figure A.2. You would like to find the input combination that is *capable* of producing 4,000 boxes (an input combination *on the isoquant* Q = *4,000*), with the lowest possible cost (an input combination on the lowest possible isocost line). As you can see in the diagram, there is only one input combination that satisfies both requirements: point *C*. At this point, the firm uses 5 hectares of land, and 5 workers, for a total cost of 5 × \$1,000 + 5 × \$500 = \$7,500. As you can see, while there are other input combinations that can also produce 4,000 boxes, such as point *J* or point *K*, each of these lie on a higher isocost line (TC = \$10,000) and will require a greater total outlay than the least-cost combination at point *C*.

The least-cost combination will always be found where the isocost line is *tangent* to the isoquant. This is where the two lines touch each other at a single point, and both lines *have the same slope*.

The least-cost input combination for producing any level of output is found at the point where an isocost line is tangent to the isoquant for that output level.

This result will prove very useful. We already know that the slope of the isoquant at any point is equal to $-MPN/MPL$. And we know that the slope of the isocost line is equal to $-P_N/P_L$. Putting the two together, we know that when you have found the least-cost input combination for any output level,

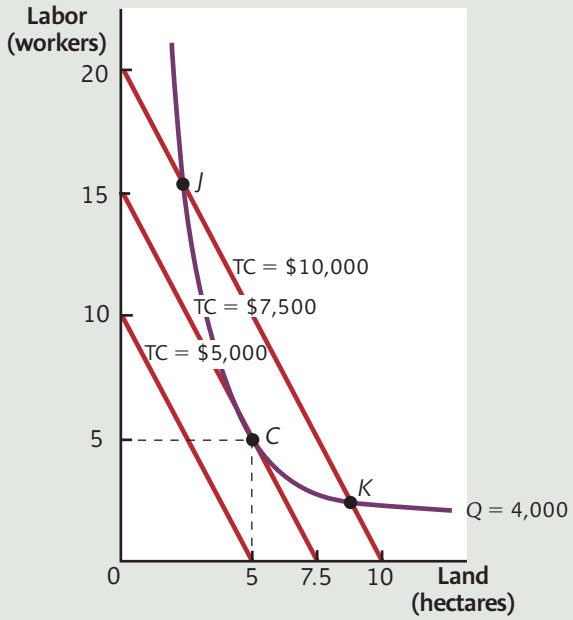
$$-MPN/MPL = -P_N/P_L$$

or

$$MPN/MPL = P_N/P_L.$$

The term on the left-hand side is just the $MRTS_{L,N}$—the marginal rate of technical substitution between labor and land. We conclude that:

FIGURE A.3 The Least-Cost Input Combination for a Given Output Level



To produce any given level of output at the least possible cost, the firm should use the input combination where the isoquant for that output level is tangent to an isocost line. In the figure, the input combinations at points J, C, *and* K *can all be used to produce 4,000 units of output. But the combination at point* C *(5 workers and 5 hectares of land), where the isoquant is tangent to the isocost line, is the least expensive input combination for that output level.*

When a firm is using the least-cost combination of two inputs (L and N) for a particular output level, the firm's MRTS *between the two inputs* (MPN/MPL) *will equal the ratio of input prices* (P_N/P_L).

In our example, P_N/P_L = \$1,000/\$500 = 2. This tells us that, at point *C*, the ratio $MPN/MPL = 2$ as well.

Finally, we can rearrange the equation $MPN/MPL = P_N/P_L$ to get:

$$MPN/P_N = MPL/P_L.$$

This form of the equation gives us another insight. It says that when you have found the least-cost input mix for any output level, the marginal product of land divided by the price of land will be equal to the marginal product of labor divided by the price of labor.

How can we interpret the marginal product of an input divided by its price? It gives us the additional output from spending one more *dollar* on the input. For example, if the (monthly) price of a hectare of land is \$1,000, and using one more hectare increases your output by 21 boxes ($MPN = 21$), then an additional *dollar* spent on land will give you 1/1,000 of a hectare, which, in turn, will increase your output by (1/1,000) × 21 = 21/1,000 or .021 boxes. So, MPN/P_N = 21/1,000 is the additional output from one more dollar spent on land. As a kind of shorthand, we'll call MPN/P_N the "marginal product per dollar" of land.

Using this language, we can state our result this way:

When a firm is using the least-cost combination of land and labor for any output level, the marginal product per dollar of land (MPN/P_N) *must equal the marginal product per dollar of labor* (MPL/P_L).

In the next section, where this result is stated more generally, you will learn the intuition behind it.

Generalizing to the Case of More than Two Inputs

When a firm can vary three or more inputs, we cannot illustrate isoquants and isocost lines on a two-dimensional graph. Nevertheless, the conclusions we reached for the two-input case can be generalized to any number of inputs.

Suppose a firm has several variable inputs, which we can label A, B, C, . . . , with marginal products MPA, MPB, MPC, . . . and input prices P_A, P_B, P_C, ... Then for any level of output, the least-cost combination of all of these inputs will always satisfy:

$$MPA/P_A = MPB/P_B = MPC/P_C = \ldots$$

That is,

> *When a firm with many variable inputs has found its least-cost input mix, the marginal product per dollar of any input will be equal to the marginal product per dollar of any other input.*

How do we know this must always be true? First, remember that MPA/P_A tells us the additional output the firm will produce *per additional dollar spent on input A*. Next, suppose we have two inputs, A and B, for which MPA/P_A is *not* equal to MPB/P_B. Then we can show that the firm can always shift its spending from one input to another, lowering its cost while leaving its output unchanged.

Let's take a specific example. Suppose that $MPA/P_A = 2$, and $MPB/P_B = 3$. Then the firm can easily save money by shifting dollars away from input A toward input B. Each dollar shifted away from A causes output to decrease by 2 units, while each dollar shifted toward input B causes output to rise by 3 units. Thus, the firm could shift dollars away from input A, and use only *some* of those dollars to increase the amount of input B, and still keep its production unchanged.

The same holds for any other two inputs we might compare: Whenever the marginal product per dollar is different for any two inputs, the firm can always shift its spending from the input with the lower marginal product per dollar to the input with the higher marginal product per dollar, achieving lower total cost with no change in output.

CHAPTER 8

How Firms Make Decisions: Profit Maximization

In July, 2008, Apple began selling its new iPhone 3G. The new model had all the features that made its original iPhone so popular, but it now came with much higher data speeds, more memory, and the ability to connect to the Internet via Wi-Fi.

But well before the launch date, Apple had several important decisions to make. From which companies should it buy the components? Where should the new product be assembled? How much should Apple spend on advertising? And finally: What price should it charge for the product, and what rate of production should it plan for?

These last decisions—what price to charge and how much to produce—are the focus of this chapter. In the end, Apple decided to charge $399 for its basic model[1], and—according to industry analysts—it was planning to produce and sell about 15 million phones during the product's first year. But why didn't Apple charge a lower price that would allow it to sell even more units? Or a higher price that would give it more profit on each unit sold?

Although this chapter concentrates on firms' decisions about price and output level, the tools you will learn apply to many other firm decisions. How much should MasterCard spend on advertising? How late should Starbucks keep its coffee shops open? How many copies should the *Wall Street Journal* give away free to potential subscribers? Should movie theaters offer Wednesday afternoon showings that only a few people attend? This chapter will help you understand how firms answer these sorts of questions.

The Goal of Profit Maximization

To analyze decision making at the firm, let's start with a very basic question: What is the firm trying to maximize?

Economists have given this question a lot of thought. Some firms—especially large ones—are complex institutions in which many different groups of people work together. A firm's owners will usually want the firm to earn as much profit as possible. But the workers and managers who actually run the firm may have other agendas. They may try to divert the firm away from profit maximization in order to benefit themselves. For now, let's assume that workers and managers are faithful servants of the firm's owners. That is,

> *We will view the firm as a single economic decision maker whose goal is to maximize its owners' profit.*

[1] If you bought an iPhone from AT&T (the exclusive U.S. carrier at the time), you may have paid less. But the discount was paid for by AT&T, not Apple, which still received its $399 (or more) for each phone.

Why do we make this assumption? Because it has proven so *useful* in understanding how firms behave. True, this assumption leaves out the details of these other agendas that often are present in real-world firms. But remember that every economic model *abstracts* from reality. To stay simple and comprehensible, it leaves out many real-world details and includes only what is relevant for the purpose at hand. If the purpose is to explain conflict within the firm or deviations from profit-maximizing behavior, then the differing goals of managers and owners should be a central element of the model. We'll consider situations like these in Chapter 15.

But when the purpose is to explain how firms decide what price to charge and how much to produce, or whether to temporarily shut down the firm or continue operating, or whether to enter a new market or permanently leave a current one, the assumption of profit maximization has proven to be very useful. It explains what firms actually do with reasonable—and sometimes remarkable—accuracy.

Why? Part of the reason is that managers who deviate *too* much from profit maximizing for *too* long are typically replaced. The managers may be sacked either by the current dissatisfied owners or by other firms that acquire the underperforming firm.

Another reason is that so many managers are well trained in the tools of profit maximization. This is in contrast to our model of consumer behavior, in which we asserted that consumers act *as if* they are using the model's graphs and calculations—although we recognize that most consumers never actually do. The basic economic model of the firm's behavior, however, is well understood *and used* by most managers, who have often taken several economics courses as part of their management education. In fact, economists' thinking about firm behavior has so permeated the language and culture of modern business that it's sometimes hard to distinguish where theory ends and practice begins.

Understanding Profit

Profit is defined as the firm's *sales revenue* minus its *costs of production*. There is widespread agreement over how to measure the firm's revenue—the flow of money into the firm. But there are two different conceptions of the firm's costs, and each of them leads to a different definition of profit.

Two Definitions of Profit

One conception of costs is the one used by accountants. With a few exceptions, accountants consider only *explicit* costs, where money is actually paid out.[2] If we deduct only the costs recognized by accountants, we get one definition of profit:

Accounting profit Total revenue minus accounting costs.

Accounting profit = Total revenue − Accounting costs.

But economics, as you have learned, has a much broader view of cost—*opportunity cost*. For the firm's owners, opportunity cost is the total value of *everything* sacrificed to produce output. This includes not only the explicit costs recognized by accountants—such as wages and salaries and outlays on raw materials—but also *implicit costs*, when something is given up but no money changes hands. For example, if an owner contributes his own time or money to the firm, there will be foregone wages or foregone investment income—both implicit costs for the firm.

[2] One exception is *depreciation*, a charge for the gradual wearing out of the firm's plant and equipment. Accountants include this as a cost even though no money is actually paid out.

This broader conception of costs leads to a second definition of profit:

Economic profit = Total revenue − *All* costs of production
= Total revenue − (Explicit costs + Implicit costs)

Economic profit Total revenue minus all costs of production, explicit and implicit.

The difference between economic profit and accounting profit is an important one; when they are confused, some serious (and costly) mistakes can result. An example might help make the difference clear.

Suppose you own a firm that produces T-shirts and you want to calculate your profit over the year. Your bookkeeper provides you with the following information:

Total Revenue from Selling T-shirts		**$300,000**
Cost of raw materials	$ 80,000	
Wages and salaries	150,000	
Electricity and phone	20,000	
Advertising cost	40,000	
Total Explicit Cost		**290,000**
Accounting Profit		**$ 10,000**

From the looks of things, your firm is earning a profit, so you might feel pretty good. Indeed, if you look only at money coming in and money going out, you have indeed earned a profit: $10,000 for the year . . . an accounting profit.

But suppose that in order to start your business you invested $100,000 of your own money—money that *could* be earning $6,000 in interest if you sold the business and got it back. Also, you are using two extra rooms in your own house as a factory—rooms that *could* be rented out for $4,000 per year. Finally, you are managing the business full-time, without receiving a separate salary, and you could instead be working at a job earning $40,000 per year. All of these costs—the interest, rent, and salary you *could* have earned—are implicit costs that have not been taken into account by your bookkeeper. They are part of the opportunity cost of your firm because they are sacrifices you made to operate your business.

Now let's look at this business from the economist's perspective and calculate your *economic* profit.

Total Revenue from Selling T-shirts		**$300,000**
Cost of raw materials	$ 80,000	
Wages and salaries	150,000	
Electricity and phone	20,000	
Advertising cost	40,000	
Total Explicit Costs	**$290,000**	
Investment income foregone	$ 6,000	
Rent foregone	4,000	
Salary foregone	40,000	
Total Implicit Costs	**$ 50,000**	
Total Costs		**$340,000**
Economic Profit		**−$ 40,000**

From an economic point of view, your business is not profitable at all, but is actually losing $40,000 per year! But wait—how can we say that your firm is suffering a loss when it takes in more money than it pays out? Because, as we've seen, your *opportunity cost*—the value of what you are giving up to produce your output—includes more than just money costs. When *all* costs are considered—implicit as well as explicit—your total revenue is not sufficient to cover what you have sacrificed to run your business. You would do better by shifting your time, your money, and your spare room to some alternative use.

Which of the two definitions of profit is the correct one? Either one of them, depending on the reason for measuring it. For tax purposes, the government is interested in profits as measured by accountants. The government cares only about the money you've earned, not what you *could* have earned had you done something else with your money or your time.

However, for our purposes—understanding the behavior of firms—economic profit is clearly better. Should your T-shirt factory stay in business? Should it expand or contract in the long run? Will other firms be attracted to the T-shirt industry? It is economic profit that will help us answer these questions, because it is economic profit that you and other owners care about.

The proper measure of profit for understanding and predicting the behavior of firms is economic profit. *Unlike accounting profit, economic profit recognizes all the opportunity costs of production—both explicit costs and implicit costs.*

Why Are There Profits?

When you look at the income received by households in the economy, you see a variety of payments. Those who provide firms with land receive *rent*—the payment for land. Those who provide labor receive a wage or salary. And those who lend firms money so they can purchase capital equipment receive interest. The firm's profit goes to its owners. But what do the owners of the firm provide that earns them this payment?

Economists view profit as a payment for two contributions of entrepreneurs, which are just as necessary for production as are land, labor, or machinery. These two contributions are *risk taking* and *innovation*.

Consider a restaurant that happens to be earning profit for its owner. The land, labor, and capital the restaurant uses to produce its meals did not simply come together magically. Someone—the owner—had to be willing to take the initiative to set up the business, and this individual assumed the risk that the business might fail and the initial investment be lost. Because the consequences of loss are so severe, the reward for success must be large in order to induce an entrepreneur to establish a business.

On a larger scale, when two Stanford students (Larry Page and Sergey Brin) started Google, they spent considerable time designing an effective search algorithm and planning their future company. As entrepreneurs, their contribution was *innovation*. But *other* entrepreneurs—the individuals and venture capital partners that provided funds to launch the new company—played the role of *risk-takers*. Had Google not been successful, they would have lost part or all of their investments. In hindsight, the fact that Google was such a good investment seems inevitable, as

successful investments often do. But in the early days, Google—like any startup—was a risky venture.

Innovation and risk taking can also be more subtle, and they are more common than you might think. When you pass by a successful laundromat, you may not give it a second thought. But someone, at some time, had to be the first one to realize, "I bet a laundromat in this neighborhood would do well"—an innovation. And someone had to risk the funds needed to get the business up and running.

The Firm's Constraints

If the firm were free to earn whatever level of profit it wanted, it would earn virtually infinite profit. This would make the owners very happy. Unfortunately for owners, though, the firm is not free to do this; it faces *constraints* on both its revenue and its costs.

The Demand Curve Facing the Firm

The constraint on the firm's revenue arises from a familiar concept: the demand curve. This curve always tells us the quantity of a good buyers wish to buy at different prices. But which buyers? And from which firms are they buying? Depending on how we answer these questions, we might be talking about different types of demand curves.

Market demand curves—like the ones you studied in Chapters 3 and 4—tell us the quantity demanded by *all* consumers from *all* firms in a market. In this chapter, we look at another kind of demand curve:

> *The **demand curve facing the firm** tells us, for different prices, the quantity of output that customers will choose to purchase from that firm.*

Demand curve facing the firm
A curve that indicates, for different prices, the quantity of output that customers will purchase from a particular firm.

Notice that this new demand curve—the demand curve facing the firm—refers to only *one* firm, and to *all buyers* who are potential customers of that firm.

Let's consider the demand curve faced by Ned, the owner and manager of Ned's Beds, a manufacturer of bed frames. Figure 1 lists the different prices that Ned could charge for each bed frame and the number of them (per day) he can sell at each price. The figure also shows a graph of the demand curve facing Ned's firm. For each price (on the vertical axis), it shows us the quantity of output the firm can sell (on the horizontal axis). Notice that, like the other types of demand curves we have studied, the demand curve facing the firm slopes downward. In order to sell more bed frames, Ned must lower his price.[3]

The definition of the demand curve facing the firm suggests that once it selects a price, the firm has also determined how much output it will sell. But, as you saw

[3] The downward-sloping demand curve tells us that Ned's Beds sells its output in an *imperfectly competitive market,* a market where the firm can *set* its price. Most firms operate in this type of market. If a manager thinks, "I'd like to sell more output, but then I'd have to lower my price, so let's see if it's worth it," we know he operates in an imperfectly competitive market. In a *perfectly competitive market,* by contrast, the firm would have to accept the market price as given. We assumed that markets were perfectly competitive in Chapter 3. In the next chapter, we'll examine perfect competition in more detail.

Like many other firms, a furniture manufacturer must determine either its price for its products, or its level of output; once it chooses price, its level of output is determined, and vice versa.

Total revenue The total inflow of receipts from selling a given amount of output.

a few chapters ago, we can also flip the demand relationship around: Once the firm has selected an output level, it has also determined the maximum price it can charge. This leads to an alternative definition:

> *The demand curve facing the firm shows us the maximum price the firm can charge to sell any given amount of output.*

Looking at Figure 1 from this perspective, we see that the horizontal axis shows alternative levels of output and the vertical axis shows the price Ned should charge if he wishes to sell each quantity of output.

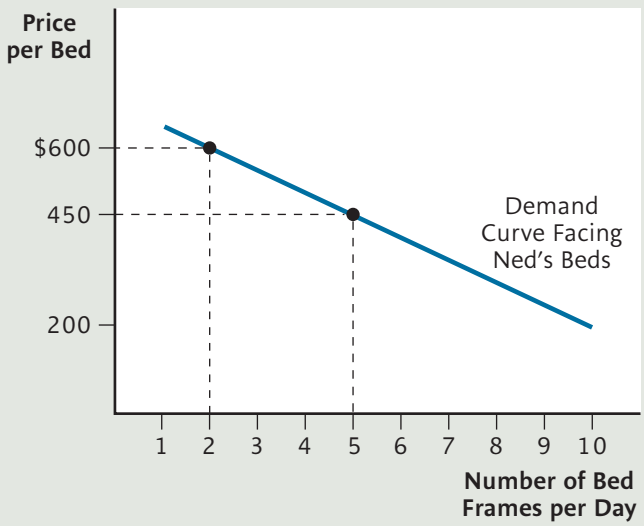
These two different ways of defining the firm's demand curve show us that it is, indeed, a constraint for the firm. The firm can freely determine *either* its price *or* its level of output. But once it makes the choice, the other variable is automatically determined by the firm's demand curve. Thus, the firm has only *one* choice to make. Selecting a particular price *implies* a level of output, and selecting an output level *implies* a particular price. Economists typically focus on the choice of output level, with the price implied as a consequence. We will follow that convention in this textbook.

Demand and Total Revenue

A firm's **total revenue** is the total inflow of receipts from selling output. Each time the firm chooses a level of output, it also determines its total revenue. Why? Because once we know the level of output, we also know the highest price the firm can charge. Total revenue, which is the number of units of output times the price per unit, follows automatically.

FIGURE 1 The Demand Curve Facing the Firm

(1) Price	(2) Output	(3) Total Revenue	(4) Total Cost	(5) Profit
>$650	0	0	$ 300	–$ 300
$650	1	$ 650	$ 700	–$ 50
$600	2	$1,200	$ 900	$ 300
$550	3	$1,650	$1,000	$ 650
$500	4	$2,000	$1,150	$ 850
$450	5	$2,250	$1,350	$ 900
$400	6	$2,400	$1,600	$ 800
$350	7	$2,450	$1,900	$ 550
$300	8	$2,400	$2,250	$ 150
$250	9	$2,250	$2,650	–$ 400
$200	10	$2,000	$3,100	–$1,100



The table presents information about Ned's Beds. Data from the first two columns are plotted in the figure to show the demand curve facing the firm. At any point along that demand curve, the product of price and quantity equals total revenue, which is given in the third column of the table.

The third column in Figure 1 lists the total revenue of Ned's Beds. Each entry is calculated by multiplying the quantity of output (column 2) by the price per unit (column 1). For example, if Ned's firm produces 2 bed frames per day, he can charge \$600 for each of them, so total revenue will be $2 \times \$600 = \$1{,}200$. If Ned increases output to 3 units, he must lower the price to \$550, earning a total revenue of $3 \times \$550 = \$1{,}650$. Because the firm's demand curve slopes downward, Ned must lower his price each time his output increases, or else he will not be able to sell all he produces. With more units of output, but each one selling at a lower price, total revenue could rise or fall. Scanning the total revenue column, we see that for this firm, total revenue first rises and then begins to fall. This will be discussed in greater detail later on.

The Cost Constraint

Every firm struggles to reduce costs, but there is a limit to how low costs can go. These limits impose a second constraint on the firm. Where do the limits come from? They come from concepts that you learned about in Chapter 6. Let's review them briefly.

First, the firm has a given production technology, which determines the different combinations of inputs the firm can use to produce its output.

Second, the firm must pay *prices* for each of the inputs that it uses, and we assume there is nothing the firm can do about those prices. Together, the firm's technology and the prices of the inputs determine the cheapest way to produce any given level of output. And once the firm finds this least-cost method, it has driven the cost of producing that output level as low as it can go.

The fourth column of Figure 1 lists Ned's total cost—the lowest possible cost of producing each quantity of output. More output always means greater costs, so the numbers in this column are always increasing. For example, at an output of zero, total cost is \$300. This tells us we are looking at costs in the short run, over which some of the firm's costs are *fixed*. (What would be the cost of producing 0 units if this were the long run?) If output increases from 0 to 1 bed frame, total cost rises from \$300 to \$700. This increase in total costs—\$400—is caused by an increase in *variable* costs, such as labor and raw materials.

The Profit-Maximizing Output Level

In this section, we ask a very simple question: How does a firm find the level of output that will earn it the greatest possible profit? We'll look at this question from several angles, each one giving us further insight into the behavior of the firm.

The Total Revenue and Total Cost Approach

At any given output level, the data in Figure 1 tell us (1) how much revenue the firm will earn and (2) its cost of production. We can then easily see how much profit the firm earns at each output level, which is the difference between total revenue (*TR*) and total cost (*TC*).

In the total revenue and total cost approach, we see the firm's profit as the difference between TC *and* TR *at each output level. The firm chooses the output level where profit is greatest.*

Let's see how this works for Ned's Beds. Column 5 of Figure 1 lists total profit at each output level. If the firm were to produce no bed frames at all, total revenue (*TR*) would be 0, while total cost (*TC*) would be $300. Total profit would be *TR* − *TC* = 0 − $300 = −$300. We would say that the firm earns a profit of negative $300 or a **loss** of $300 per day. Producing one bed frame would raise total revenue to $650 and total cost to $700, for a loss of $50. Not until the firm produces 2 bed frames does total revenue rise above total cost and the firm begin to make a profit. At 2 bed frames per day, *TR* is $1,200 and *TC* is $900, so the firm earns a profit of $300. Remember that as long as we have been careful to include *all* costs in *TC*—implicit as well as explicit—the profits and losses we are calculating are *economic* profits and losses.

Loss The difference between total cost (*TC*) and total revenue (*TR*), when *TC* > *TR*.

In the total revenue and total cost approach, locating the profit-maximizing output level is straightforward: We just scan the numbers in the profit column until we find the largest value, $900, and the output level at which it is achieved, 5 units per day. We conclude that the profit-maximizing output for Ned's Beds is 5 units per day.

The Marginal Revenue and Marginal Cost Approach

There is another way to find the profit-maximizing level of output. This approach, which uses *marginal* concepts, gives us some powerful insights into the firm's decision-making process. It is also closer to the trial-and-error procedure at some firms, in which small experimental changes are made to determine the impact on profit.

Recall that *marginal* cost is the *change* in total cost per unit increase in output. Now, let's consider a similar concept for revenue.

Marginal revenue The change in total revenue from producing one more unit of output.

Marginal revenue **(MR)** *is the change in the firm's total revenue* (ΔTR) *divided by the change in its output* (ΔQ)*:*

$$MR = \frac{\Delta TR}{\Delta Q}$$

MR *tells us how much revenue rises per unit increase in output.*

Table 1 reproduces the *TR* and *TC* columns from Figure 1, but adds columns for marginal revenue and marginal cost. (In the table, output is always changing by one unit, so we can use Δ*TR* alone as our measure of marginal revenue.) For example, when output changes from 2 to 3 units, total revenue rises from $1,200 to $1,650. For this output change, *MR* = $450. As usual, marginals are placed *between* different output levels because they tell us what happens as output *changes* from one level to another.

There are two important things to notice about marginal revenue. First, when *MR* is *positive,* an increase in output causes total revenue to *rise.* In the table, *MR* is positive for all increases in output from 0 to 7 units. When *MR* is *negative,* an increase in output causes total revenue to *fall,* as occurs for all increases beyond 7 units.

The second thing to notice about *MR* is a bit more complicated: Each time output increases, *MR* is *smaller* than the price the firm charges at the new output level. For example, when output increases from 2 to 3 units, the firm's total revenue rises by $450—even though it

dangerous curves

Maximize Profit, Not Revenue You may be tempted to forget about profit and think that the firm should produce where its total revenue is maximized. As you can see in Figure 1 (column 3), total revenue is greatest when the firm produces 7 units per day, but at this output level, profit is not as high as it could be. The firm does better by producing only 5 units. True, revenue is lower at 5 units, but so are costs. It is the difference between revenue and cost that matters, not revenue alone.

TABLE 1

More Data for Ned's Beds

Output	Total Revenue	Marginal Revenue	Total Cost	Marginal Cost	Profit
0	0		$ 300		−$ 300
		$650		$400	
1	$ 650		$ 700		−$ 50
		$550		$200	
2	$1,200		$ 900		$ 300
		$450		$100	
3	$1,650		$1,000		$ 650
		$350		$150	
4	$2,000		$1,150		$ 850
		$250		$200	
5	$2,250		$1,350		$ 900
		$150		$250	
6	$2,400		$1,600		$ 800
		$ 50		$300	
7	$2,450		$1,900		$ 550
		−$ 50		$350	
8	$2,400		$2,250		$ 150
		−$150		$400	
9	$2,250		$2,650		−$ 400
		−$250		$450	
10	$2,000		$3,100		−$1,100

sells the third unit for a price of $550. This may seem strange to you. After all, if the firm increases output from 2 to 3 units, and it gets $550 for the third unit of output, why doesn't its total revenue rise by $550?

The answer is found in the firm's downward-sloping demand curve, which tells us that to sell more output, the firm must cut its price. Look back at Figure 1 of this chapter. When output increases from 2 to 3 units, the firm must lower its price from $600 to $550. Moreover, the new price of $550 will apply to *all three* units the firm sells.[4] This means it *gains* some revenue—$550—by selling that third unit. But it also *loses* some revenue—$100—by having to lower the price by $50 on each of the two units of output it could have otherwise sold at $600. Marginal revenue will always equal the *difference* between this gain and loss in revenue—in this case, $550 − $100 = $450.

[4] Some firms can charge two or more different prices for the same product. We'll explore some examples in Chapter 9.

When a firm faces a downward-sloping demand curve, each increase in output causes a revenue gain, from selling additional output at the new price, and a revenue loss, from having to lower the price on all previous units of output. Marginal revenue is therefore less than the price of the last unit of output.[5]

Using *MR* and *MC* to Maximize Profits

Now we'll see how marginal revenue, together with marginal cost, can be used to find the profit-maximizing output level. The logic behind the *MC* and *MR* approach is this:

An increase in output will always raise profit as long as marginal revenue is greater than marginal cost (MR > MC).

Notice the word *always*. Let's see why this rather sweeping statement must be true. Table 1 tells us that when output rises from 2 to 3 units, *MR* is \$450, while *MC* is \$100. This change in output causes both total revenue and total cost to rise, but it causes revenue to rise by *more* than cost (\$450 > \$100). As a result, profit must increase. Indeed, looking at the profit column, we see that increasing output from 2 to 3 units *does* cause profit to increase, from \$300 to \$650.[6]

The converse of this statement is also true:

An increase in output will always lower profit whenever marginal revenue is less than marginal cost (MR < MC).

For example, when output rises from 5 to 6 units, *MR* is \$150, while *MC* is \$250. For this change in output, both total revenue and total cost rise, but cost rises *more,* so profit must go down. In Table 1, you can see that this change in output does indeed cause profit to decline, from \$900 to \$800.

These insights about *MR* and *MC* lead us to the following simple guideline the firm should use to find its profit-maximizing level of output:

To find the profit-maximizing output level, the firm should increase output whenever MR > MC, *and decrease output when* MR < MC.

Let's apply this rule to Ned's Beds. In Table 1 we see that when moving from 0 to 1 unit of output, *MR* is \$650, while *MC* is only \$400. Since *MR* is larger than *MC*, making this move will increase profit. Thus, if the firm is producing 0 beds, it should always increase to 1 bed. Should it stop there? Let's see. If it moves from 1 to 2 beds, *MR* is \$550, while *MC* is only \$200. Once again, *MR* > *MC,* so the

[5] There is a connection between the behavior of total revenue, marginal revenue, and the price elasticity of demand you learned about in Chapter 5. When demand for a specific firm's output is elastic, a fall in price—and the associated rise in quantity—causes total revenue to rise. In Table 1, total revenue rises (marginal revenue is positive) for all changes in output between 0 and 7 units. Therefore, we know that demand is elastic along the interval of the demand curve between 0 and 7 units. Similarly, demand for this firm's output is *inelastic* along the interval of the demand curve between 7 and 10 units.

[6] You may have noticed that the rise in profit (\$350) is equal to the difference between *MR* and *MC* in this example. This is no accident. *MR* tells us the *rise* in revenue; *MC* tells us the *rise* in cost. The difference between them will always be the *rise* in profit.

firm should increase to 2 beds. You can verify from the table that if the firm finds itself producing 0, 1, 2, 3, or 4 beds, $MR > MC$ for an increase of 1 unit, so it will always make greater profit by increasing production.

Until, that is, output reaches 5 beds. At this point, the picture changes: From 5 to 6 beds, *MR* is \$150, while *MC* is \$250. For this move, $MR > MC$, so profits would decrease. Thus, if the firm is producing 5 beds, it should *not* increase to 6. The same is true at every other output level beyond 5 units: The firm should *not* raise its output, since $MR < MC$ for each increase. We conclude that Ned maximizes his profit by producing 5 beds per day—the same answer we got using the *TR* and *TC* approach earlier.[7]

Profit Maximization Using Graphs

Both approaches to maximizing profit (using totals or using marginals) can be seen even more clearly when we use graphs. In Figure 2(a) and (b), the data from Table 1 have been plotted—the *TC* and *TR* curves in the upper panel, and the *MC* and *MR* curves in the lower one.

The marginal revenue curve has an important relationship to the total revenue curve. As you can see in Figure 2(a), total revenue (*TR*) is plotted on the vertical axis, and quantity (*Q*) on the horizontal axis, so the slope along any interval is just $\Delta TR/\Delta Q$. But this is exactly the definition of marginal revenue.

> *The marginal revenue for any change in output is equal to the* slope *of the total revenue curve along that interval.*

Thus, as long as the *MR* curve lies above the horizontal axis ($MR > 0$), *TR* must be increasing and the *TR* curve must slope upward. In the figure, $MR > 0$, and the *TR* curve slopes upward from zero to 7 units. When the *MR* curve dips below the horizontal axis ($MR < 0$), *TR* is decreasing, so the *TR* curve begins to slope downward. In the figure, this occurs beyond 7 units of output. As output increases in Figure 2, *MR* is first positive and then turns negative, so the *TR* curve will first *rise* and then *fall*.

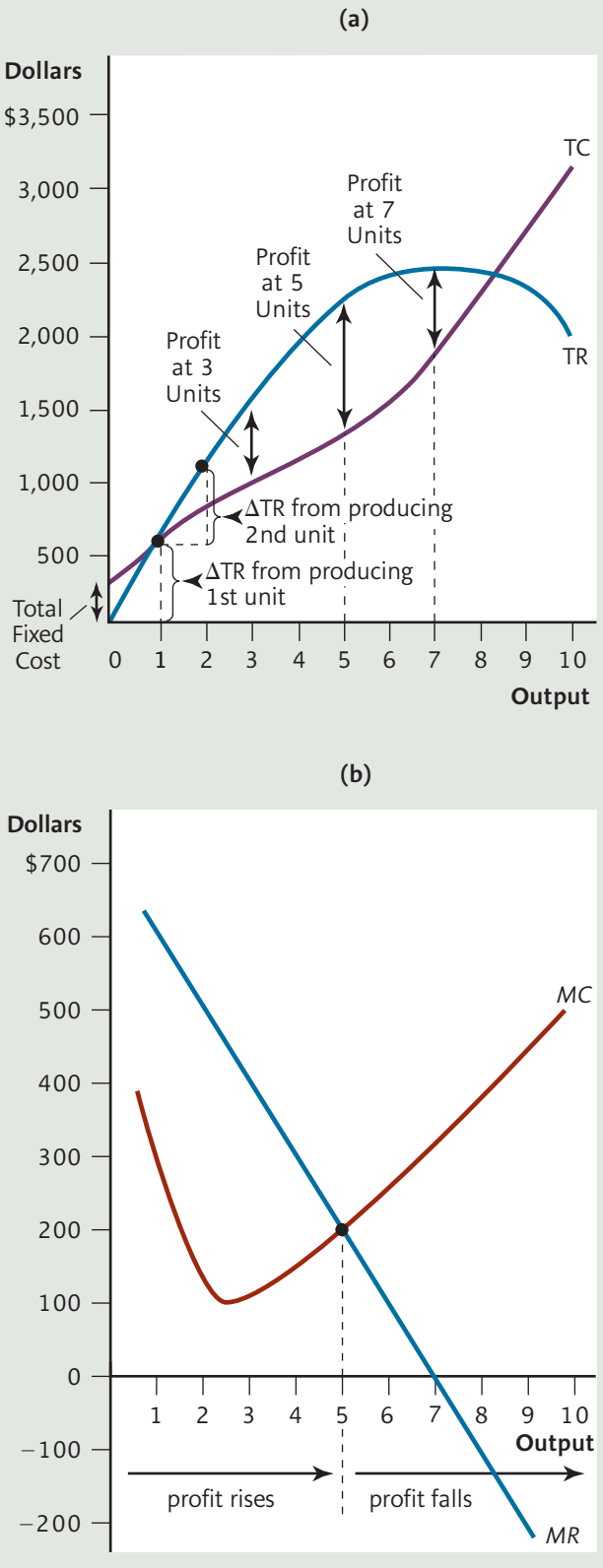
The *TR* and *TC* Approach Using Graphs

Now let's see how we can use the *TC* and *TR* curves to guide the firm to its profit-maximizing output level. We know that the firm earns a profit at any output level where $TR > TC$—where the *TR* curve lies *above* the *TC* curve. In Figure 2(a), you can see that all output levels from 2 through 8 units are profitable for the firm. The *amount* of profit is simply the *vertical distance* between the *TR* and *TC* curves, whenever the *TR* curve lies above the *TC* curve. Since the firm cannot sell part of a bed frame, it must choose whole numbers for its output, so the profit-maximizing output level is simply the whole-number quantity at which this vertical distance is greatest—5 units of output. Of course, the *TR* and *TC* curves in Figure 2 were plotted from the data in Table 1, so we should not be surprised to find the same profit-maximizing output level—5 units—that we found before when using the table.

[7] It sometimes happens that *MR* is precisely equal to *MC* for some change in output, although this does not occur in Table 1. In this case, increasing output would cause *both* cost and revenue to rise by equal amounts, so there would be *no* change in profit. The firm should not care whether it makes this change in output or not.

FIGURE 2 **Profit Maximization**

Panel (a) shows the firm's total revenue (TR) *and total cost* (TC) *curves. Profit is the vertical distance between the two curves at any level of output. Profit is maximized when that vertical distance is greatest—at 5 units of output. Panel (b) shows the firm's marginal revenue* (MR) *and marginal cost* (MC) *curves. (As long as* MR *lies above the horizontal axis, the* TR *curve slopes upward.) Profit is maximized at the level of output closest to where the two curves cross—at 5 units of output.*



We can sum up our graphical rule for using the *TR* and *TC* curves this way:

To maximize profit, the firm should produce the quantity of output where the vertical distance between the TR *and* TC *curve is greatest and the* TR *curve lies above the* TC *curve.*

The *MR* and *MC* Approach Using Graphs

Figure 2 also illustrates the *MR* and *MC* approach to maximizing profits. As usual, the marginal data in panel (b) are plotted *between* output levels, since they tell us what happens as output changes from one level to another.

In the diagram, as long as output is less than 5 units, the *MR* curve lies above the *MC* curve ($MR > MC$), so the firm should produce more. For example, if we consider the move from 4 to 5 units, we compare the *MR* and *MC* curves at the midpoint between 4 and 5. Here, the *MR* curve lies above the *MC* curve, so increasing output from 4 to 5 will increase profit.

But now suppose the firm is producing 5 units and considering a move to 6. At the midpoint between 5 and 6 units, the *MR* curve has already crossed the *MC* curve, and now it lies *below* the *MC* curve. For this move, $MR < MC$, so raising output would *decrease* the firm's profit. The same is true for every increase in output beyond 5 units: The *MR* curve always lies below the *MC* curve, so the firm will decrease its profits by increasing output. Once again, we find that the profit-maximizing output level for the firm is 5 units.

Notice that the profit-maximizing output level—5 units—is the level closest to where the *MC* and *MR* curves cross. This is no accident. For each change in output that *increases* profit, the *MR* curve will lie above the *MC* curve. The first time that an output change *decreases* profit, the *MR* curve will cross the *MC* curve and dip below it. Thus, the *MC* and *MR* curves will always cross closest to the profit-maximizing output level.

With this graphical insight, we can summarize the *MC* and *MR* approach this way:

To maximize profit, the firm should produce the quantity of output closest to the point where MC = MR*—that is, the quantity of output at which the* MC *and* MR *curves intersect.*

This rule is very useful, since it allows us to look at a diagram of *MC* and *MR* curves and *immediately* identify the profit-maximizing output level. In this text, you will often see this rule. When you read, "The profit-maximizing output level is where *MC* equals *MR*," translate to "The profit-maximizing output level is closest to the point where the *MC* curve crosses the *MR* curve."

A Proviso. There is, however, one important exception to this rule. Sometimes the *MC* and *MR* curves cross at two different points. In this case, the profit-maximizing output level is the one at which the *MC* curve crosses the *MR* curve *from below*.

dangerous curves

Misusing the Gap Between *MR* AND *MC* A common error is to think a firm should produce the level of output at which the difference between *MR* and *MC* is as large as possible, like 2 or 3 units of output in Figure 2. Let's see why this is wrong. If the firm produces 2 or 3 units, it would leave many profitable increases in output unexploited—increases where $MR > MC$. As long as *MR* is even a tiny bit larger than *MC*, it pays to increase output, since doing so will add more to revenue than to cost. The firm should be satisfied only when the difference between *MR* and *MC* is as *small* as possible, not as *large* as possible.

FIGURE 3 Two Points of Intersection

Sometimes the MR *and* MC *curves intersect twice. The profit-maximizing level of output is always found where* MC *crosses* MR *from below.*

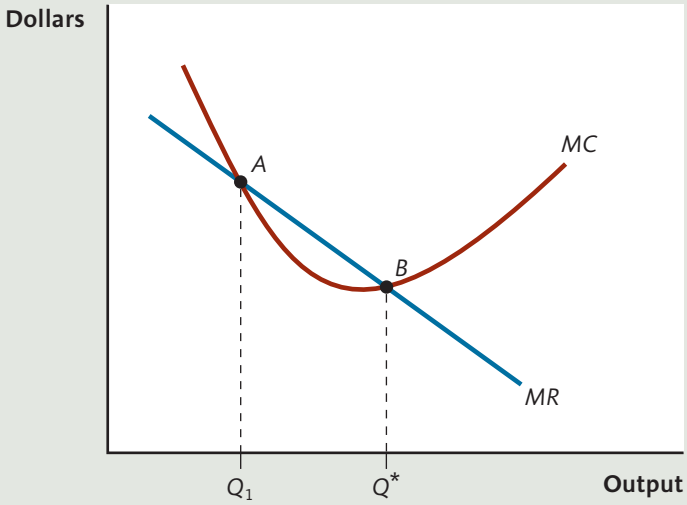


Figure 3 shows why. At point *A*, the *MC* curve crosses the *MR* curve from *above*. Our rule tells us that the output level at this point, Q_1, is *not* profit maximizing. Why not? Because at output levels lower than Q_1, $MC > MR$, so profit *falls* as output increases toward Q_1. Also, profit *rises* as output increases *beyond* Q_1, since $MR > MC$ for these moves. Since it never pays to increase *to* Q_1, and profit rises when increasing *from* Q_1, we know that Q_1 cannot possibly maximize the firm's profit.

But now look at point *B*, where the *MC* curve crosses the *MR* curve from below. You can see that when we are at an output level lower than Q^*, it always pays to increase output, since $MR > MC$ for these moves. You can also see that, once we have arrived at Q^*, further increases will reduce profit, since $MC > MR$. Q^* is thus the profit-maximizing output level for this firm—the output level at which the *MC* curve crosses the *MR* curve *from below*.

What about Average Costs?

You may have noticed that this chapter has discussed *most* of the cost concepts introduced in Chapter 7. But it has not yet referred to *average* cost. There is a good reason for this. We have been concerned about how much the firm should produce if it wishes to earn the greatest possible level of profit. To achieve this goal, the firm should produce more output whenever doing so *increases* profit, and it needs to know only *marginal* cost and *marginal* revenue for this purpose. The different types of average cost (*ATC, AVC,* and *AFC*) are simply irrelevant. Indeed, a common error—sometimes made even by business managers—is to use *average* cost in place of *marginal* cost in making decisions.

For example, suppose a yacht maker wants to know how much his total cost will rise in the short run if he produces another unit of output. It is tempting—*but wrong*—for the yacht maker to reason this way: "My cost per unit (*ATC*) is currently $50,000 per yacht. Therefore, if I increase production by 1 unit, my total cost

will rise by $50,000; if I increase production by 2 units, my total cost will rise by $100,000, and so on."

There are two problems with this approach. First, *ATC* includes many costs that are *fixed* in the short run—including the cost of all fixed inputs such as the factory and equipment and the design staff. These costs will *not* increase when additional yachts are produced, and they are therefore irrelevant to the firm's decision making in the short run.

Second, *ATC changes* as output increases. The cost per yacht may rise above $50,000 or fall below $50,000, depending on whether the *ATC* curve is upward or downward sloping at the current production level. Note that the first problem—fixed costs—could be solved by using *AVC* instead of *ATC*. The second problem—changes in average cost—remains even when *AVC* is used.

The correct approach, as we've seen in this chapter, is to use the *marginal cost* of a yacht and to consider increases in output one unit at a time. The firm should produce the output level where its *MC* curve crosses its *MR* curve from below. Average cost doesn't help at all; it only confuses the issue.

Does this mean that all of your efforts to master *ATC* and *AVC*—their definitions, their relationship to each other, and their relationship to *MC*—were a waste of time? Far from it. As you'll see, average cost will prove *very* useful in the chapters to come. You'll learn that whereas marginal values tell the firm *what* to do, averages can tell the firm *how well* it has done. But average cost should *not* be used in place of marginal cost as a basis for decisions.

The Marginal Approach to Profit

The *MC* and *MR* approach for finding the profit-maximizing output level is actually a very specific application of a more general principle:

*The **marginal approach to profit** states that a firm should take any action that adds more to its revenue than to its costs.*

Marginal approach to profit A firm maximizes its profit by taking any action that adds more to its revenue than to its cost.

In this chapter, the action being considered is whether to increase output by 1 unit. We've learned that the firm should take this action whenever $MR > MC$.

But the same logic can be applied to *any other decision* facing the firm. Should a restaurant owner take out an ad in the local newspaper? Should a convenience store that currently closes at midnight stay open 24 hours instead? Should a private kindergarten hire another teacher? Should an inventor pay to produce an infomercial for her new gizmo? Should a bank install another ATM? The answer to all of these questions is yes—*if* the action would add more to revenue than to costs. In future chapters, we'll be using the marginal approach to profit to analyze some other types of firm decisions.

Dealing with Losses

So far, we have dealt only with the pleasant case of profitable firms and how they select their profit-maximizing output level. But what about a firm that cannot earn a positive profit at *any* output level? What should it do? The answer depends on what time horizon we are looking at.

The Short Run and the Shutdown Rule

In the short run, the firm must pay for its fixed inputs, because there is not enough time to sell them or get out of lease and rental agreements. But the firm can *still* make decisions about production. And one of its options is to *shut down*—to stop producing output, at least temporarily.

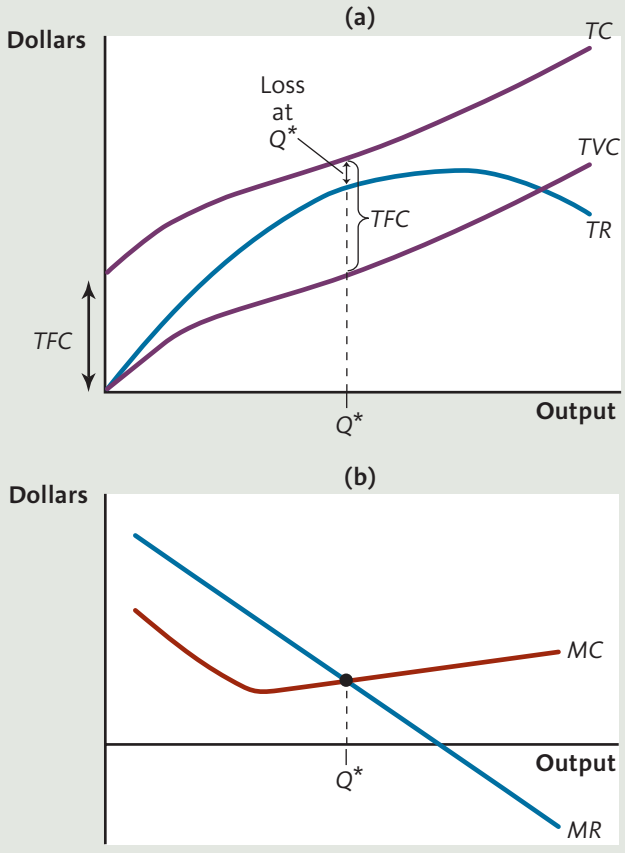
At first glance, you might think that a loss-making firm should always shut down its operation in the short run. After all, why keep producing if you are not making any profit? In fact, it makes sense for some unprofitable firms to continue operating.

Imagine a firm with the *TC* and *TR* curves shown in the upper panel of Figure 4 (ignore the *TVC* curve for now). No matter what output level the firm produces, the *TC* curve lies above the *TR* curve, so it will suffer a loss—a negative profit. For this firm, the goal is still profit maximization. But now, the highest profit will be the one with the *least negative value*. In other words, profit maximization becomes *loss minimization*.

If the firm keeps producing, then the smallest possible loss is at an output level of Q^*, where the distance between the *TC* and *TR* curves is smallest. Q^* is also the output level we would find by using our marginal approach to profit (increasing

FIGURE 4 Loss Minimization

The firm shown here cannot earn a positive profit at any level of output. If it produces anything, it will minimize its loss by producing where the vertical distance between TR *and* TC *is smallest. Because* TR *exceeds* TVC *at* Q*, *the firm will produce there in the short run.*



output whenever that adds more to revenue than to costs). This is why, in the lower panel of Figure 4, the *MC* and *MR* curves must intersect at (or very close to) Q^*.

The question is: Should this firm produce at Q^* and suffer a loss? The answer is yes—*if* the firm would lose even *more* if it stopped producing and shut down its operation. Remember that, in the short run, a firm must continue to pay its total fixed cost (*TFC*) no matter what level of output it produces—even if it produces nothing at all. If the firm shuts down, it will therefore have a loss equal to its *TFC,* since it will not earn any revenue. But if, by producing some output, the firm can cut its loss to something *less* than *TFC,* then it should stay open and keep producing.

To understand the shutdown decision more clearly, let's think about the firm's total variable costs. Business managers often call *TVC* the firm's *operating cost,* since the firm only pays these variable costs when it continues to operate. If a firm, by staying open, can earn *more* than enough revenue to cover its operating costs, then it is making an *operating profit* (*TR* > *TVC*). It should not shut down because its operating profit can be used to help pay its fixed costs. But if the firm cannot even cover its operating cost when it stays open—that is, if it would suffer an *operating loss* (*TR* < *TVC*)—it should definitely shut down. Continuing to operate only *adds* to the firm's loss, increasing the total loss beyond fixed costs.

This suggests the following guideline—called the **shutdown rule**—for a loss-making firm:

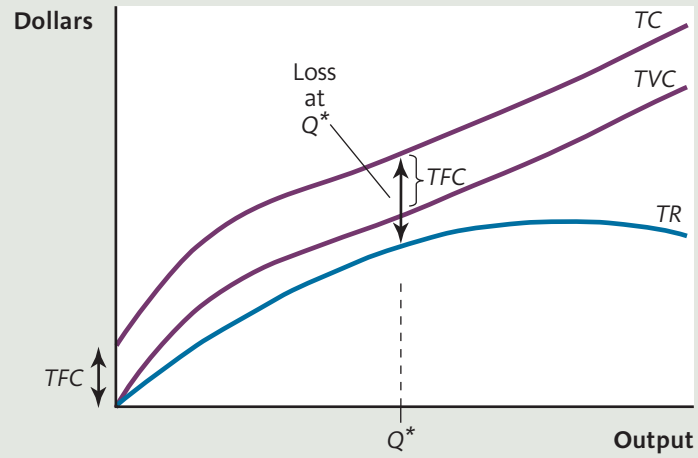
Shutdown rule In the short run, the firm should continue to produce if total revenue exceeds total variable costs; otherwise, it should shut down.

> *Let* Q* *be the output level at which* MR = MC. *Then, in the short run:*
> *If* TR > TVC *at* Q*, *the firm should keep producing.*
> *If* TR < TVC *at* Q*, *the firm should shut down.*
> *If* TR = TVC *at* Q*, *the firm should be indifferent between shutting down and producing.*

Look back at Figure 4. At Q^*, the firm is making an operating profit, since its *TR* curve is above its *TVC* curve. This firm, as we've seen, should continue to operate.

Figure 5 is drawn for a different firm, one that has different cost curves and a different *TR* curve than the firm in Figure 4. This firm *cannot* earn an operating

FIGURE 5 Shut Down



At Q*, *this firm's total variable cost exceeds its total revenue. The best policy is to shut down, produce nothing, and suffer a loss equal to* TFC *in the short run.*

profit, since its *TR* curve lies below its *TVC* curve everywhere—even at Q^*. This firm should shut down.

The shutdown rule is a powerful predictor of firms' decisions to stay open or cease production in the short run. It tells us, for example, why some seasonal businesses—such as ice cream shops in summer resort areas—shut down in the winter, when *TR* drops so low that it becomes smaller than *TVC*. And it tells us why producers of steel, automobiles, agricultural goods, and television sets will often keep producing output for some time even when they are losing money.

The Long Run and The Exit Decision

The shutdown rule applies only in the short run, a time horizon too short for the firm to escape its commitments to pay for fixed inputs such as plant and equipment. In fact, we only use the term *shut down* when referring to the short run.

Exit A permanent cessation of production when a firm leaves an industry.

But a firm can also decide to stop producing in the long run. In that case, we say the firm has decided to **exit** the industry.

The long-run decision to exit is different than the short-run decision to shut down. That's because in the long run, there *are* no fixed costs, since all inputs can be varied. Therefore, a firm that exits, by reducing all of its inputs to zero, will have *zero* costs (an option not available in the short run). And since exit also means zero revenue, a firm that exits will earn zero profit. When would a firm decide to exit and earn zero profit? When its only other alternative is to earn *negative* profit.

> *A firm should exit the industry in the long run when—at its best possible output level—it has any loss at all.*

We will look more closely at the exit decision and other long-run considerations in the next chapter.

Using the Theory

GETTING IT WRONG AND GETTING IT RIGHT: TWO CLASSIC EXAMPLES

Today, almost all managers have a good grasp of the concepts you've learned in this chapter, largely because microeconomics has become an important part of every business school curriculum. But if we go back a few decades—to when fewer managers had business degrees—we can find two examples of how management's failure to understand the basic theory of the firm led to serious errors. In one case, ignorance of the theory caused a large bank to go bankrupt; in the other, an airline was able to outperform its competitors because *they* remained ignorant of the theory. Even though these examples are old ones, they are classic.

Getting It Wrong: The Failure of Franklin National Bank

In the mid-1970s, Franklin National Bank—one of the largest banks in the United States—went bankrupt. The bank's management had made several errors, but we will focus on the most serious one.

First, a little background. A bank is very much like any other business firm: It produces output (in this case a service, making loans) using a variety of inputs (including the funds it lends out). The price of the bank's output is the interest rate it charges to borrowers. For example, with a 5 percent interest rate, the price of each dollar in loans is 5 cents per year.

Unfortunately for banks, they must also *pay* for the dollars they lend. The largest source of funds is customer deposits, for which the bank must pay interest. If a bank wants to lend out *more* than its customers have deposited, it can obtain funds from a second source, the *federal funds market,* where banks lend money to one another. To borrow money in this market, the bank will usually have to pay a higher interest rate than it pays on customer deposits.

In mid-1974, John Sadlik, Franklin's chief financial officer, asked his staff to compute the average cost to the bank of a dollar in loanable funds. At the time, Franklin's funds came from three sources, each with its own associated interest cost:

Source	Interest Cost
Checking Accounts	2.25 percent
Savings Accounts	4 percent
Borrowed Funds	9–11 percent

What do these numbers tell us? First, each dollar deposited in a Franklin *checking* account cost the bank 2.25 cents per year,[8] while each dollar in a *savings* account cost Franklin 4 cents. Also, Franklin, like other banks at the time, had to pay between 9 and 11 cents on each dollar borrowed in the federal funds market. When Franklin's accountants were asked to figure out the average cost of a dollar in loans, they divided the total cost of funds by the number of dollars they had lent out. The number they came up with was 7 cents.

This average cost of 7 cents per dollar is an interesting number, but, as we know, it should have *no relevance to a profit-maximizing firm's decisions*. And this is where Franklin went wrong. At the time, all banks, including Franklin, were charging interest rates of 9 to 9.5 percent to their best customers. But Sadlik decided that since money was costing an *average* of 7 cents per dollar, the bank could make a tidy profit by lending money at 8 percent—earning 8 cents per dollar. Accordingly, he ordered his loan officers to approve any loan that could be made to a reputable borrower at 8 percent interest. Needless to say, with other banks continuing to charge 9 percent or more, Franklin National Bank became a very popular place from which to borrow money.

But where did Franklin get the additional funds it was lending out? That was a problem for the managers in *another* department at Franklin, who were responsible for *obtaining* funds. It was not easy to attract additional checking and savings

[8] This cost for checking accounts was not technically an "interest" payment, since in the 1970s banks generally did not pay interest on checking accounts. But banks *did* provide free services such as check clearing, monthly statements, free coffee, and even gifts to their checking account depositors, and the cost of these freebies was computed to be 2.25 cents per dollar of deposits.

account deposits, since, in the 1970s, the interest rate banks could pay was regulated by the government. That left only one alternative: the federal funds market. And this is exactly where Franklin went to obtain the funds pouring out of its lending department. Of course, these funds were borrowed not at 7 percent, the average cost of funds, but at 9 to 11 percent, the cost of borrowing in the federal funds market.

To understand Franklin's error, let's look again at the average cost figure it was using. This figure included an irrelevant cost: the cost of funds obtained from customer deposits. This cost was irrelevant to the bank's lending decisions, since *additional* loans would not come from these deposits, but rather from the more expensive federal funds market. Further, this average figure was doomed to rise as Franklin expanded its loans. How do we know this? The *marginal* cost of an additional dollar of loans—9 to 11 cents per dollar—was greater than the *average* cost—7 cents. As you know, whenever the marginal is greater than the average, it pulls the average up. Thus, Franklin was basing its decisions on an average cost figure that not only included irrelevant sunk costs but was bound to increase as its lending expanded.

More directly, we can see Franklin's error through the lens of the marginal approach. The *marginal revenue* of each additional dollar lent out at 8 percent was 8 cents, while the *marginal cost* of each additional dollar—since it came from the federal funds market—was 9 to 11 cents. *MC* was greater than *MR,* so Franklin was actually losing money each time its loan officers approved another loan! Not surprisingly, these loans—which never should have been made—caused Franklin's profits to *decrease,* and within a year the bank had lost hundreds of millions of dollars. This, together with other management errors, caused the bank to fail.

Getting It Right: Continental Airlines

In the early 1960s, Continental Airlines was doing something that seemed like a horrible mistake. All other airlines at the time were following a simple rule: They would offer a flight only if, on average, 65 percent of the seats could be filled with paying passengers, since only then could the flight break even. Continental, however, was flying jets filled to just 50 percent of capacity and was actually expanding flights on many routes. When word of Continental's policy leaked out, its stockholders were angry, and managers at competing airlines smiled knowingly, waiting for Continental to fail. Yet Continental's profits—already higher than the industry average—continued to grow. What was going on?

There *was,* indeed, a serious mistake being made, but by the *other* airlines, not Continental. This mistake should by now be familiar to you: using average cost instead of marginal cost to make decisions. The "65 percent of capacity" rule used throughout the industry was derived more or less as follows: The total cost of the airline for the year (*TC*) was divided by the number of flights during the year (*Q*) to obtain the average cost of a flight ($TC/Q = ATC$). For the typical flight, this came to about $4,000. Since a jet had to be 65 percent full in order to earn ticket sales of $4,000, the industry regarded any flight that repeatedly took off with less than 65 percent as a money loser and canceled it.

As usual, there are two problems with using *ATC* in this way. First, an airline's average cost per flight includes many costs that are irrelevant to the decision to add or subtract a flight. These *sunk costs* include the cost of running the reservations system, paying interest on the firm's debt, and fixed fees for landing rights at airports—none of which would change if the firm added or subtracted a flight. Also, average cost ordinarily *changes* as output changes, so it is wrong to assume it is constant in decisions about *changing* output.

Continental's management, led by its vice-president of operations, had decided to try the marginal approach to profit. Whenever a new flight was being considered, every department within the company was asked to determine the *additional* cost they would have to bear. Of course, the only additional costs were for additional *variable* inputs, such as additional flight attendants, ground crew personnel, in-flight meals, and jet fuel. These additional costs came to only about $2,000 per flight. Thus, the *marginal* cost of an additional flight—$2,000—was significantly less than the marginal revenue of a flight filled to 65 percent of capacity—$4,000. The marginal approach to profits tells us that when $MR > MC$, output should be increased, which is just what Continental was doing. Indeed, Continental correctly drew the conclusion that the marginal revenue of a flight filled at even 50 percent of capacity—$3,000—was *still* greater than its marginal cost, and so offering the flight would increase profit. This is why Continental was expanding routes even when it could fill only 50 percent of its seats.

In the early 1960s, Continental was able to outperform its competitors by using a secret—the marginal approach to profits. Today, of course, the secret is out, and all airlines use the marginal approach when deciding which flights to offer.[9]

SUMMARY

In economics, we view the firm as a single economic decision maker with the goal of maximizing the owners' profit. Economic profit is total revenue minus *all* costs of production, explicit and implicit. In their pursuit of maximum profit, firms face two constraints. One is embodied in the demand curve the firm faces; it indicates the maximum price the firm can charge to sell any amount of output. This constraint determines the firm's revenue at each level of production. The other constraint is imposed by costs: More output always means greater costs. In choosing the profit-maximizing output, the firm must consider both revenues and costs.

One approach to choosing the optimal level of output is to measure profit as the difference between total revenue and total cost at each level of output, and then select the output level at which profit is greatest. An alternate approach uses *marginal revenue* (*MR*), the change in total revenue from producing one more unit of output, and *marginal cost* (*MC*), the change in total cost from producing one more unit. The firm should increase output whenever $MR > MC$, and lower output when $MR < MC$. The profit-maximizing output level is the one closest to the point where $MR = MC$.

If profit is negative, but total revenue exceeds total variable cost, the firm should continue producing in the short run. Otherwise, it should shut down and suffer a loss equal to its fixed cost. A firm with negative profit in the long run should exit the market.

PROBLEM SET

Answers to even-numbered Questions and Problems can be found on the text website at www.cengage.com/economics/hall.

1. You have a part-time work/study job at the library that pays $10 per hour, 3 hours per day on Saturdays and Sundays. Some friends want you to join them on a weekend ski trip leaving Friday night and returning Monday morning. They estimate your share of the gas, motel, lift tickets, and other expenses to be around $30. What is your total cost (considering both explicit and implicit costs) for the trip?

2. Until recently, you worked for a software development firm at a yearly salary of $35,000. Now, you decide to open your own business. You quit your job, cash in a $10,000 savings account (which pays 5 percent interest), and use the money to buy computer hardware to use in your business. You also convert a basement apartment in your house, which you have

[9] For more information about Continental's strategy, see "Airline Takes the Marginal Bone," *Business Week*, April 20, 1963, pp. 111–114.

been renting for $250 a month, into a workspace for your new software firm.

You lease some office equipment for $3,600 a year and hire two part-time programmers, whose combined salary is $25,000 a year. You also figure it costs around $50 a month to provide heat and light for your new office.

a. What are the total annual explicit costs of your new business?
b. What are the total annual implicit costs?
c. At the end of your first year, your accountant cheerily informs you that your total sales for the year amounted to $55,000. She congratulates you on a profitable year. Are her congratulations warranted? Why or why not?

3. The following data are price/quantity/cost combinations for Titan Industry's mainframe computer division:

Quantity	Price per Unit	Total Cost of Production
0	above $225,000	$200,000
1	$225,000	$250,000
2	$175,000	$275,000
3	$150,000	$325,000
4	$125,000	$400,000
5	$ 90,000	$500,000

a. What is the marginal revenue if output rises from 2 to 3 units? (Hint: Calculate total revenue at each output level first.) What is the marginal cost if output rises from 4 to 5 units?
b. What quantity should Titan produce to maximize total revenue? Total profit?
c. What is Titan's fixed cost? How do Titan's marginal costs behave as output increases?

4. Each entry in this table shows marginal revenue and marginal cost when a firm increases output to the given quantity:

Quantity	MR	MC
10		
	30	25
11		
	29	23
12		
	27	22
13		
	25	25
14		
	23	27
15		
	21	29
16		
	19	31
17		

What is the profit-maximizing level of output?

5. The following tables give information about demand and total cost for two firms. In the short run, how much should each produce?

Firm A

Quantity	Price	Total Cost
0	above $125	$250
1	$125	$400
2	$100	$500
3	$ 75	$550
4	$ 50	$600
5	$ 25	$700

Firm B

Quantity	Price	Total Cost
0	above $500	$ 500
1	$500	$ 700
2	$400	$ 900
3	$300	$1,100
4	$200	$1,300
5	$100	$1,500

6. At its best possible output level, a firm has total revenue of $3,500 per day and total cost of $7,000 per day. What should this firm do in the short run if:
a. the firm has total *fixed* costs of $3,000 per day?
b. the firm has total *variable* costs of $3,000 per day?

7. Suppose you own a restaurant that serves only dinner. You are trying to decide whether or not to rent out your dining room and kitchen during mornings to another firm, The Breakfast Club, Inc., that will serve only breakfast. Your restaurant currently has the following monthly costs:

Rent on building:	$ 2,000
Electricity:	$ 1,000
Wages and salaries:	$15,000
Advertising:	$ 2,000
Purchases of food and supplies:	$ 8,000
Your foregone labor income:	$ 4,000
Your foregone interest:	$ 1,000

a. Which of your current costs are implicit, and which are explicit?

b. Suppose The Breakfast Club, Inc., offers to pay $800 per month to use the building. They promise to use only their own food, and also to leave the place spotless when they leave each day. If you believe them, should you rent out your restaurant to them? Or does it depend? Explain.

8. Suppose that, due to a dramatic rise in real estate taxes, Ned's Beds' total fixed cost rises from $300 to $1,300 per day. Use the data of Table 1 to answer the following:
   a. What does the tax hike do to Ned's *MC* and *MR* curves?
   b. In the short run, how many beds should Ned produce after the rise in taxes?

9. A firm's *marginal profit* can be defined as the change in its profit when output increases by one unit.
   a. Compute the marginal profit for each change in Ned's Beds' output in Table 1.
   b. State a complete rule for finding the profit-maximizing output level in terms of marginal profit.

More Challenging

10. Suppose Ned's Beds does *not* have to lower the price in order to sell more beds. Specifically, suppose Ned can sell all the beds he wants at a price of $275 per bed.
    a. What will Ned's *MR* curve look like? (Hint: How much will his revenue rise for each additional bed he sells?)
    b. In Table 1, how would you change the numbers in the marginal revenue column to reflect the constant price for beds?
    c. Using the marginal cost and *new* marginal revenue numbers in Table 1, find the number of beds Ned should sell.

11. Howell Industries specializes in precision plastics. Their latest invention promises to revolutionize the electronics industry, and they have already made and sold 75 of the miracle devices. They have estimated average costs as given in the following table:

Unit	*ATC*
74	$10,000
75	$12,000
76	$14,000

Backus Electronics has just offered Howell $150,000 if it will produce the 76th unit. Should Howell accept the offer and manufacture the additional device?

CHAPTER 9

Perfect Competition

When we observe buyers and sellers in action, we see that different goods and services are sold in vastly different ways. Consider advertising. Every day, we are inundated with sales pitches for a long list of products: toothpaste, perfume, automobiles, cat food, banking services, and more. But not everyone advertises what they have for sale. Farmers don't advertise when they want to sell their wheat or corn. Nor do shareholders or bondholders advertise when they want to sell stocks or bonds. Why, in a world full of advertising, don't the sellers of wheat, shares of stock, corn, crude oil, copper or foreign currency advertise what they are selling?

Or consider profits. Anyone starting a business hopes to make as much profit as possible. Yet some companies—Microsoft, Quaker Oats, and PepsiCo, for example—earn sizable profits for their owners year after year, while at other companies, such as Delta Air Lines, Ford, and most small businesses, economic profit may fluctuate from year to year, but on average it is very low.

When economists turn their attention to these observed differences in trading, they think immediately about *market structure,* the subject of this and the next two chapters. We've used this term informally before, but now it's time for a formal definition:

Market structure The characteristics of a market that influence how trading takes place.

> *By **market structure**, we mean all the characteristics of a market that influence the behavior of buyers and sellers when they come together to trade.*

In microeconomics, we can divide markets for goods and services into four basic kinds of market structure: *perfect competition, monopoly, monopolistic competition,* or *oligopoly*. The subject of this chapter is perfect competition. In the next two chapters, we'll look carefully at the other market structures.

What Is Perfect Competition?

The phrase "perfect competition" should sound familiar, because you encountered it earlier, in Chapter 3. There you learned (briefly) that the famous supply and demand model explains how prices are determined in *perfectly competitive markets*. Now we're going to take a much deeper and more comprehensive look at perfectly competitive markets. By the end of this chapter, you will understand very clearly how perfect competition and the supply and demand model are related.

Let's start with the word *competition* itself. When you hear that word, you may think of an intense, personal rivalry, like that between two boxers competing in a ring or two students competing for the best grade in a small class. But there are other, less personal forms of competition. If you took the SAT exam to get into

college, you were competing with thousands of other test takers in rooms just like yours, all across the country. But the competition was *impersonal:* You were trying to do the best that you could do, trying to outperform others in general, but not competing with any one individual in the room. In economics, the term "competition" is used in the latter sense. It describes a situation of diffuse, impersonal competition in a highly populated environment.

The Four Requirements of Perfect Competition

Perfect competition *is a market structure with four important characteristics:*

1. *There are large numbers of buyers and sellers, and each buys or sells only a tiny fraction of the total quantity in the market.*
2. *Sellers offer a standardized product.*
3. *Sellers can easily enter into or exit from the market.*
4. *Buyers and sellers are well-informed.*

Perfect competition A market structure in which there are many buyers and sellers, the product is standardized, sellers can easily enter or exit the market, and buyers and sellers are well-informed.

These four conditions probably raise more questions than they answer, so let's see what each one really means.

A Large Number of Buyers and Sellers

In perfect competition, there must be many buyers and sellers. How many? It would be nice if we could specify a number—like 1,000—for this requirement. Unfortunately, we cannot, since what constitutes a large number of buyers and sellers can be different under different conditions. What is important is this:

In a perfectly competitive market, the number of buyers and sellers is so large that no individual decision maker can significantly affect the price of the product by changing the quantity it buys or sells.

Think of the world market for wheat. On the selling side, there are hundreds of thousands of individual wheat farmers—more than 250,000 in the United States alone. Each of these farmers produces only a tiny fraction of the total market quantity. If any one of them were to double, triple, or even quadruple production, the impact on total market quantity and market price would be negligible. The same is true on the buying side: There are so many small buyers that no one of them can affect the market price by increasing or decreasing its quantity demanded. Most agricultural markets conform to the large-number-of-small-firms requirement, as do markets for many commodities, such as gold or silver.

Now think about the market for athletic shoes. In 2007, four large firms—Nike, Adidas (including Reebok), New Balance, and Puma—accounted for about 70 percent of worldwide sales in this market. If any one of these producers decided to change its output by even 10 percent, the impact on total quantity supplied—and market price—would be *very* noticeable. The market for athletic shoes fails the large-number-of-small-firms requirement, so it is not an example of perfect competition.

A Standardized Product Offered by Sellers

In a perfectly competitive market, buyers do not perceive differences between the products of one seller and another. For example, buyers of wheat will ordinarily have no preference for one farmer's wheat over another's, so wheat would surely

pass the standardized product test. The same is true of many other agricultural products—for example, corn syrup and soybeans. It is also true of commodities like crude oil or pork bellies, precious metals like gold or silver, and financial instruments such as the stocks and bonds of a particular firm. (One share of eBay stock is indistinguishable from another.)

When differences among firms' products matter to buyers, the market is not perfectly competitive. For example, most consumers perceive differences among the various brands of coffee on the supermarket shelf and may have strong preferences for one particular brand. Coffee, therefore, fails the standardized product test of perfect competition. Other goods and services that would fail this test include automobiles, colleges, and magazines.

Easy Entry into and Exit from the Market

Entry into a market is rarely free; a new seller must always incur *some* costs to set up shop, begin production, and establish contacts with customers. But a perfectly competitive market has no *significant* barriers or special costs to discourage new entrants: Any firm wishing to enter can do business on the same terms as firms that are already there. For example, anyone who wants to start a wheat farm can do so, facing the same costs for land, farm equipment, seeds, fertilizer, and hired labor as existing farms. The same is true of anyone wishing to open up a dry cleaning shop, restaurant, or dog-walking service. Each of these examples would pass the easy-entry test of perfect competition.

In many markets, however, there are significant barriers to entry. For example, local zoning laws may place strict limits on how many businesses—movie theaters, supermarkets, hotels—can operate in a local area. The brand loyalty enjoyed by existing producers of breakfast cereals, instant coffee, and soft drinks would require a new entrant to wrest customers away from existing firms—a very costly undertaking. We will discuss these and other barriers to entry in more detail in later chapters.

Perfect competition also requires easy *exit:* A firm suffering a long-run loss must be able to sell off its plant and equipment and leave the industry without obstacles. Some markets satisfy this requirement, and some do not. Plant-closing laws or union agreements can require lengthy advance notice and high severance pay when workers are laid off. Or capital equipment may be so highly specialized—perhaps designed to produce just one type of automobile—that it cannot be sold off if the firm decides to exit the market. These and other barriers to exit do not conform to the assumptions of perfect competition.

Well-Informed Buyers and Sellers

In perfect competition, both buyers and sellers have all information relevant to their decision to buy or sell. For example, they know about the quality of the product, and the prices being charged by competitors. In most agricultural and commodities markets—wheat, copper, crude oil, or platinum—traders are well informed in this way. In most other types of markets, buyers and sellers are *reasonably* well informed. For example, if you buy a shirt at the Gap, you have a very good idea what you are getting, and you can easily find out the prices being charged for similar items in other stores.

But in some markets, this assumption may not be realistic. When individuals sell used cars to other individuals (via Internet sites like autotrader.com or classified listings in the local paper), the seller knows if he is unloading a lemon. But a buyer usually doesn't—unless he or she goes through considerable time and expense to have a mechanic examine the car. In the market for home insurance against fire and theft, *you*

may know whether you smoke in bed or leave your door unlocked when not at home. But—unless you have a record of past insurance claims—the insurance company has no easy way of finding this out. Markets for used cars and home insurance thus violate the easily obtained information requirements of perfect competition. We'll explore the consequences of these and other information problems in Chapter 16.

Is Perfect Competition Realistic?

The four assumptions a market must satisfy to be perfectly competitive (or just "competitive," for short) are rather restrictive. Do any markets satisfy all these requirements? How broadly can we apply the model of perfect competition when we think about the real world?

In some cases, the model fits remarkably well. We have seen that the market for wheat, for example, passes all four tests for a competitive market: many buyers and sellers, standardized output, easy entry and exit, and well-informed buyers. Indeed, most agricultural markets satisfy the strict requirements of perfect competition quite closely, as do many financial markets and some markets for consumer goods and services.

But in the vast majority of markets, one or more of the assumptions of perfect competition will, in a strict sense, be violated. This might suggest that the model can be applied only in a few limited cases. Yet when economists look at real-world markets, they use the perfect competition model more than any other market model. Why is this?

First, the model of perfect competition is powerful. Using simple techniques, it leads to important predictions about a market's response to changes in consumer tastes, technology, and government policies. While other types of market structure models also yield valuable predictions, they are often more cumbersome and their predictions less definitive.

Second, many markets, while not strictly perfectly competitive, come *reasonably* close. The more closely a real-world market fits the model, the more accurate our predictions will be when we use it.

Perfect competition approximates conditions and yields accurate-enough predictions in a wide variety of markets. This is why you will often find economists using the model to analyze the markets for crude oil, consumer electronic goods, fast-food meals, medical care, and higher education, even though in each of these cases one or more of the requirements is not strictly satisfied.

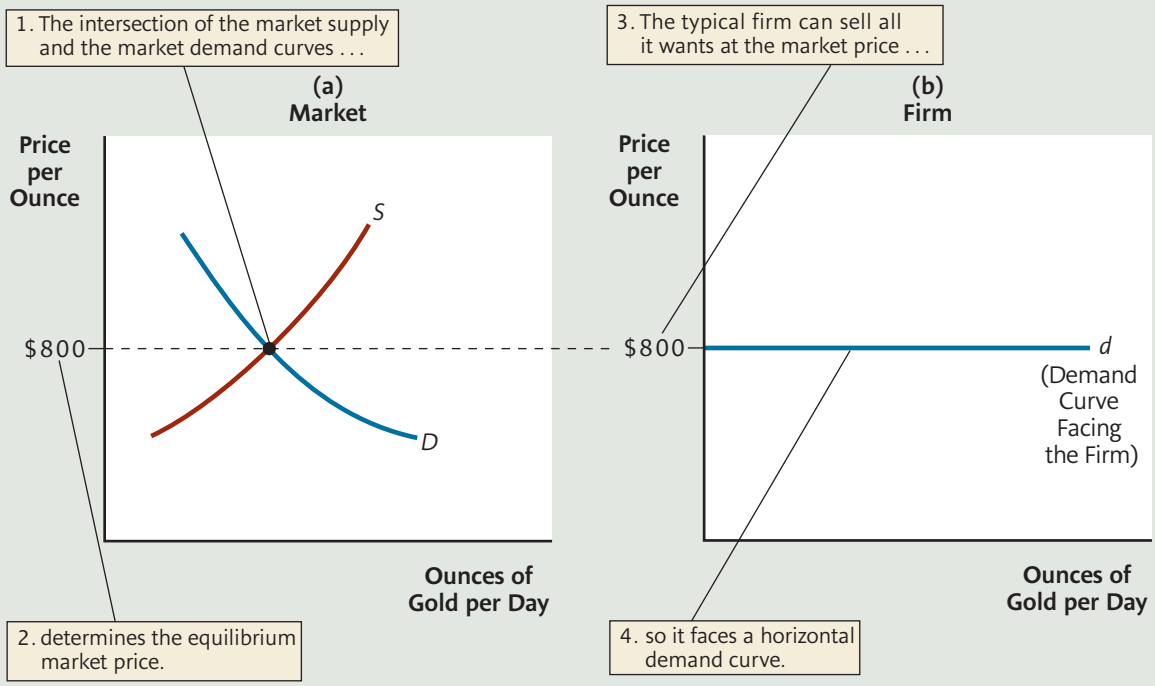
The Perfectly Competitive Firm

A market is a collection of individual decision makers. The decisions made by individuals, collectively, affect the market. And the market, in turn, influences the choices made by individuals. This is why, in learning about perfectly competitive markets, we'll be going back and forth between the competitive firm and the market in which it operates.

In Figure 1(a), we start with the market—specifically, the competitive market in which gold is produced and sold. The intersection of the market supply and demand curves determines the market price of gold which, in the figure, is $800 per troy ounce.[1] This is all familiar territory, which you learned in Chapter 3. But now, let's turn our attention to one of the many sellers in this market: Small Time Gold Mines, a small mining company.

[1] Gold is sold by the troy ounce, which is about 10 percent heavier than a regular ounce.

FIGURE 1 The Competitive Industry and Firm



The Competitive Firm's Demand Curve

Panel (b) of Figure 1 shows the demand curve facing Small Time Gold Mines. Notice the special shape of this curve: It is horizontal, or perfectly price elastic. This tells us that no matter how much gold Small Time produces, it will always sell it at the same price—$800 per troy ounce.

A perfectly competitive firm faces a demand curve that is horizontal (perfectly elastic) at the market price.

Why should this be?

First, in perfect competition, output is standardized—buyers do not distinguish the gold of one mine from that of another. If Small Time were to charge a price even a tiny bit higher than other producers, it would lose all of its customers; they would simply buy from Small Time's competitors. The horizontal demand curve captures this effect. It tells us that if Small Time raises its price above $800, it will not just sell *less* output, it will sell *no* output.

Second, Small Time is only a tiny producer relative to the entire gold market. No matter how much it produces and sells, it cannot make a noticeable difference in market quantity supplied, and so it cannot affect the market price. Once again, the horizontal demand curve describes this effect very well: The firm can increase its production without having to lower its price.

All of this means that Small Time has no control over the price of its output—it simply accepts the market price as given.

*In perfect competition, the firm is a **price taker**: It treats the price of its output as given.*

Price taker A firm that treats the price of its product as given and beyond its control.

The horizontal demand curve facing the firm and the corresponding price-taking behavior of firms are hallmarks of perfect competition. When a manager thinks, "If we produce more output, we will have to lower our price in order to sell it" then the firm faces a *downward sloping* demand curve and it is *not* a competitive firm. The manager of a competitive firm will instead think, "We can sell all the output we want at the going price, so how much should we produce?"

Notice that, since a competitive firm takes the market price as given, its only decision is *how much output to produce and sell*. And that decision will determine the firm's cost of production, as well as its total revenue. Let's see how this works in practice with Small Time Gold Mines.

dangerous curves

Two Demand Curves in Perfect Competition In the model of perfect competition, there are *two different* demand curves, as shown in Figure 1. One is the familiar *market demand curve* (labeled with an uppercase *D*). The other is the *demand curve facing the individual firm* (labeled with a lowercase *d*). If you forget the distinction, you'll get confused. For example, you might start thinking that perfectly competitive markets are for special products with perfectly elastic demand. True, the demand curve facing the individual firm is perfectly elastic, but this has nothing to do with buyers' elasticity of demand for the product itself. The elasticity of demand for the product is given by the (downward sloping) *market* demand curve.

Cost and Revenue Data for a Competitive Firm

Table 1 shows cost and revenue data for Small Time. In the first two columns are different quantities of gold that Small Time could produce each day and the selling

TABLE 1

Cost and Revenue Data for Small Time Gold Mines

(1) Output (Troy Ounces of Gold per Day)	(2) Price (per Troy Ounce)	(3) Total Revenue	(4) Marginal Revenue	(5) Total Cost	(6) Marginal Cost	(7) Profit
0	$800	$ 0		$1,100		−$1,100
			$800		$ 900	
1	$800	$ 800		$2,000		−$1,200
			$800		$ 400	
2	$800	$1,600		$2,400		−$ 800
			$800		$ 100	
3	$800	$2,400		$2,500		−$ 100
			$800		$ 200	
4	$800	$3,200		$2,700		$ 500
			$800		$ 300	
5	$800	$4,000		$3,000		$1,000
			$800		$ 500	
6	$800	$4,800		$3,500		$1,300
			$800		$ 700	
7	**$800**	**$5,600**		**$4,200**		**$1,400**
			$800		$ 900	
8	$800	$6,400		$5,100		$1,300
			$800		$1,100	
9	$800	$7,200		$6,200		$1,000
			$800		$1,300	
10	$800	$8,000		$7,500		$ 500

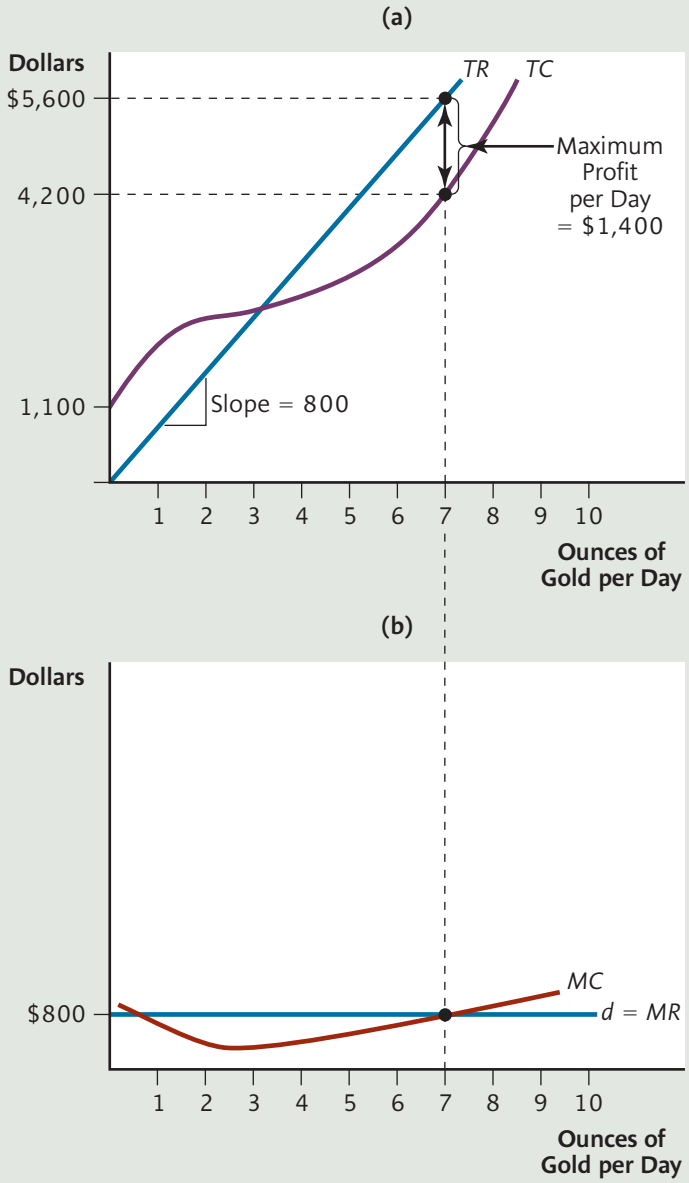
price per ounce. Because Small Time is a competitive firm (a price taker), the price remains constant at $800 per ounce, no matter *how* much gold it produces.

Run your finger down the total revenue and marginal revenue columns. Since price is always $800, each time the firm produces another ounce of gold, total revenue rises by $800. This is why marginal revenue—the additional revenue from selling one more ounce of gold—remains constant at $800.

Figure 2 plots Small Time's total revenue and marginal revenue. Notice that the total revenue (*TR*) curve in the upper panel is a *straight line* that slopes upward;

FIGURE 2 Profit Maximization in Perfect Competition

Panel (a) shows a competitive firm's total revenue (TR) *and total cost* (TC) *curves.* TR *is a straight line with slope equal to the market price. Profit is maximized at 7 ounces per day, where the vertical distance between* TR *and TC is greatest. Panel (b) shows that profit is maximized where the marginal cost* (MC) *curve intersects the marginal revenue* (MR) *curve, which is also the firm's demand curve.*



each time output increases by one unit, *TR* rises by the same $800. That is, the slope of the *TR* curve is equal to the price of output.

The marginal revenue (*MR*) curve in the lower panel is a *horizontal* line at the market price. In fact, the *MR* curve is the same horizontal line as the demand curve facing the firm. Why? Remember that marginal revenue is the additional revenue the firm earns from selling an additional unit of output. For a price-taking competitive firm, that additional revenue will always be the unchanging price it gets for each unit—in this case, $800.

For a competitive firm, marginal revenue is the same as the market price. For this reason, the marginal revenue curve and the demand curve facing the firm are the same: a horizontal line at the market price.

In panel (b), we have labeled the horizontal line "$d = MR$," since this line is both the firm's demand curve (*d*) *and* its marginal revenue curve (*MR*).

Columns 5 and 6 of Table 1 show total cost and marginal cost for Small Time. There is nothing special about cost data for a competitive firm. In Figure 2, you can see that marginal cost (*MC*)—as usual—first falls and then rises. Total cost, therefore, rises first at a decreasing rate and then at an increasing rate. (You may want to look at Chapter 7 to review why this cost behavior is so common.)

Finding the Profit-Maximizing Output Level

A competitive firm—like any other firm—wants to earn the highest possible profit, and to do so, it should use the principles you learned in Chapter 7. Although the diagrams look a bit different for competitive firms, the ideas behind them are the same. We can use either Table 1 or Figure 2 to find the profit-maximizing output level. And we can use the techniques you have already learned: the total revenue and total cost approach, or the marginal revenue and marginal cost approach.

The Total Revenue and Total Cost Approach

The *TR* and *TC* approach is the most direct way of viewing the firm's search for the profit-maximizing output level. Quite simply, at each output level, subtract total cost from total revenue to get total profit:

$$\text{Total Profit} = TR - TC$$

Then we just scan the different output levels to see which one gives the highest number for profit.

In Table 1, total profit is shown in the last column. A simple scan of that column tells us that $1,400 is the highest daily profit that Small Time Gold Mines can earn. To earn this profit, the first column tells us that Small Time must produce 7 ounces per day, its profit-maximizing output level.

The same approach to maximizing profit can be seen graphically, in the upper panel of Figure 2. There, total profit at any output level is the distance between the *TR* and *TC* curves. As you can see, this distance is greatest when the firm produces 7 units, verifying what we found in the table.

This approach is simple and straightforward, but it hides the interesting part of the story: the way that *changes* in output cause total revenue and total cost to change. The other approach to finding the profit-maximizing output level focuses on these changes.

The Marginal Revenue and Marginal Cost Approach

In the *MR* and *MC* approach, the firm should continue to increase output as long as marginal revenue is greater than marginal cost. You can verify, using Table 1, that if the firm is initially producing 1, 2, 3, 4, 5, or 6 units, it will find that $MR < MC$ when it raises output by one unit, so producing more will raise profit. Once the firm is producing 7 units, however, $MR < MC$, so further increases in output will reduce profit.

Alternatively, using the graph in panel (b) of Figure 2, we look for the output level at which $MR = MC$. As the graph shows, there are two output levels at which the *MR* and *MC* curves intersect. However, we can rule out the first crossing point because there, the *MC* curve crosses the *MR* curve from above. Remember that the profit-maximizing output is found where the *MC* curve crosses the *MR* curve from *below*, at 7 units of output.

You can see that finding the profit-maximizing output level for a competitive firm requires no new concepts or techniques; you have already learned everything you need to know in Chapter 8. In fact, the only difference is one of appearance. Ned's Beds—our firm in Chapter 8—did *not* operate under perfect competition. As a result, both its demand curve and its marginal revenue curve sloped *downward*. Small Time, however, operates under perfect competition, so its demand and *MR* curves are the same horizontal line.

Measuring Total Profit

You have already seen one way to measure a firm's total profit on a graph: the vertical distance between the *TR* and *TC* curves. In this section, you will learn another graphical way to measure profit.

To do this, we start with the firm's *profit per unit,* which is the revenue it gets on each unit minus the cost per unit. Revenue per unit is just the price (*P*) of the firm's output, and cost per unit is our familiar average total cost, so we can write:

$$\text{Profit per unit} = P - ATC.$$

In Figure 3(a), Small Time's *ATC* curve has been plotted (calculated from the data in Table 1). When the firm is producing at the profit-maximizing output level, 7 units, its *ATC* is $TC/Q = \$4{,}200/7 = \$600$. Since the price of output is \$800, profit *per unit* $= P - ATC = \$800 - \$600 = \$200$. Graphically, this is the vertical distance between the firm's demand curve and its *ATC* curve at the profit-maximizing output level.

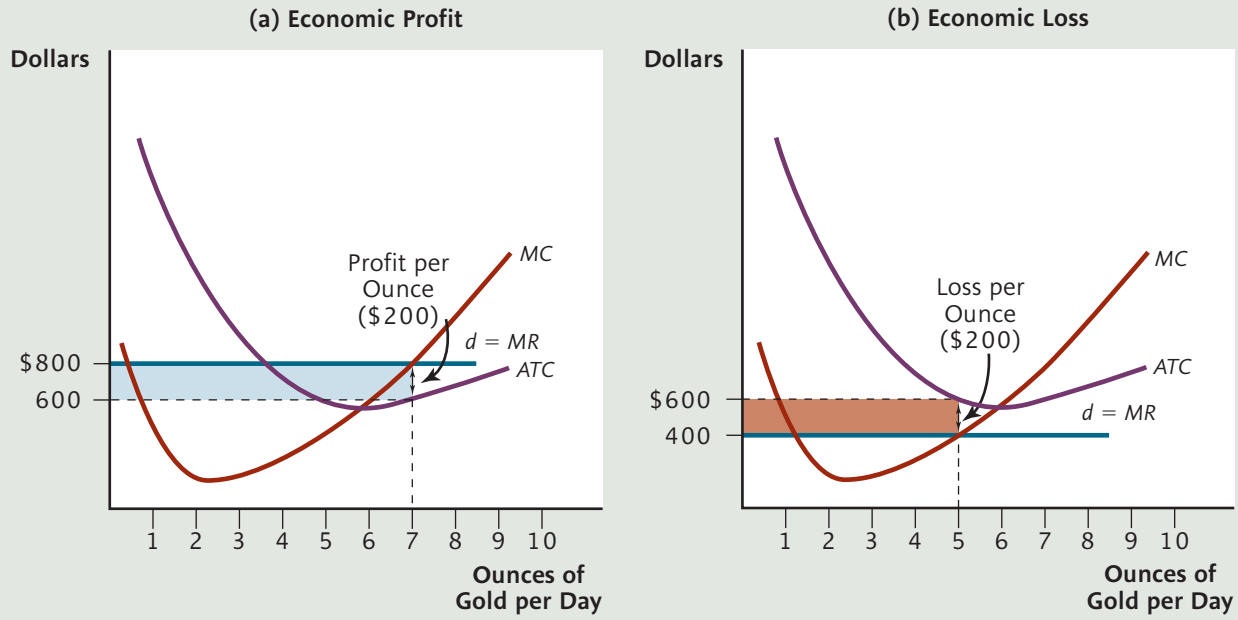
Once we know Small Time's profit per unit, it is easy to calculate its *total* profit: Just multiply profit per unit by the number of units sold. Small Time is earning \$200 profit on each ounce of gold, and it sells 7 ounces in all, so total profit is $\$200 \times 7 = \$1{,}400$.

Now look at the blue-shaded rectangle in Figure 3(a). The height of this rectangle is profit per unit, and the width is the number of units produced. The *area* of the rectangle—height × width—equals Small Time's profit:

> *A firm earns a profit whenever* P > ATC. *Its total profit at the best output level equals the area of a rectangle with height equal to the distance between* P *and* ATC, *and width equal to the quantity of output.*

In the figure, Small Time is fortunate: At a price of \$800, there are several output levels at which it can earn a profit. Its problem is to select the one that makes its profit as large as possible. (We should all wish for such problems.)

FIGURE 3 **Measuring Profit or Loss**



The competitive firm in panel (a) produces where marginal cost equals marginal revenue, or 7 units of output per day. Profit per unit at that output level is equal to revenue per unit ($800) minus cost per unit ($600), or $100 per unit. Total profit (indicated by the blue-shaded rectangle) is equal to profit per unit times the number of units sold, $200 × 7 = $1,400. In panel (b), we assume that the market price is lower, at $400 per ounce. The best the firm can do is to produce 5 ounces per day and suffer a loss shown by the red area. It loses $200 per ounce on each of those 5 ounces produced, so the total loss is $1,000—the area of the red-shaded rectangle.

But what if the price had been lower than $800—so low, in fact, that Small Time could not make a profit at *any* output level? Then the best it can do is to choose the smallest possible loss. Just as we did in the case of profit, we can measure the firm's total loss using the *ATC* curve.

Panel (b) of Figure 3 reproduces Small Time's *ATC* and *MC* curves from panel (a). This time, however, we have assumed a lower price for gold—$400—so the firm's $d = MR$ curve is the horizontal line at $400. Since this line lies everywhere below the *ATC* curve, profit per unit ($P - ATC$) is always negative: Small Time cannot make a positive profit at *any* output level.

With a price of $400, the *MC* curve crosses the *MR* curve from below at 5 units of output. Unless Small Time decides to shut down (we'll discuss shutting down for competitive firms later), it should produce 5 units. At that level of output, *ATC* is $600, and profit per unit is $P - ATC = \$400 - \$600 = -\$200$, a *loss* of $200 per unit. The total loss is loss per unit (negative profit per unit) times the number of

dangerous curves

Misusing Profit per Unit It is tempting—but *wrong*—to think that the firm should produce where profit *per unit* ($P-ATC$) is greatest. The firm's goal is to maximize *total* profit, not profit per unit. Using Table 1 or Figure 3(a), you can verify that while Small Time's profit *per unit* is greatest at 6 units of output, its *total* profit is greatest at 7 units.

units produced, or $-\$200 \times 5 = -\$1{,}000$. This is the area of the red-shaded rectangle in Figure 3(b), with height of \$200 and width of 5 units:

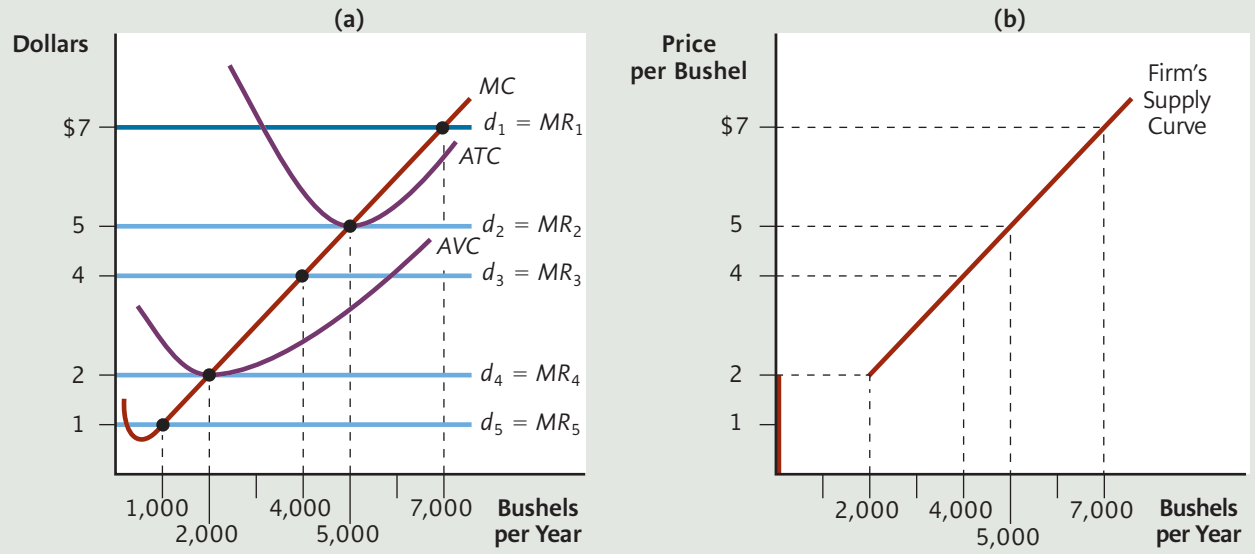
> *A firm suffers a loss whenever* P < ATC *at the best level of output. Its total loss equals the area of a rectangle with height equal to the distance between* P *and* ATC*, and width equal to the quantity of output.*

The Firm's Short-Run Supply Curve

A competitive firm is a price taker: It takes the market price as given and then decides how much output it will produce at that price. If the market price changes for any reason, the price taken as given by the firm will change as well. The firm will then have to find a new profit-maximizing output level. Let's see how the firm's profit-maximizing output changes as the market price rises or falls.

Figure 4(a) shows *ATC*, *AVC*, and *MC* curves for a competitive producer of wheat. The figure also shows five hypothetical demand curves the firm might face, each corresponding to a different market price for wheat. If the market price were \$7 per bushel, the firm would face demand curve d_1, and its profit-maximizing output level—where *MC* and *MR* intersect—would be 7,000 bushels per year. If the price dropped to \$5 per bushel, the firm would face demand curve d_2, and its profit-maximizing output level would drop to 5,000 bushels. You can see that the

FIGURE 4 **Short-Run Supply under Perfect Competition**



Panel (a) shows a typical competitive firm facing various market prices. For prices between \$2 and \$7 per bushel, the profit-maximizing quantity is found by sliding along the MC *curve. Below \$2 per bushel, the firm is better off shutting down, because* P < AVC*, Panel (b) shows that the firm's supply curve consists of two segments. Above the shutdown price of \$2 per bushel it follows the* MC *curve; below that price, it is coincident with the vertical axis.*

profit-maximizing output level is always found by traveling from the price, across to the firm's *MC* curve, and then down to the horizontal axis. In other words,

> *as the price of output changes, the firm will slide along its* MC *curve in deciding how much to produce.*

But there is one problem with this: If the firm is suffering a loss—a loss large enough to justify shutting down—then it will *not* produce along its *MC* curve; it will produce zero units instead. Thus, in order to know for certain how much output the firm will produce, we must bring in the shutdown rule.

The Shutdown Price

In Chapter 8, you learned that a firm should shut down in the short run if, at its best positive output level, it finds that $TR < TVC$. (In words, if the firm cannot even cover its operating costs, it should not continue to operate.) If $TR > TVC$, the firm should continue to operate.

But when we use a graph such as Figure 4(a), which has different prices *per unit* on the vertical axis and has curves showing cost *per unit,* it will be helpful to express this shutdown rule in "per unit" terms. Let's start with the rule from Chapter 8:

$$\text{Shut down if } TR < TVC$$

Next, with lowercase q representing the individual firm's output level, we divide both sides of the inequality by q:

$$\text{Shut down if } (TR/q) < (TVC/q)$$

Finally, we recognize that TR/q is just revenue per unit, or the price (P), and TVC/q is the firm's average variable cost (AVC), giving us

$$\text{Shut down if } P < AVC$$

Now let's apply the shutdown rule to the firms in Figure 4(a). Suppose the price drops down to \$4 per bushel. At this price, the best output level is 4,000 bushels, and the firm suffers a loss, since $P < ATC$. Should the firm shut down? Let's see. At 4,000 bushels, it is also true that $P > AVC$, since the demand curve lies above the *AVC* curve at this output level. Thus, at a price of \$4, the firm will stay open and produce 4,000 units of output.

Now, suppose the price drops all the way down to \$1 per bushel. At this price, $MR = MC$ at 1,000 bushels. But notice that here $P < AVC$. Therefore, at a price of \$1, this firm will shut down and produce *zero* units of output.

Finally, let's consider a price of \$2. At this price, $MR = MC$ at 2,000 bushels, and here we have $P = AVC$. At \$2, therefore, the firm will be indifferent between staying open and shutting down. We call this price the firm's **shut-down price,** since it will shut down at any price lower and stay open at any price higher.

Shutdown price The price at which a firm is indifferent between producing and shutting down.

Note that the shutdown price is found at the *minimum* of the *AVC* curve. Why? As the price decreases, the best output level is found by sliding along the *MC* curve, until *MC* and *AVC* cross. At that point, the firm will shut down. But—as you learned in Chapter 7—*MC* will always cross *AVC* at its minimum point.

Now let's recapitulate what we've found about the firm's output decision. For all prices above the minimum point on the *AVC* curve, the firm will stay open and will produce the level of output at which $MR = MC$. For these prices, the firm slides

along its *MC* curve in deciding how much output to produce. But for any price below the minimum *AVC*, the firm will shut down and produce zero units. We can summarize all of this information in a single curve—the **firm's supply curve**—which tells us how much output the firm will produce at any price:

Firm's supply curve A curve that shows the quantity of output a competitive firm will produce at different prices.

> *A competitive firm's supply curve is its* MC *curve for all prices above* AVC, *and a vertical line at zero units for all prices below* AVC.

In panel (b) of Figure 4, we have drawn the supply curve for our hypothetical wheat farmer. As price declines from $7 to $2, output is determined by the firm's *MC* curve. For all prices *below* $2—the shutdown price—output is zero and the supply curve coincides with the vertical axis.

Competitive Markets in the Short Run

Recall that the short run is a time period too short for the firm to vary its fixed inputs. Therefore, logically, the short run is also insufficient time for a *new* firm to acquire those fixed inputs and *enter* the market. Similarly, it is too short a period for firms to reduce their fixed inputs to zero and *exit* the market. We conclude that

> *in the short run, the number of firms in the industry is fixed.*

The Market Supply Curve

Once we know how many firms there are in a market, and we know each firm's supply curve, we can easily determine the *market supply curve*.

Market supply curve A curve indicating the quantity of output that all sellers in a market will produce at different prices in the short run.

> *To obtain the* ***market supply curve****, we add up the quantities of output supplied by all firms in the market at each price.*

To keep things simple, suppose there are 100 identical wheat farms and that each one has the supply curve shown in Figure 5(a). (This is the same supply curve we derived earlier, in Figure 4.) If the price is $7, each firm produces 7,000 bushels. With 100 such firms, the market quantity supplied is 7,000 × 100 = 700,000 bushels. If the price is $5, each firm supplies 5,000 bushels, so market supply is 500,000. Continuing in this way, we can trace out the market supply curve shown in panel (b) of Figure 5. Notice that once the price drops below $2—the shutdown price for each firm—the market supply curve jumps to zero.

The market supply curve in the figure is a *short-run* market supply curve, since it gives us the combined output level of just those firms *already* in the industry. As we move along this curve, we are assuming that two things are constant: (1) the fixed inputs of each firm and (2) the number of firms in the market.

Short-Run Equilibrium

How does a perfectly competitive market achieve equilibrium? We've already addressed this question in Chapter 3, in our study of supply and demand. But now we'll take a much closer look, paying attention to the individual firm and individual consumer as well as the market.

FIGURE 5 Deriving the Market Supply Curve

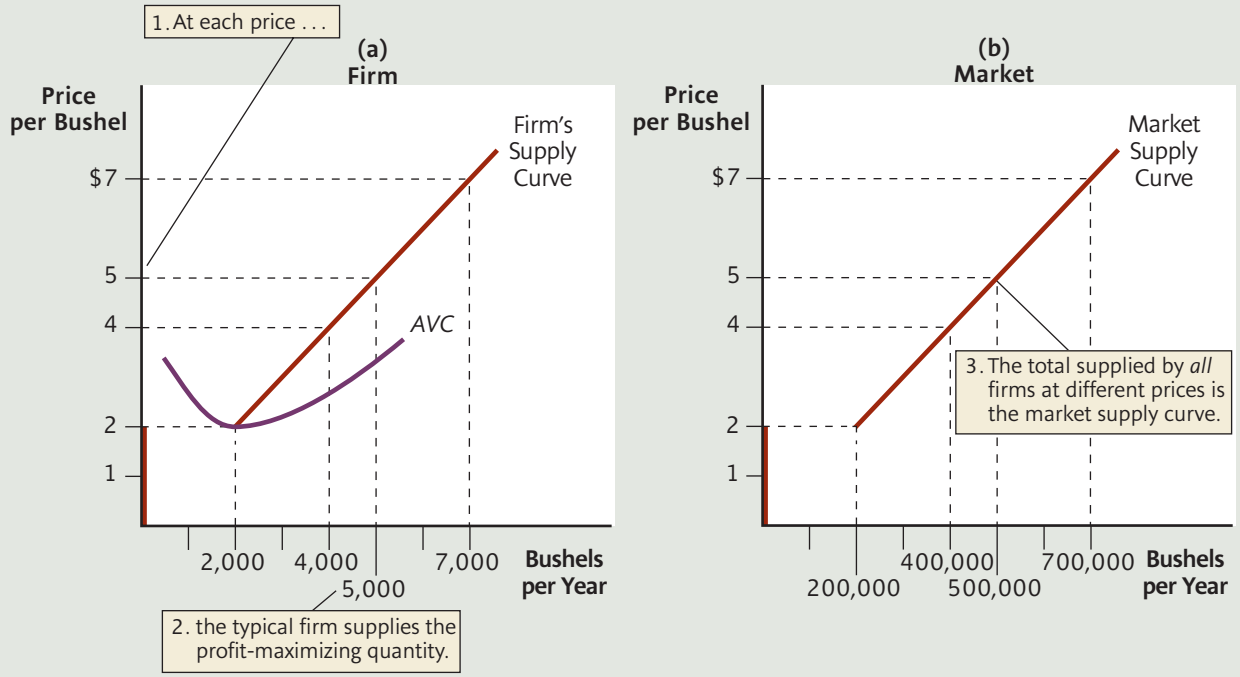


Figure 6 puts together the pieces we've discussed so far, including those from Chapter 6 on consumer choice. It paints a complete picture of how a competitive market arrives at its short-run equilibrium. On the right side, we add up the quantities supplied by all firms to obtain the market supply curve. On the left side, we add up the quantities demanded by all consumers to obtain the market demand curve.

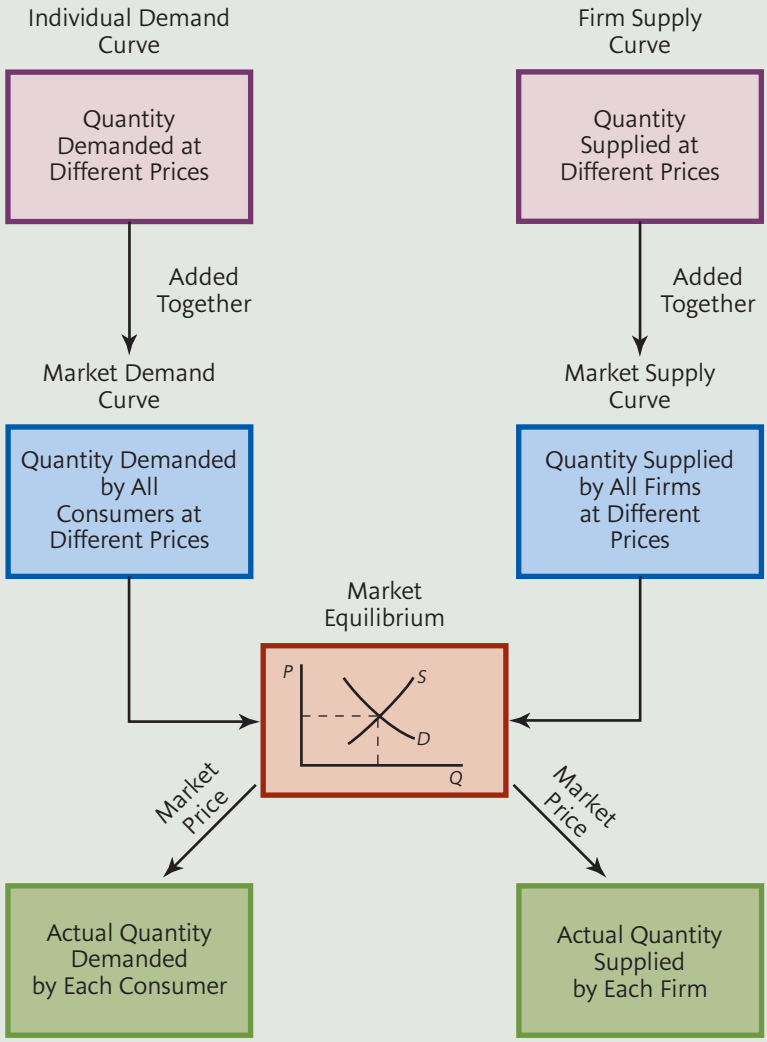
At this stage, the market supply and demand curves show if/then relationships: *If* the price were such and such, *then* firms would supply this much and consumers would buy that much. But once we bring the two curves together and find their intersection point, we know the *equilibrium* price at which trading will actually take place. Finally, when we confront each firm and each consumer with the equilibrium price, we find the actual quantity each consumer will buy and the actual quantity each firm will produce.

Figure 7 gets more specific, illustrating two possible short-run equilibriums in the wheat market, depending on the position of the market demand curve. In panel (a), if the market demand curve were D_1, the short-run equilibrium price would be \$7. Each firm would face the horizontal demand curve d_1 [panel (b)] and decide to produce 7,000 bushels. With 100 such firms, the equilibrium market quantity would be 700,000 bushels. Notice that, at a price of \$7, each firm is enjoying an economic profit, since $P > ATC$.

If the market demand curve were D_2 instead, the equilibrium price would be \$4. Each firm would face demand curve d_2 and produce 4,000 bushels. With 100 firms, the equilibrium market quantity would be 400,000. Here, each firm is suffering an economic loss, since $P < ATC$. These two examples show us that

in short-run equilibrium, competitive firms can earn an economic profit or suffer an economic loss.

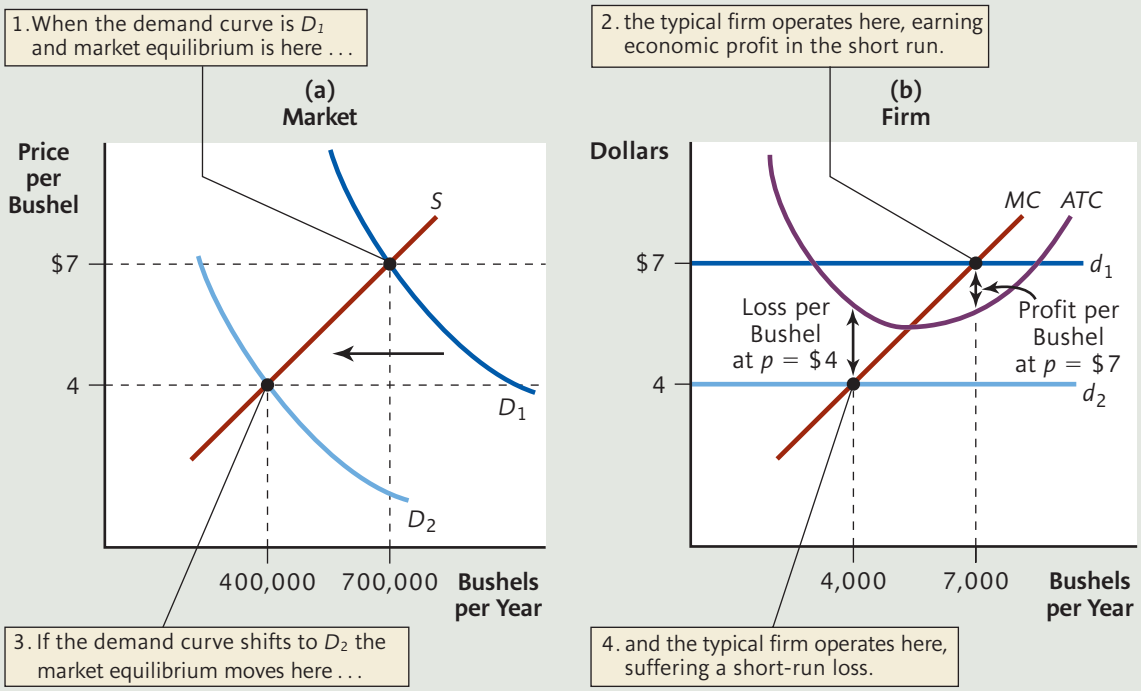
FIGURE 6 **Perfect Competition**



Equilibrium in Perspective

We are about to leave the short run and turn our attention to what happens in a competitive market over the long run. But before we do, let's look once more at how a short-run equilibrium is established. One part of this process—combining supply and demand curves to find the market equilibrium—has been familiar to you all along. But now you can see how much information is contained within each of these curves. And you can appreciate what an impressive job the market does—coordinating millions of decisions made by people who may never even meet each other.

FIGURE 7 Short-Run Equilibrium in Perfect Competition



Think about it: So many individual consumers and firms, each with its own agenda, trading in the market. Not one of them has any power to decide or even influence the market price. Rather, the price is determined by *all* of them, adjusting until *total* quantity supplied is equal to *total* quantity demanded. Then, facing this equilibrium price, each consumer buys the quantity he or she wants, each firm produces the output level that it wants, and we can be confident that all of them will be able to realize their plans. Each buyer can find willing sellers, and each seller can find willing buyers.

In perfect competition, the market sums up the buying and selling preferences of individual consumers and producers, and determines the market price. Each buyer and seller then takes the market price as given, and each is able to buy or sell the desired quantity.

This process is, from a certain perspective, a thing of beauty, and it happens each day in markets all across the world—markets for wheat, corn, barley, soybeans, apples, oranges, gold, silver, copper, and more. And something quite similar happens in other markets that do not strictly satisfy our requirements for perfect competition—markets for television sets, books, air conditioners, fast-food meals, bottled water, blue jeans. . . . The list is virtually endless.

Competitive Markets in the Long Run

The long run is a time horizon sufficiently long for firms to vary *all* of their inputs. This includes inputs that were treated as fixed in the short run, such as plant and equipment. Logically, then, the long run must be enough time for *new* firms to acquire those inputs and enter the market as *new* suppliers. And it is also long enough for existing firms to sell all such inputs and exit the market.

> *In the long run, new firms can enter a competitive market, and existing firms can exit the market.*

But what makes firms want to enter or exit a market? The driving force behind entry is economic profit, and the force behind exit is economic loss.

Profit and Loss and the Long Run

Recall that economic profit is the amount by which total revenue exceeds *all* costs of doing business. The costs we deduct include implicit costs like foregone investment income or foregone wages for an owner who devotes money or time to the business. Thus, when a firm earns positive economic profit, we know the owners are earning *more* than they could by devoting their money and time to some other activity.

© RAMIN TALAIE/CORBIS

Long run exit from a market can occur in different ways. An example is when a firm (such as this retailer) goes entirely out of business.

A temporary episode of positive economic profit will not have much impact on a competitive industry, other than the temporary pleasure it gives the owners of competitive firms. But when positive profit reflects basic conditions in the industry and is expected to continue, major changes are in the works. Outsiders, hungry for profit themselves, will want to enter the market and—since *there are no barriers to entry*—they can do so.

On the other hand, if firms already in the industry are suffering economic losses, they are not earning enough revenue to cover all their costs. There must be other opportunities that would more adequately compensate the owners for their money or time. If this situation is expected to continue over the firm's long-run planning horizon—a period long enough to vary *all* inputs—there is only one thing for the firm to do: exit the industry by selling off its plant and equipment, thereby reducing its loss to zero.

> *In a competitive market, economic profit and loss are the forces driving long-run change. The expectation of continued economic profit causes outsiders to enter the market; the expectation of continued economic losses causes firms in the market to exit.*

In the real world of business, entry and exit occur in a variety of different ways. Sometimes it involves the formation of an entirely new firm, such as JetBlue, an airline formed in 2000. Entry can also occur when an existing firm adds a new product line, as Apple did when it entered the wireless phone market in 2007. Or, in a local market, entry can occur when an existing firm creates a new branch, such as when Wal-Mart or Starbucks builds a new store. In all of these cases, the number of sellers in the market increases.

Exit, too, can occur in different ways. Sometimes, it involves a firm going entirely out of business as did Skybus Airlines in 2008, and Circuit City in 2009. But exit can also occur when a firm switches out of a particular product line, even as it continues to produce other things. In 2009, for example, Condé Nast stopped

publishing *Portfolio,* thus exiting the market for business-oriented magazines. But it continued to publish other monthlies—such as *Vogue, Allure,* and *Wired*—for other magazine markets.

Long-Run Equilibrium

Entry and exit—however they occur—are powerful forces in real-world competitive markets. They determine how these markets change over the long run, how much output will be available to consumers, and the prices they must pay. To explore these issues, let's see how entry and exit move a market to its long-run equilibrium from different starting points.

From Short-Run Profit to Long-Run Equilibrium

Suppose that the market for wheat is initially in a short-run equilibrium like point *A* in panel (a) of Figure 8, with market supply curve S_1. The initial equilibrium price is $9 per bushel. In panel (b), we see that a typical competitive firm—producing 9,000 bushels—is earning economic profit, since $P > ATC$ at that output level. As long as we remain in the short run, with no new firms entering the market, this situation will not change.

But as we enter the long run, much will change. First, economic profit will attract new entrants, increasing the number of firms in the market. Now remember (from Chapter 3) when we draw a market supply curve like S_1, we draw it for some *given* number of firms, and we hold that number constant. But in the long run, as the number of firms increases, the market supply curve will *shift rightward;* a greater quantity will be supplied at any given price. As the market supply curve shifts rightward, several things happen:

1. The market price begins to fall—from $9 to $8 to $7 and so on.
2. As market price falls, the horizontal demand curve facing each firm shifts downward.
3. Each firm—striving as always to maximize profit—will slide down its marginal cost curve, decreasing output.[2]

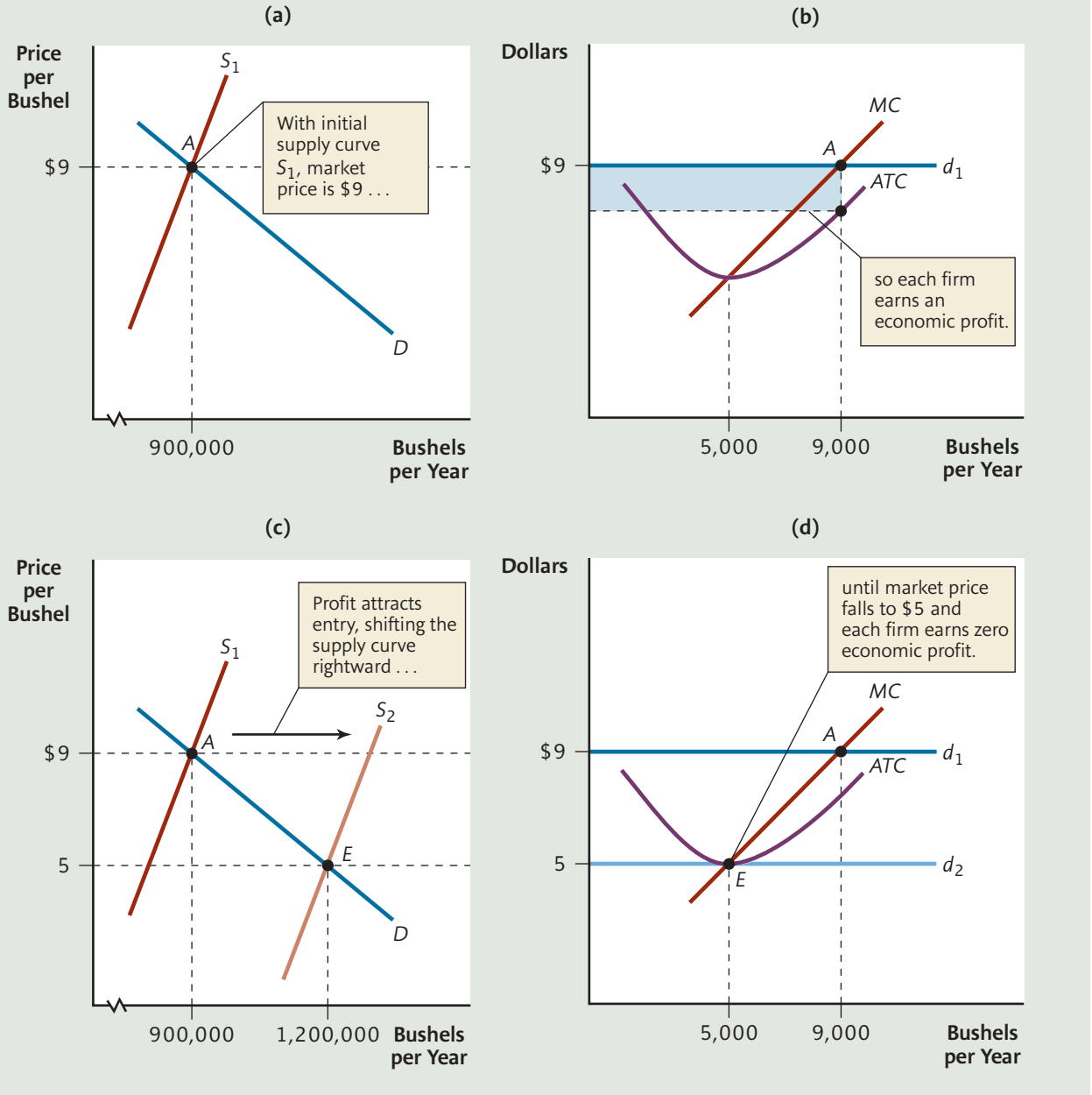
This process of adjustment, in the market and the firm, continues until the *reason* for entry—positive profit—no longer exists. That is, it will continue until the market supply curve shifts rightward enough, and the price falls enough, so that *each existing firm is earning zero economic profit*.

Panels (c) and (d) in Figure 8 show the final, long-run equilibrium. First, look at panel (c), which shows long-run market equilibrium at point *E*. The market supply curve has shifted to S_2, and the price has fallen to $5 per bushel. Next, look at panel (d), which tells us why the market supply curve stops shifting when it reaches S_2. With that supply curve, each firm is producing at the lowest point of its *ATC* curve, with $P = ATC = \$5$, and each is earning zero economic profit. With no economic profit, there is no further reason for entry, and no further shift in the market supply curve.

In a competitive market, positive economic profit continues to attract new entrants until economic profit is reduced to zero.

[2] There is one other possible consequence that we ignore here: Entry into the industry, which changes the demand for the industry's inputs, may also change input prices. If this occurs, firms' cost curves can shift as well. We'll explore this possibility a few pages later.

FIGURE 8 From Short-Run Profit to Long-Run Equilibrium



Now you can see the role played by one of our assumptions about competitive markets: *easy entry*. With no significant barriers to entry, we can be confident that economic profit at the typical firm will attract new firms to the industry, driving down the market price until the economic profit disappears. If a permanent barrier—legal or otherwise—prevented new firms from coming into the market, this mechanism would not work, so long-run economic profit would be possible.

Before proceeding further, take a close look at Figure 8. As the market moves to its long-run equilibrium [point *E* in panels (c) and (d)], output at each firm *decreases*

from 9,000 to 5,000 bushels. But in the market as a whole, output *increases* from 900,000 to 1,200,000 bushels. How can this be? (See if you can answer this question yourself. Hint: entry!)

From Short-Run Loss to Long-Run Equilibrium

We have just seen how, beginning from a position of short-run profit at the typical firm, a competitive market will adjust until the profit is eliminated. But what if we begin from a position of loss? As you might guess, the same type of adjustments will occur, only in the opposite direction.

This is a good opportunity for you to test your own skill and understanding. Study Figure 8 carefully. Then see if you can draw a similar diagram that illustrates the adjustment from short-run *loss* to long-run equilibrium. Use the same demand curve as in Figure 8, but draw in a new, appropriate market supply curve to create an initial equilibrium price of $3. Then let the market work. Show what happens in the market, and at each firm, as economic loss causes some firms to exit. If you do this correctly, you'll end up once again at a market price of $5.00, with each firm earning zero economic profit. Your graph will illustrate the following conclusion:

> *In a competitive market, economic losses continue to cause exit until the losses are reduced to zero.*

Notice the role played by our assumption of *easy exit* in competitive markets. When there are no significant barriers to exit, we can be confident that economic loss will eventually drive firms from the industry, raising the market price until the typical firm breaks even again. Significant barriers to exit (such as a local law forbidding a plant from closing down) would prevent this mechanism from working, and economic losses could persist even in the long run.

The Notion of Zero Profit in Perfect Competition

From the preceding discussion, you may wonder why anyone in his or her right mind would ever want to set up shop in a competitive industry or stay there for any length of time. In the long run, after all, they can expect zero economic profit. Indeed, if you want to become a millionaire, you would be well advised *not* to buy a wheat farm. But most wheat farmers—like most other sellers in competitive markets—do not curse their fate. On the contrary, they are likely to be reasonably content with the performance of their businesses. How can this be?

Remember that zero *economic* profit is not the same as zero *accounting* profit. When a firm is making zero *economic* profit, it is still making some accounting profit. In fact, the accounting profit is just enough to cover all of the owner's implicit costs, including compensation for any foregone investment income or foregone salary. Suppose, for example, that a farmer paid $100,000 for land and works 40 hours per week. Suppose, too, that the $100,000 *could* be invested in some other way and earn $6,000 per year, and the farmer *could* work equally pleasantly elsewhere and earn $50,000 per year. Then the farm's implicit costs will be $56,000, and zero economic profit means that the farm is earning $56,000 in *accounting profit* each year. This won't make a farmer ecstatic, but it will make it worthwhile to keep working the farm. After all, if the farmer quits and takes up the next best alternative, he or she will do no better.

Normal profit Another name for zero economic profit.

To emphasize that zero economic profit is not an unpleasant outcome, economists often replace it with the term **normal profit,** which is a synonym for "zero economic profit," or "just enough accounting profit to cover implicit costs." Using this language, we can summarize long-run conditions at the typical firm this way:

> *In the long run, the competitive firm will earn normal profit—that is, zero economic profit.*

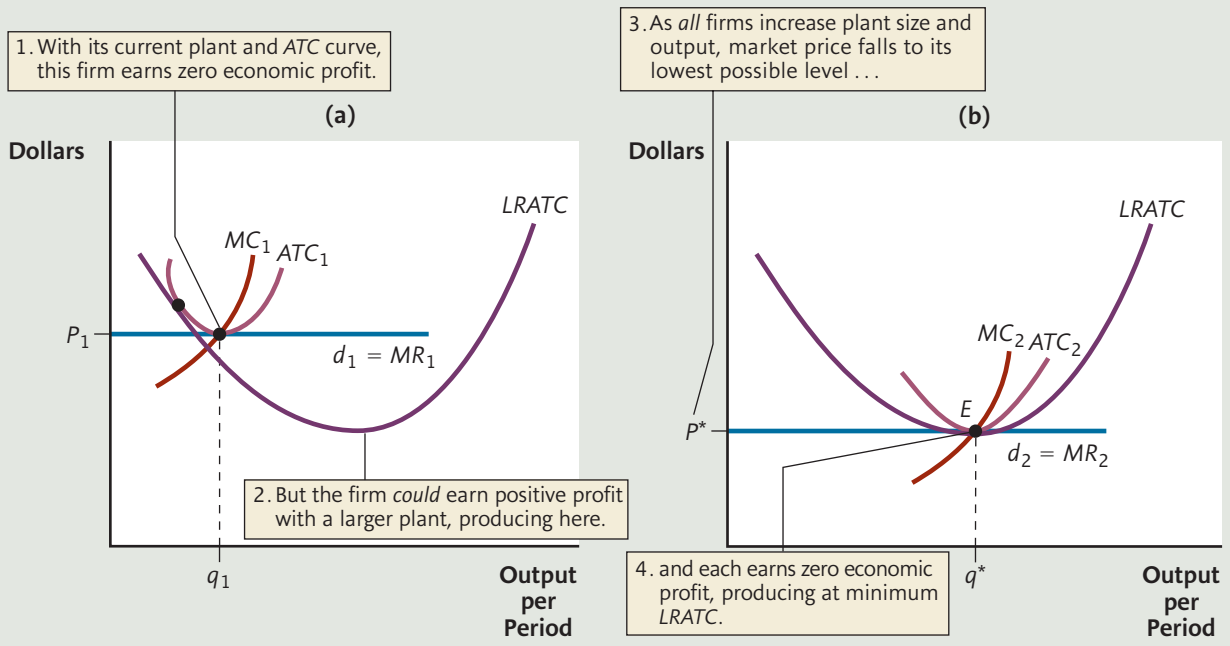
Perfect Competition and Plant Size

There is one more characteristic of competitive markets in the long run that we have not yet discussed: plant size. It turns out that the same forces—entry and exit—that cause all firms to earn zero economic profit *also* determine the size of each firm's plant.

> *In long-run equilibrium, a competitive firm will operate with the plant and output level that bring it to the bottom of its* LRATC *curve.*

To see why, let's consider what would happen if this condition were violated. Figure 9(a) illustrates a firm in a perfectly competitive market. The firm faces a market price of P_1 and produces quantity q_1, where $MC_1 = MR_1$. With its current plant, the firm has average costs given by ATC_1. Note that the firm is earning zero profit, since average cost is equal to P_1 at the best output level.

FIGURE 9 Perfect Competition and Plant Size



But panel (a) does *not* show a true long-run equilibrium. How do we know this? First, in the long run, the typical firm will want to expand. Why? Because by increasing its plant size, it could slide down its *LRATC* curve and produce more output at a lower cost per unit. Since it is a perfectly competitive firm—a small participant in the market—it can expand in this way *without* worrying about affecting the market price. As a result, the firm, after expanding, could operate on a new, lower *ATC* curve, so that *ATC* is less than *P*. That is, by expanding, the firm could potentially earn an economic profit.

Second, this same opportunity to earn positive economic profit will attract new entrants that will establish larger plants from the outset.

Expansion by existing firms and entry by new ones increase market output and bring down the market price. (This would be illustrated by a rightward shift of the market supply curve, which is not shown in the figure.) The process will stop—and a long-run equilibrium will be established—only when there is no potential to earn positive economic profit with *any* plant size. As you can see in panel (b), this condition is satisfied only when each firm is operating at the minimum point on its *LRATC* curve, using the plant represented by ATC_2, and producing output of q^*. Entry and expansion must continue in this market until the price falls to P^* because only then will each firm—doing the best that it can do—earn zero economic profit. (*Question:* In the long run, what would happen to a firm if it refused to increase its plant size?)

A Summary of the Competitive Firm in the Long Run

Panel (b) of Figure 9 summarizes everything you have learned about the competitive firm in long-run equilibrium. The typical firm, taking the market price P^* as given, produces the profit-maximizing output level q^*, where $MR = MC$. Since this is the long run, each firm will be earning zero economic profit, so we also know that $P^* = ATC$. But since $P^* = MC$ and $P^* = ATC$, it must also be true that $MC = ATC$. As you learned in Chapter 7, *MC* and *ATC* are equal only at the minimum point of the *ATC* curve. Thus, we know that each firm must be operating at the lowest possible point on the *ATC* curve for the plant it is operating. Finally, each firm selects the plant that makes its *LRATC* as low as possible, so each operates at the minimum point on its *LRATC* curve.

As you can see, there is a lot going on in Figure 9(b). But we can put it all together with a very simple statement:

> *In long-run equilibrium, the competitive firm operates where* MC = *minimum* ATC = *minimum* LRATC = P.

In Figure 9(b), this equality is satisfied when the firm produces at point *E,* where its demand, marginal cost, *ATC,* and *LRATC* curves all intersect. This is a figure well worth remembering, since it summarizes so much information about competitive markets in a single picture. (Here is a useful self-test: Close the book, put away your notes, and draw a set of diagrams in which one curve at a time does *not* pass through the common intersection point of the other three. Then explain which principle of firm or market behavior is violated by your diagram. Do this separately for all four curves.)

Figure 9(b) also explains one of the important ways in which perfect competition benefits consumers: In the long run, each firm is driven to the plant size and output level at which its cost per unit is as low as possible. This lowest possible cost per unit

is also the price per unit that consumers will pay. If price were any lower than P^*, it would not be worthwhile for firms to continue producing the good in the long run. Thus, given the *LRATC* curve faced by each firm in this industry—a curve that is determined by the technology of production and the prices of its inputs—P^* is the lowest possible price that will ensure the continued availability of the good. In perfect competition, consumers are getting the best deal they could possibly get.

What Happens When Things Change?

So far, you've learned how competitive firms make decisions, how these decisions lead to a short-run equilibrium in the market, and how the market moves from short- to long-run equilibrium through entry and exit. Now, it's time to ask: *What happens when things change?*

A Change in Demand

In Figure 10, panel (a) shows a competitive market that is initially in long-run equilibrium at point *A*, where the market demand curve D_1 and supply curve S_1 intersect. Panel (b) shows conditions at the firm, which faces demand curve d_1 and produces the profit-maximizing quantity q_1.

But now suppose that the market demand curve shifts rightward to D_2 and remains there. (This shift could be caused by any one of several factors. If you can't list some of them, turn back to Chapter 3.) Panels (c) and (d) show what happens. In the *short run*, the shift in demand moves the market equilibrium to point *B*, with market output Q_{SR} and price P_{SR}. At the same time, the demand curve facing each firm shifts upward, and each firm raises output to the new profit-maximizing level q_{SR}. At this output level, $P > ATC$, so each firm is earning economic profit. Thus, the short-run impact of an increase in demand is (1) a rise in market price, (2) a rise in market quantity, and (3) economic profits.

Constant cost industry An industry in which the long-run supply curve is horizontal because each firm's cost curves are unaffected by changes in industry output.

When we turn to the long run, we know that economic profit will attract the entry of new firms. And, as you learned a few pages ago, an increase in the number of firms shifts the market supply curve rightward, which drives down the price until the economic profit is eliminated. But how far must the price fall in order to bring this about? That is, how far can we expect the market supply curve to shift? That depends on whether or not the expansion of the industry causes each firm's cost curves to shift.

A Constant Cost Industry

Let's assume, for now, that a change in industry output (such as when new firms enter) has no impact on the cost curves of the individual firm. This is called a **constant cost industry**. Then in panel (c), entry will continue—and the supply curve will continue shifting rightward—until the price returns to P_1, its original level. (At any higher price, each firm would still be

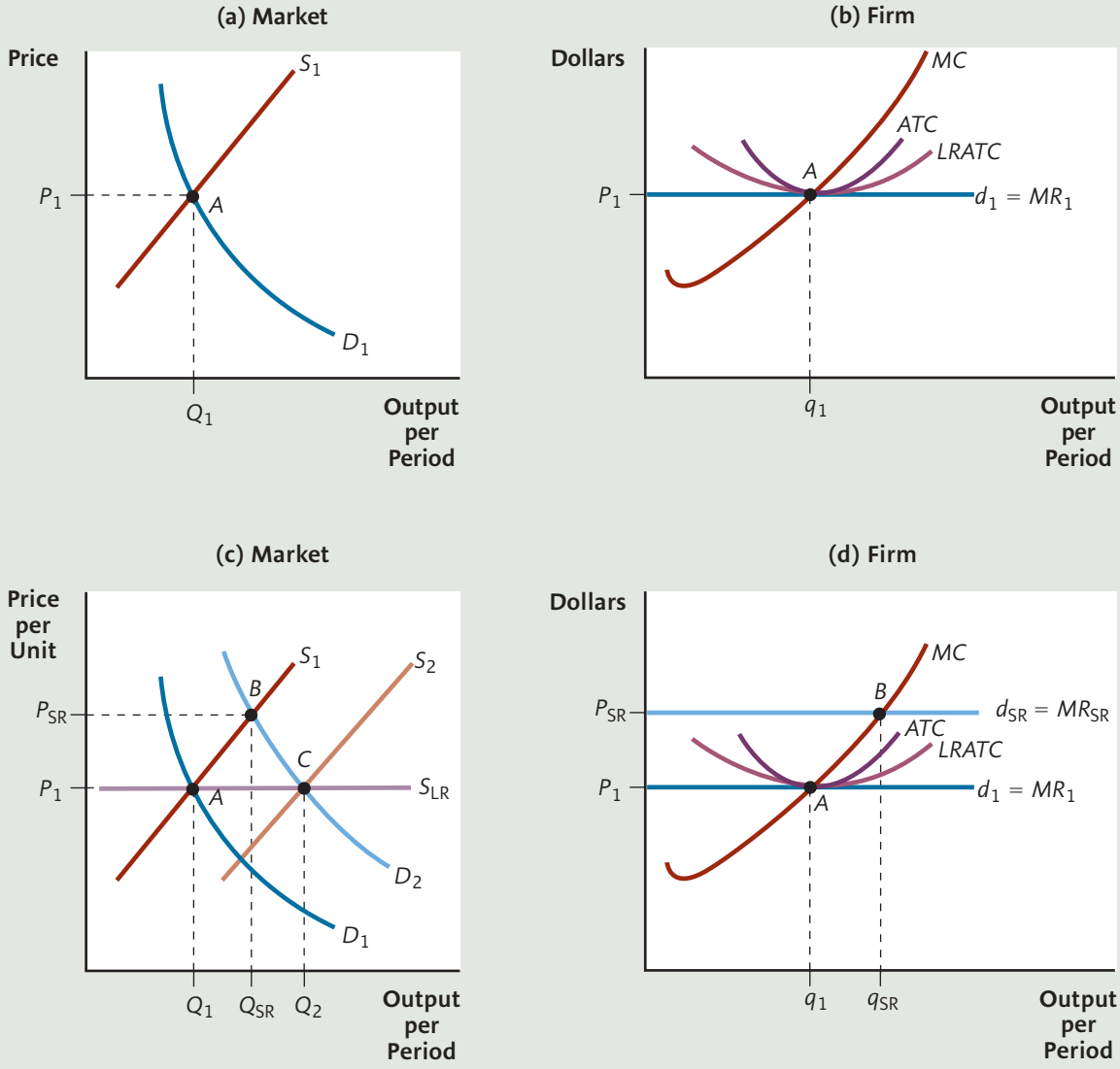
dangerous curves

Do Demand Shifts Cause Supply Shifts? In Chapter 3, you learned that a rightward shift in demand does *not* cause a rightward shift in supply. Instead, it raises the price and causes *a movement along* the supply curve. But in Figure 10, the demand curve shifts rightward from D_1 to D_2, the price rises, and then . . . the supply curve shifts rightward! So, now you may be wondering whether demand shifts *do* cause supply shifts.

The answer is: They *don't*—at least, not directly, and not at all in the short run. But demand shifts indirectly cause supply shifts in the long run. In the figure, you can see that the shift in demand first raises the price, moving us *along* the supply curve S_1 from point *A* to point *B* in the short run. This is just as you learned in Chapter 3, which only dealt with the short run. (In that chapter, we assumed the number of firms was constant.)

But now, we're extending our analysis further, into the long run. The rise in price causes new firms to enter by creating profit for firms already in the industry. An *increase in the number of firms* is a shift variable for the supply curve. This is why, in the long run, the supply curve shifts from S_1 to S_2.

FIGURE 10 A Constant Cost Industry



At point A *in panel (a), the market is in long-run equilibrium. The typical firm in panel (b) operates at the minimum of its* ATC *and* LRATC *curves, and earns zero economic profit. The lower panels show what happens if demand increases. In the short run, the market reaches a new equilibrium at point* B *in panel (c), and the typical firm in panel (d) earns economic profit at the higher price* P_{SR}. *In the long run, profit attracts entry, increasing market supply and lowering price. Entry continues until economic profit at the typical firm in panel (d) is reduced to zero, which requires the price to drop to* P_1, *its original level. In panel (d), the typical firm returns to point* A, *and in panel (c), the new long-run market equilibrium is point* C. *The increase in demand raises output, but leaves price unchanged, as shown by the horizontal long-run supply curve connecting points* A *and* C.

earning economic profit, and new firms would still be entering.) Our new long-run equilibrium occurs at point C, with the supply curve S_2, price P_1, and market quantity Q_2. Panel (d) shows what happens at the typical firm: The price moves back to P_1, so the demand curve facing the firm shifts back to d_1, and the typical firm returns to its original level of output q_1.

There is a lot going on in Figure 10. But we can make the story simpler if we *skip over* the short-run equilibrium at point *B*, and just ask: What happens in the *long run* after the demand curve shifts rightward? The answer is: The market equilibrium will move from point *A* to point *C*. A line drawn through these two points tells us, in the long run, the market price we can expect for any quantity the market provides. In Figure 10, this is the thin line, which is called the *long-run supply curve* (S_{LR}).

Long-run supply curve A curve indicating price and quantity combinations in an industry after all long-run adjustments have taken place.

*The **long-run supply curve** shows the relationship between market price and market quantity produced after all long-run adjustments have taken place.*

Notice that, because we are dealing with a constant cost industry, the long-run supply curve is horizontal.

In a constant cost industry, in which industry output has no effect on individual firms' cost curves, the long-run supply curve is horizontal. In the long-run, the industry will supply any amount of output demanded at an unchanged price.

An Increasing Cost Industry

Our trip through Figure 10 illustrated the impact of an increase in demand for a constant cost industry, in which a rise in industry output has no impact on the cost curves of individual firms. But a constant cost industry is just one possible case.

Increasing cost industry An industry in which the long-run supply curve slopes upward because each firm's *LRATC* curve shifts upward as industry output increases.

In an **increasing cost industry**, the entry of new firms that use the same inputs as existing firms drives up input prices. This, in turn, causes each firm's *LRATC* curve to shift upward.

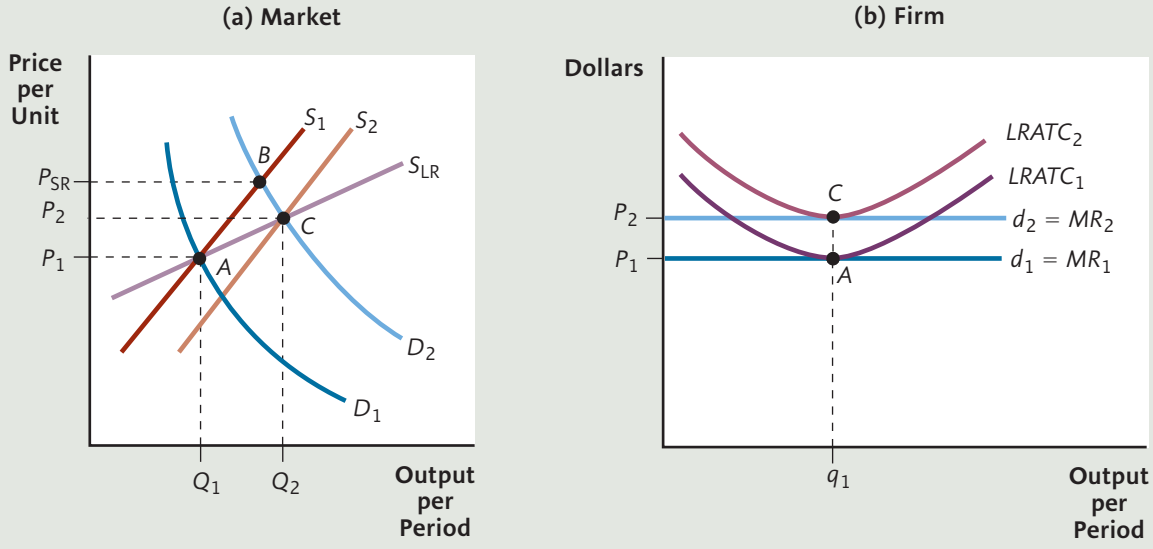
For example, wheat farming uses a great deal of land in the Midwestern United States. If the demand for wheat increased significantly, existing wheat farms would expand, and new farms would enter the industry. The price of farmland would rise. Because *every* farm in the industry—the existing ones as well as new entrants—would have to pay more for farmland, their *LRATC* curves would shift upward (greater cost per unit at each output level).

Let's see how this changes the graphical analysis of an increase in demand. Panel (a) in Figure 11 shows a competitive market in an initial long-run equilibrium at point *A*. Panel (b) shows the situation of a single competitive firm in this market, facing demand curve d_1 and producing output level q_1. To keep the diagram simple, we've left out the *MC* and *ATC* curves for the firm and show the only cost curve that will matter to our analysis: the *LRATC* curve. Initially, the firm operates at the minimum point of $LRATC_1$.

Now suppose the demand curve shifts rightward to D_2 [panel (a)]. As a result, the short-run market equilibrium moves to point *B*, and price rises to P_{SR}. Because the typical firm enjoys economic profit (not shown), entry will occur in the long run, and the market supply curve shifts rightward. As usual, the supply curve will continue shifting rightward until economic profit is eliminated.

But this time, the entry of new firms and the rise in industry output causes the typical firm's *LRATC* curve to shift *upward* to $LRATC_2$. With higher long-run average cost, zero profit will occur at a price *higher* than the original price P_1. In Figure 11, the supply curve stops shifting when the price reaches P_2, with the new market equilibrium at point *C*. As panel (b) shows, once the price reaches P_2, the typical firm—facing the horizontal demand curve d_2—operates at the minimum point on $LRATC_2$, earning zero economic profit.

FIGURE 11 An Increasing Cost Industry



Point A *in both panels shows the initial long-run market equilibrium, with the typical firm earning zero economic profit. After demand increases, the market reaches a new short-run equilibrium at point* B *in panel (a). At the higher price, the typical firm earns economic profit (not shown). In the long run, profit attracts entry, supply increases and price begins to fall. But in an increasing cost industry, the rise in industry output also causes costs to rise, shifting up the* LRATC *curve. In the final, long-run market equilibrium (point* C *in both panels), price at* P_2 *is higher than originally, and the typical firm once again earns zero economic profit. The increase in demand raises both output and price, as shown [in panel (a)] by the upward-sloping long-run supply curve.*

Let's now concentrate on just the long-run impact of the change in demand, which moves the equilibrium from point *A* to point C. Connecting these two equilibrium points gives us the long-run supply curve for this industry. As you can see, the curve slopes *upward,* telling us that the industry will supply greater output, but only with a higher price.

In an increasing cost industry, *a rise in industry output shifts up each firm's* LRATC *curve, so that zero economic profit occurs at a higher price. The long-run supply curve slopes upward.*

The long-run supply curve tells us that an increasing cost industry will deliver more output, but only at a higher price. It also tells us that, if industry output decreases, the price will drop. This is because a decrease in output would cause each firm's *LRATC* curve to shift *downward* so that zero profit would be established at a *lower* price than initially.

A Decreasing Cost Industry

Decreasing cost industry An industry in which the long-run supply curve slopes downward because each firm's *LRATC* curve shifts downward as industry output increases.

In a **decreasing cost industry,** a rise in industry output causes input prices to *fall,* and the *LRATC* curve to shift downward at each firm. This might occur for a number of reasons. As an industry expands, there might be more workers in the area with

dangerous curves

Similar-Sounding Terms, but Different Concepts You've learned a number of different terms having to do with rising costs. A direct comparison can help to prevent confusing them.

Diseconomies of scale (from Chapter 7) refers to a rise in long-run average cost due to an increase in a firm's *own* output. This is illustrated by a movement *along* the upward sloping portion of the firm's *LRATC* curve.

Increasing-cost industry (from this chapter) refers to an upward *shift* of the entire *LRATC* curve, due to higher input prices caused by entry.

Increasing marginal cost refers to the upward-sloping portion of the firm's marginal cost curve. It is a short run concept, associated with diminishing marginal returns (see Chapter 7).

Each of these terms has its opposite, having to do with *falling* costs: economies of scale, decreasing cost industry, and decreasing marginal cost. You might want to test yourself by explaining how these three terms differ from one another.

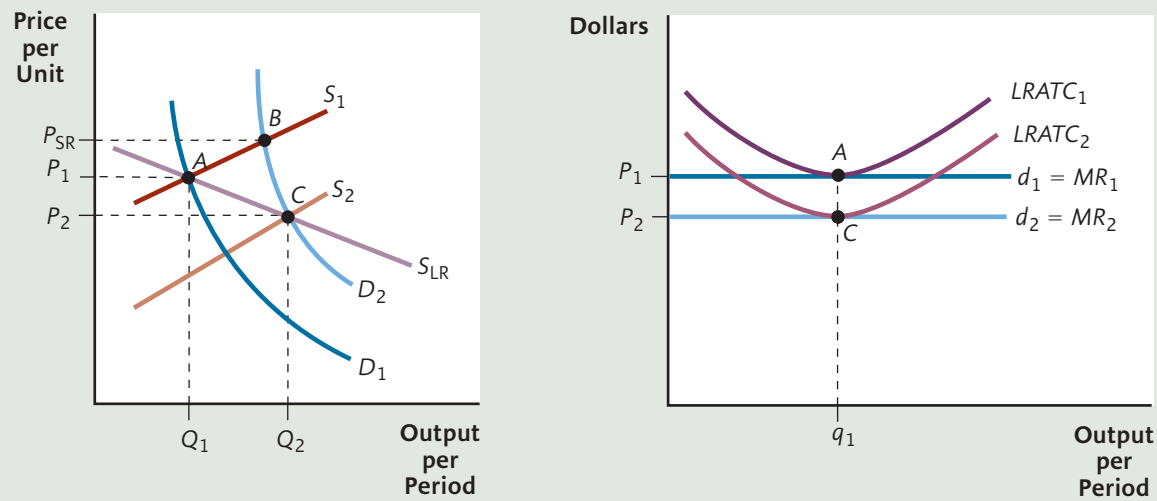
the needed skills, making it easier and less expensive for each firm to find and recruit qualified employees. Or transportation costs might decrease.

For example, suppose that a modest size city has just a few sushi restaurants. Periodically, a partially loaded truck makes a special trip from a distant larger city to deliver raw fish, nori seaweed, wasabi, and other special ingredients to these few restaurants. Transportation costs—part of the price of the ingredients—will be rather high.

Now suppose that demand for sushi meals increases. Profits at the existing restaurants attract entry. With more restaurants ordering ingredients, the same delivery truck makes the same trip, but now it is fully loaded and the transportation costs are shared among more restaurants. As a result, transportation costs at *each* restaurant decrease—and each restaurant's *LRATC* curve shifts down. Competition among the restaurants then ensures that prices will drop to match the lower *LRATC*. As a result, the long-run effect of an increase in demand is a *lower* price for eating sushi at a restaurant—a downward sloping long-run supply curve.

Figure 12 illustrates how a decreasing cost industry behaves after an increase in demand. In panel (a), after the demand curve shifts rightward, the market equilibrium

FIGURE 12 A Decreasing Cost Industry



Point A *in both panels shows the initial long-run market equilibrium, with the typical firm earning zero economic profit. After demand increases, the market reaches a new short-run equilibrium at point B in panel (a). At the higher price, the typical firm earns economic profit (not shown). In the long run, profit attracts entry, supply increases and price begins to fall. But in a decreasing cost industry, the rise in industry output causes costs to fall, shifting down the* LRATC *curve. In the final, long-run market equilibrium (point* C *in both panels), price at* P_2 *is lower than originally, and the typical firm once again earns zero economic profit. The increase in demand raises output but lowers price, as shown [in panel (a)] by the downward-sloping long-run supply curve.*

moves from *A* to *B* in the short run. The typical firm earns economic profit (not shown). In the long run, profit causes entry. But now, as the industry expands, the *LRATC* curve at each firm shifts *downward*. With lower cost per unit, zero economic profit occurs at a long-run equilibrium price *lower than the original price*. In the figure, the market reaches its new long-run equilibrium at point *C*, at the new, lower price P_2.

When we draw a line through the initial equilibrium at point *A* and the new long-run equilibrium at point *C*, we get the long-run supply curve for this industry. As you can see, the curve slopes downward: In a decreasing cost industry, as industry output rises, the *price drops*.

In a decreasing cost industry, *a rise in industry output shifts down each firm's* LRATC *curve, so that zero economic profit occurs at a lower price. The long-run supply curve slopes downward.*

The long-run supply curve tells us that in a decreasing cost industry, the more output produced, the lower the price. On the other hand, if industry output were to fall, the price would rise. This is because a decrease in output would cause each firm's *LRATC* curve to shift *upward*, so that zero profit would be established at a *higher* price than initially.

Market Signals and the Economy

The previous discussion of changes in demand included a lot of details, so let's take a moment to go over it in broad outline. You've seen that an *increase* in demand always leads to an *increase* in market output in the short run, as existing firms raise their output levels, and an even *greater* increase in output in the long run, as new firms enter the market.

We could have also analyzed what happens when demand *decreases*, but you'll be asked to do this on your own in the end-of-chapter problem set. If you do it correctly, you'll find that the leftward shift of the demand curve will cause a drop in output in the short run and an even greater drop in the long run. The effect on price will depend on the nature of the industry, (i.e., whether it is a constant, increasing, or decreasing cost industry).

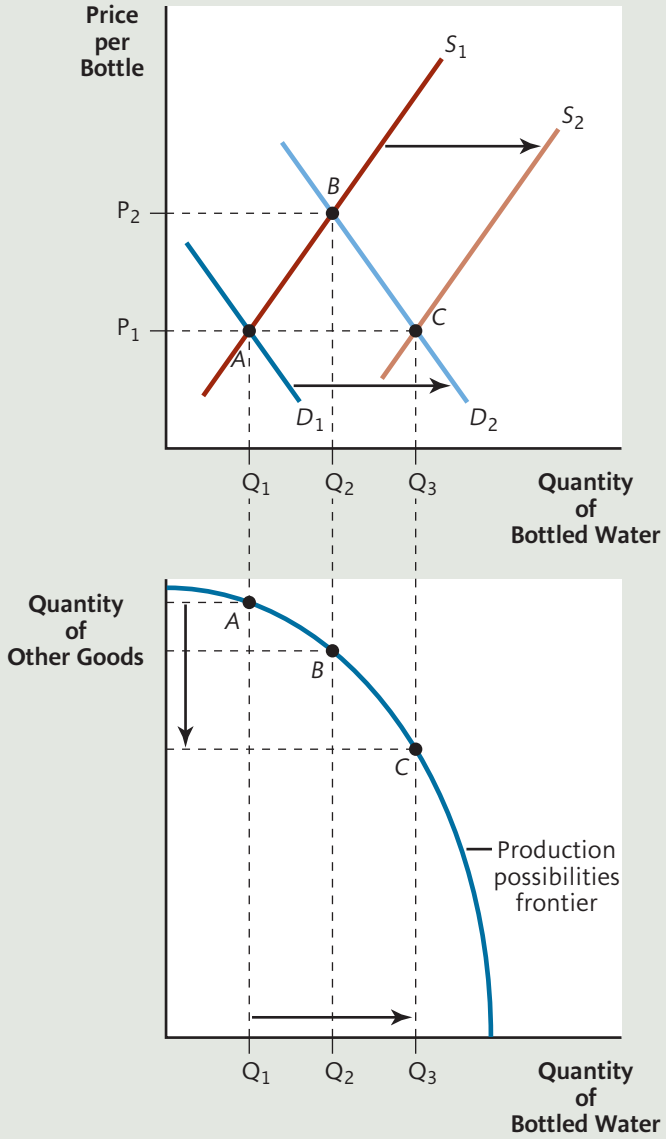
But now let's step back from these details and see what they really tell us about the economy. We can start with a simple fact: In the real world, the demand curves for different goods and services are constantly shifting. For example, over the last couple of decades, Americans have developed an increased taste for bottled water. The average American gulped down 8 gallons of the stuff in 1990, and almost three times that much—22 gallons—in 2006. As a consequence, the *production* of bottled water has increased dramatically. This seems like magic: Consumers want more bottled water and, presto!, the economy provides it. Our model of perfect competition shows us the workings behind this apparent magic, the logical sequence of events leading from our desire to consume more bottled water and its appearance on store shelves.

The secret—the trick up the magician's sleeve—is this: As demand increases or decreases in a market, *prices change*. And price changes act as *signals* for firms to enter or exit an industry. How do these signals work? As you've seen, when demand increases, the price tends to initially *overshoot* its long-run equilibrium value during the adjustment process, creating sizable temporary profits for existing firms. Similarly, when demand decreases, the price falls *below* its long-run equilibrium

FIGURE 13 How a Change in Demand Reallocates Resources

In the upper panel, an increased desire for bottled water shifts the market demand curve rightward, from D_1 to D_2. Price and quantity rise in the short run, and we move from A to B along short-run supply curve S_1. The lower panel shows the corresponding short-run movement from A to B along the economy's PPF: Greater production of bottled water, less production of other things.

In the long run, the higher price creates economic profit, attracting new firms, and shifting the supply curve rightward (upper panel). Price falls and quantity rises further. In the figure, we assume bottled water is a constant cost industry, so entry brings the price back to its initial value of P_1 at point C. In the lower panel, the further long-run increase in bottled water production moves us along the PPF, from B to C.



value, creating sizable losses for existing firms. These exaggerated, temporary movements in price, and the profits and losses they cause, are almost irresistible forces, pulling new firms into the market or driving existing firms out. In this way, the economy is driven to produce whatever collection of goods consumers prefer.

Figure 13 illustrates the process. In the upper panel, as Americans shifted their tastes toward bottled water, the market demand curve for this good shifted rightward from D_1 to D_2. Initially, the price rose *above* its new long-run equilibrium value, to P_2, leading to high profits at existing bottled water firms such as Poland Spring and Arrowhead. High profits, in the long run, attracted entry—especially the

entry of new brands from established firms not previously selling bottled water, such as Pepsi's Aquafina and Coke's Dasani. Entry shifted the supply curve rightward, to S_2, bringing the price back down to P_1. (We are viewing bottled water as a constant cost industry.) As a result, production expanded to match the increase in demand by consumers. More of our land, labor, capital, and entrepreneurial skills are now used to produce bottled water. Where did these resources come from?

In large part, they were freed up from those industries that experienced a *decline* in demand. In these industries, lower prices have caused exit, freeing up land, labor, capital, and entrepreneurship to be used in other, expanding industries, such as the bottled water industry. The lower panel of Figure 13 shows a production possibilities frontier (PPF) for bottled water and other goods. As production of bottled water increases from Q_1 to Q_2 to Q_3, the production of other things decreases. (If we wanted to illustrate economic growth at the same time, the entire PPF would shift outward. In that case, increased production of bottled water would mean that production of other things would *increase* by less than otherwise.)

This leads us to an important observation:

In a market economy, price changes act as ***market signals****, ensuring that the pattern of production matches the pattern of consumer demands. When demand increases, a rise in price signals firms to enter the market, increasing industry output. When demand decreases, a fall in price signals firms to exit the market, decreasing industry output.*

Market signals Price changes that cause changes in production to match changes in consumer demand.

Importantly, in a market economy, no single person or government agency directs this process. There is no central command post where information about consumer demand is assembled, and no one tells firms how to respond. Instead, existing firms and new entrants, in their *own* search for higher profits, respond to market signals and help move the overall market in the direction it needs to go. This is what Adam Smith meant when he suggested that individual decision makers act—as if guided by an *invisible hand*—for the overall benefit of society, even though, as individuals, they are merely trying to satisfy their own desires.

A Change in Technology

You've just learned how production in perfectly competitive markets responds to changes in consumer demand. Demand for a good increases, and production rises. Demand decreases; production falls.

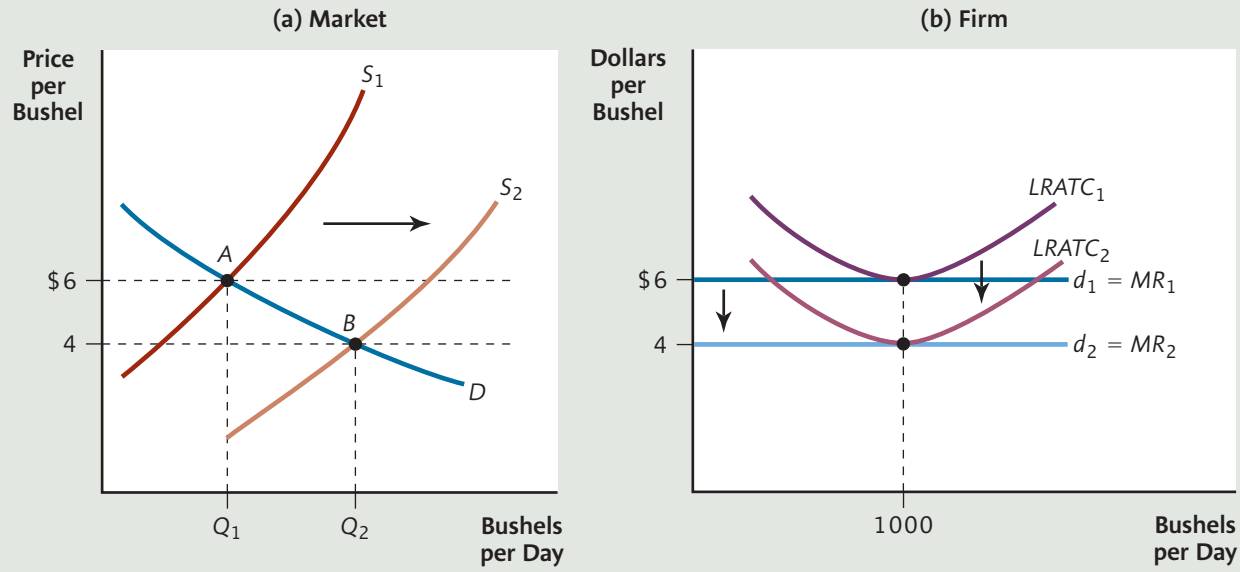
But the service of competitive markets extends to other types of changes as well. In this section, we'll explore how competitive markets ensure that the benefits of technological advances are enjoyed by consumers.

One industry that has experienced especially rapid technological changes in the 1990s and 2000s is farming. By using genetically altered seeds, farmers are able to grow crops that are more resistant to insects and more tolerant of herbicides. This lowers the total—and average—cost of producing any given amount of the crop.

Figure 14 illustrates the market for corn, but it could just as well be the market for soybeans, cotton, or many other crops. In panel (a), the market begins at point *A*, where the price of corn is $6 per bushel. In panel (b), the typical farm produces 1,000 bushels per year and—with long-run average cost curve $LRATC_1$—earns zero economic profit.

Now let's see what happens when new, higher-yield corn seeds are made available. Suppose first that only one farm uses the new technology. This farm will enjoy

FIGURE 14 Technological Change in Perfect Competition



Technological change may reduce LRATC. *In panel (b), the first farms that adopt new technology will earn economic profit if they can sell at the old market price of $6 per bushel. That profit will lead its competitors to adopt the same technology and will also attract new entrants. As market supply increases, price falls until each farm is once again earning zero economic profit.*

a downward shift in its *LRATC* curve from $LRATC_1$ to $LRATC_2$. Since it is so small relative to the market, it can produce all it wants and continue to sell at $6. Although we have not drawn in the farm's *MC* curve, you can see that the farm has several output levels from which to choose where price exceeds cost per unit and it can earn economic profit.

But not for long. In the long run, economic profit at this farm will cause two things to happen. First, all other farmers in the market will have a powerful incentive to adopt the new technology—to plant the new, genetically engineered seed themselves. Under perfect competition, they can do so; there are no barriers that prevent any farmer from using the same technology as any other. As these farms adopt the new seed technology, their *LRATC* curves, too, will drop down to $LRATC_2$.

Second, outsiders will have an incentive to enter this industry, using the new technology, shifting the market supply curve rightward (from S_1 to S_2) and driving down the market price. The process will stop only when the market price has reached the level at which *farms using the new technology* earn zero economic profit. In Figure 14, this occurs at a price of $4 per bushel.

From this example, we can draw two conclusions about technological change under perfect competition. First, what will happen to a farmer who is reluctant to use the new technology? As *other* farms make the change, and the market price falls from $6 to $4, the reluctant farmer will suffer an economic loss, since the farm's average cost will remain at $6. Thus, a farmer that refuses to adopt the new technology will be forced to exit the industry. In the end, *all* farms that remain in the market must use the new technology.

Second, who benefits from the new technology in the long run? Not the farmers who adopt it. *Some* farmers—the earliest adopters—may enjoy *short-run* profit before the price adjusts completely. But in the long run, all farmers will be right back where they started, earning zero economic profit. The gainers are *consumers* of corn, since they benefit from the lower price.

Although the data in this example are hypothetical, the story is not. The average American farmer today feeds far more people than a decade ago, mostly due to technological advances in farming. And as our example suggests, powerful forces push farmers to adopt new productivity-enhancing technology. From 2000 to 2008, the fraction of U.S. corn acreage planted with genetically modified seeds increased from 17 to 80 percent.

More generally, we can summarize the impact of technological change as follows:

Under perfect competition, a technological advance leads to a rightward shift of the market supply curve, decreasing market price. In the short run, early adopters may enjoy economic profit, but in the long run, all adopters will earn zero economic profit. Firms that refuse to use the new technology will not survive.

Technological advances in many competitive industries—mining, lumber, communication, entertainment, and others—have indeed spread quickly, shifting market supply curves rapidly and steadily rightward over the past 100 years. Competitive firms in these industries have had to continually adapt to new technologies in order to survive, leading to huge rewards for consumers.

Using the Theory

© CHINCH GRYNIEWICZ; ECOSCENE/CORBIS

SHORT- AND LONG-RUN ADJUSTMENT IN THE SOLAR POWER INDUSTRY

Sometimes an industry that is constant-cost in the *very* long run will, for some time, behave like an increasing-cost industry. This can occur when there are long lags in stepping up production of needed inputs. The recent behavior of the solar-panel industry provides a good example.

Look first at panel (a) in Figure 15, which shows the market for solar panels. The horizontal axis measures the annual quantity of solar panels produced, measured in megawatts of electricity-generating capacity. The vertical axis measures the price of solar panels in dollars per watt of capacity with the sun at full strength.[3] Point *A* shows the initial equilibrium of the industry in 2003, with Q_1 panels supplied and demanded, at an average price of $4.30 per watt of capacity. In panel (b), we show the typical firm operating at the bottom of its long-run average cost curve, $LRATC_{2003}$, earning normal profit (zero economic profit).

[3] Average retail prices, weighted by module capacity. Source: www.solarbuzz.com/Moduleprices.htm

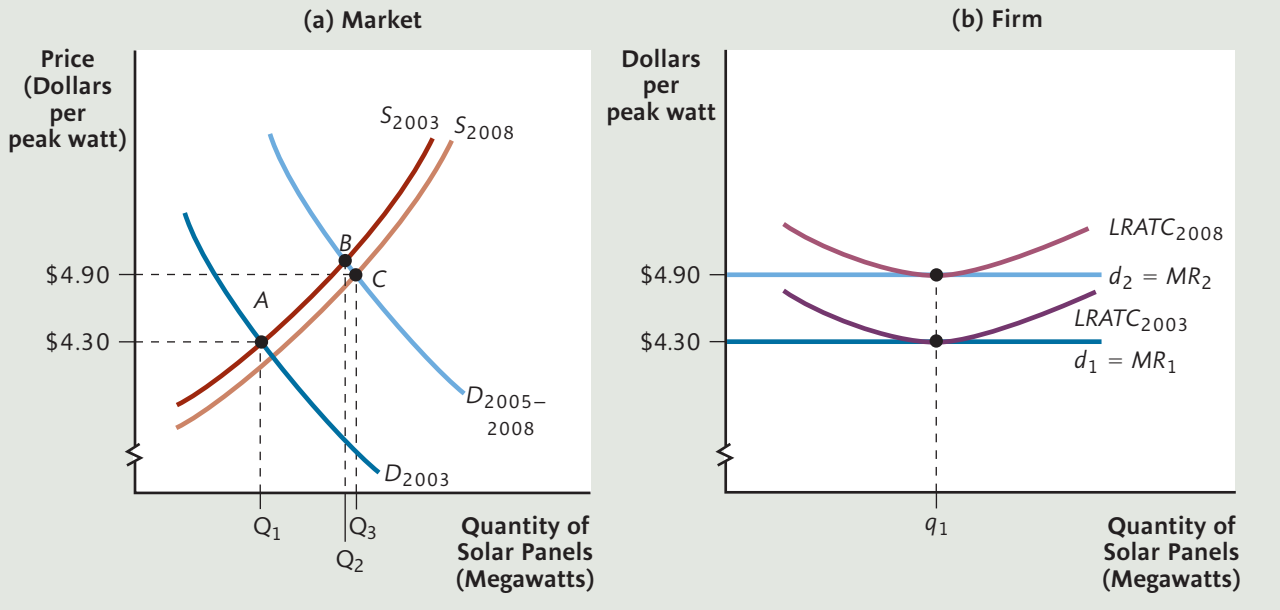
The market might have remained a point like *A* for some time, except that around 2004, several national governments intervened, subsidizing the installation of household solar systems and offering other financial incentives. In the United States, both the federal government and several individual states offered subsidies as well. The result was a significant increase in the demand curve for solar panels, to $D_{2005–2008}$ in panel (a). (The subscript is 2005–2008 because the demand curve shifted to its new position by 2005, and we're assuming it remained in roughly the same position through 2008). This moved the industry to a new short-run equilibrium at point *B* in 2005. The new short-run equilibrium occurs along the original supply curve S_{2003} because entry had not yet occurred, so the supply curve had not yet shifted. In the short run, production rose (to Q_2) and so did the price—to a bit above $4.90. The typical panel-maker was earning economic profit (not shown).

As you've learned, in the long run profits attract entry. And as new firms around the world began entering the solar panel market, the supply curve shifted rightward, reaching S_{2008} a few years later. This increase in supply brought the market price down—a bit. But notice that in the new equilibrium at point *C*, the price of solar panels remained high—about $4.90 per peak watt of capacity.

Why didn't the entry of new solar-panel makers continue until the price returned to its original level? The answer is that the expansion of the industry shifted the typical firm's *LRATC* curve upward, to $LRATC_{2008}$, as shown in panel (b). That is, the solar panel industry behaved as an *increasing-cost* industry during this time. But why did the *LRATC* curves of solar-panel makers shift upward? To answer, we'll need to switch our analysis to another product: *polysilicon*.

Polysilicon in raw form is one of the most abundant raw materials on the earth. But before it can be used by the solar panel industry to make solar cells, it must be processed. Figure 16 shows the market for processed polysilicon.

FIGURE 15 **The Global Market for Solar Panels**



The demand curve, D_{2003}, shows the initial demand by the two main industries that use polysilicon as an input: the semiconductor industry and the solar panel industry. Before the recent increase in demand for solar panels, the polysilicon industry had adjusted its manufacturing capacity to the demands of these two industries, reaching a long-run equilibrium at point *E*. The price for processed polysilicon was about $32 per kilogram.

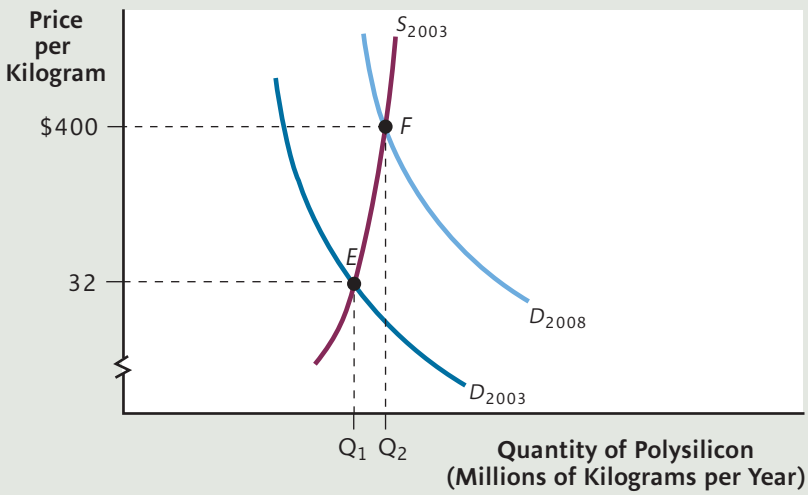
As new solar panel manufacturers began demanding more processed polysilicon, the demand curve for it shifted rightward, to D_{2008}. By this time, the world's polysilicon manufacturers were already operating at full capacity (the near-vertical portion of the supply curve), so further increases in demand resulted in a higher price ($400), but little additional output (point *F*).

Of course, with the price of polysilicon rising, the typical producer of *polysilicon* was earning economic profit (not shown). And as you have probably guessed, this caused entry into the polysilicon-processing industry. But in this case, the entry occurred only after a long time lag. One reason is that it takes more than a year—and a large investment—to set up a processing plant.

Moreover, building such a plant was viewed as highly risky. After all, the new government subsidies that were driving the demand for solar panels could easily be reversed. Moreover, the semiconductor industry is notoriously cyclical; its demand for polysilicon could easily dry up. Firms did not want to take the risk of building new plants to manufacture polysilicon, increasing market supply, only to find that demand had decreased back to its original level. (In that case, with an increase in supply, and a return to the original market demand curve, the price of polysilicon would drop *below* its starting price of $32, and the typical polysilicon firm would experience a loss.) Because of this uncertainty, polysilicon firms refused to expand their polysilicon capacity for several years, and there was little entry of new firms. The supply curve for polysilicon in Figure 16 stayed put at S_{2003}, and the market remained at a point like *F* for several years.

Eventually, the prospect of higher profits, and some legal guarantees from the solar-panel industry, convinced entrepreneurs to build more polysilicon plants. And build they did. From 2003 to 2007, only nine companies produced almost all of the

FIGURE 16 The Global Market for Polysilicon



world's polysilicon. Over the next two years, 60 new producers had entered the market. The result, by mid-2009, was a rightward shift of the supply curve for polysilicon, driving its price down to $75 per kilogram (not shown).

When would the process stop? As of mid-2009, market analysts expected the price of polysilicon to eventually fall back to its 2003 price—around $32 per kilogram—as more processing plants came on line. That is, the expectation is that polysilicon processing will be a constant cost industry. And in that case, the solar panel industry would, in the end, conform to the constant-cost industry model as well. (As an exercise, draw in the new, final equilibrium in Figure 15(a) and (b), as well as Figure 16, to illustrate this constant-cost result for these two industries. If you do this correctly, you will have to pencil in just two entirely new curves).

One last observation: In the constant cost case, would we really expect the price of solar panels to go back to its initial level of $4.30 per watt? Yes . . . *if* the only change in the market were the rightward shift in demand that started off our analysis. But in fact, the solar panel industry undergoes continual *technological* change as well, which drives down production costs. When we classify an industry as a constant-cost industry, we consider the impact of a change in demand *only* (the *ceteris paribus* assumption). If we want to forecast what will *actually* happen to a price, we should take account of *every* relevant change affecting the market—not just a shift in demand. You'll be asked to think about this further in an end-of-chapter problem.

SUMMARY

Perfect competition is a market structure in which (1) there are large numbers of buyers and sellers and each buys or sells only a tiny fraction of the total market quantity; (2) sellers offer a standardized product; (3) sellers can easily enter or exit from the market; and (4) buyers and sellers are well-informed. While few real markets satisfy these conditions precisely, the model is still useful in a wide variety of cases.

Each perfectly competitive firm faces a horizontal demand curve; it can sell as much as it wishes at the market price. The firm chooses its profit-maximizing output level by setting marginal cost equal to the market price. Its *short-run supply curve* is that part of its *MC* curve that lies above the average variable cost curve. Total profit is profit per unit ($P - ATC$) times the profit-maximizing quantity.

In the short run, market price is determined where the market supply curve—the horizontal sum of all firms' supply curves—crosses the market demand curve. In short-run equilibrium, existing firms can earn a profit (in which case new firms will enter) or suffer a loss (in which case existing firms will exit). Entry or exit will continue until, in the long run, each firm is earning zero economic profit. At each competitive firm in long-run equilibrium, price = marginal cost = minimum average total cost = minimum long-run average total cost.

When demand curves shift, prices change more in the short run than in the long run. The temporary, exaggerated price movements act as market signals, ensuring that output expands and contracts in each industry to match the pattern of consumer preferences.

In the long run, an increase in demand can result in a higher, lower, or unchanged market price, depending on whether the good is produced, respectively, in an *increasing cost industry, decreasing cost industry,* or *constant cost industry*. The long-run supply curve slopes upward in an increasing cost industry and slopes downward for a decreasing cost industry. In a constant cost industry, the long-run supply curve will be horizontal.

A technological advance in a perfectly competitive market causes the equilibrium price to fall and equilibrium quantity to rise. Each competitive firm must use the new technology in order to survive, but consumers reap all the benefits by paying a lower price.

PROBLEM SET

Answers to even-numbered questions and problems can be found on the text website at www.cengage.com/economics/hall.

1. Assume that the market for cardboard is perfectly competitive (if not very exciting). In each of the following scenarios, should a typical firm continue to produce or should it shut down in the short run? Draw a diagram that illustrates the firm's situation in each case.
   a. Minimum *ATC* = $2.00
      Minimum *AVC* = $1.50
      Market price = $1.75
   b. *MR* = $1.00
      Minimum *AVC* = $1.50
      Minimum *ATC* = $2.00

2. Suppose that a perfectly competitive firm has the following total variable costs (*TVC*):

Quantity:	0	1	2	3	4	5	6
TVC:	$0	$6	$11	$15	$18	$22	$28

   It also has total fixed costs (*TFC*) of $6. If the market price is $5 per unit:
   a. Find the firm's profit-maximizing quantity using the marginal revenue and marginal cost approach.
   b. Check your results by re-solving the problem using the total revenue and total cost approach. Is the firm earning a positive profit, suffering a loss, or breaking even?

3. "*A profit-maximizing* competitive firm will produce the quantity of output at which price *exceeds* cost per unit by the greatest possible amount." True or false? Explain briefly. [Hint: See Figure 3(a).]

4. The following table gives quantity supplied and quantity demanded at various prices in the perfectly competitive meat-packing market:

Price (per lb.)	Q_S **(in millions of lbs.)**	Q_D
$1.00	10	100
$1.25	15	90
$1.50	25	75
$1.75	40	63
$2.00	55	55
$2.25	65	40

   Assume that each firm in the meat-packing industry faces the following cost structure:

Pounds	***TC***
60,000	$110,000
61,000	$111,000
62,000	$112,000
63,000	$115,000

   a. What is the profit-maximizing output level for the typical firm? (Hint: Calculate *MC* for each change in output, then find the equilibrium price, and calculate *MR* for each change in output.)
   b. Is this market in long-run equilibrium? Why or why not? (Hint: Calculate *ATC*.)
   c. What do you expect to happen to the number of meat-packing firms over the long run? Why?

5. Assume that the kitty litter industry is perfectly competitive and is presently in long-run equilibrium:
   a. Draw diagrams for both the market and a typical firm, showing equilibrium price and quantity for the market, and *MC, ATC, AVC, MR,* and the demand curve for the firm.
   b. Your friend has always had a passion to get into the kitty litter business. If the market is in long-run equilibrium, will it be profitable for him to jump in headfirst (so to speak)? Why or why not?
   c. Suppose people begin to prefer dogs as pets, and cat ownership declines. Show on your diagrams from part (a) what happens in the industry and the firm in the long run, assuming that this is a constant cost industry.

6. In a perfectly competitive, increasing cost industry, is the long-run supply curve always flatter than the short-run market supply curve? Explain.

7. A student says, "My economics professor must be confused. First he tells us that in perfect competition, the demand curve is completely flat—horizontal. But then he draws a supply and demand diagram that has a downward-sloping demand curve. What gives?" Resolve this student's problem in a single sentence.

8. Assume that the firm shown in the following table produces output using one fixed input and one variable input.

Output	**Price**	**Total Revenue**	**Marginal Revenue**	**Total Cost**	**Marginal Cost**	**Profit**
0	$50	$0		$5		
					$35	
1	$50					
					$15	
2	$50					
					$35	
3	$50					
					$55	
4	$50					
					$65	

a. Complete this table and use it to find this firm's short-run profit-maximizing quantity of output. How much profit will this firm earn?
b. Redo the table and find the profit-maximizing quantity of output, if the price of the firm's fixed input rose from $5 to $10. How much profit will this firm earn now?
c. Now redo the original table and find the profit-maximizing quantity of output, if the price of the firm's variable input rose so that *MC* increased by $20 at each level of output. How much profit will this firm earn in this case?

9. Assume that the firm shown in the following table produces output using one fixed input and one variable input.
   a. Complete this table and use it to find this firm's short-run profit-maximizing quantity of output. How much profit will this firm earn?
   b. Redo the table and find the profit-maximizing quantity of output, if the price of the firm's fixed input fell by half. How much profit will this firm earn now?
   c. Now redo the original table and find the profit-maximizing quantity of output, assuming the price of the variable input drops, and each *MC* value is 50% lower than before. How much profit will the firm earn in this case?

Output	Price	Total Revenue	Marginal Revenue	Total Cost	Marginal Cost	Profit
0	$3500	$0		$1000		
					$ 4000	
1	$3500					
					$ 3000	
2	$3500					
					$ 2000	
3	$3500					
					$ 1000	
4	$3500					
					$ 3000	
5	$3500					
					$ 4000	
6	$3500					
					$ 9000	
7	$3500					
					$36,000	
8	$3500					

10. Figure 11 shows the short-run and long-run adjustment process for an increasing cost industry responding to an increase in demand. Draw a similar two-panel diagram, illustrating the response of an increasing cost industry to a *decrease* in demand. Draw in the long-run supply curve. In the long run, will the price be higher or lower, compared to the initial price?

11. Figure 12 shows the short-run and long-run adjustment process for a decreasing cost industry responding to an increase in demand. Draw a similar two-panel diagram, illustrating the response of a decreasing cost industry to a *decrease* in demand. Draw in the long-run supply curve. In the long run, will the price be higher or lower, compared to the initial price?

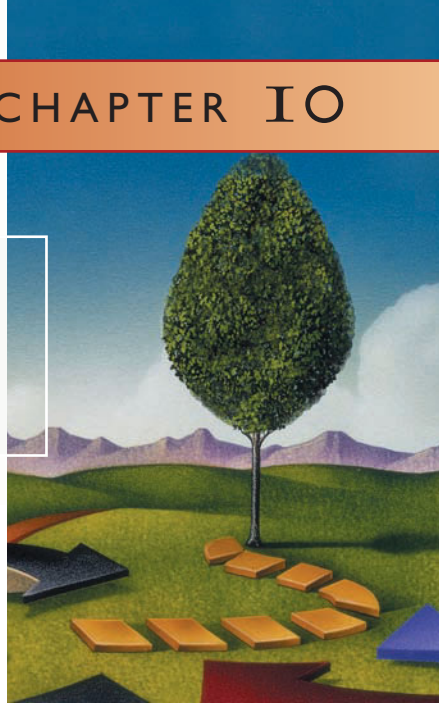
More Challenging

12. In the Using the Theory section, we suggested that the solar panel industry, over the very long run, ultimately conforms to the constant-cost industry model. But it also experiences cost-lowering technological change over the long run.
   a. With diagrams similar to those in Figure 15(a) and (b), show what happens after an increase in demand for solar panels in the short run and the long run. But (unlike in the figure) assume solar panels are a constant cost industry, and there is also long-run technological change.
   b. In the new, final long-run equilibrium, will the price of solar panels be higher or lower than it was initially?

13. Draw a diagram for a perfectly competitive firm in long run equilibrium. Include only the demand curve facing the firm and its *LRATC* curve. Then show the impact of an excise tax (some number of dollars per unit) imposed by the government *on this firm only* but not on any other firm in the market. Can we say what this firm will do in the long-run?

14. In Chapter 4, you learned that when an excise tax is imposed on buyers or sellers in a competitive market, the equilibrium price rises, and the tax payment is shared between buyers and sellers. To obtain that result, we used the (short-run) market supply curve. Now let's extend the analysis to the long run. Draw a two panel diagram: one panel for the market (demand and short-run supply curves only), the other panel for the typical firm (demand and *LRATC* curves only). Suppose an excise tax (some number of dollars per unit) is imposed on *all* sellers (firms) in this market.
   a. Show what will happen in the *market* in the short run.
   b. Show what will happen in *both* diagrams (market and typical firm) in the long run, assuming that this is a *constant cost industry.* [Hint: After the tax is imposed, will the typical firm earn a profit or suffer a loss? Will entry or exit occur?]
   c. In the long run, do both buyers and sellers share in the payment of the excise tax? Explain briefly.

CHAPTER 10



Monopoly

"Monopoly" is as close as economics comes to a dirty word. It is often associated with thoughts of extraordinary power, unfairly high prices, and exploitation. Even in the board game *Monopoly,* when you take over a neighborhood by buying up adjacent properties, you exploit other players by charging them higher rent.

The negative reputation of monopoly is in many ways deserved. A monopoly, as the only firm in its market, has the power to act in ways that a perfectly competitive firm cannot. Adam Smith's "invisible hand"—which channels the behavior of perfectly competitive firms into a socially beneficial outcome—doesn't poke, prod, or even lay a finger on a monopoly firm. Left unchecked, it will *not* create the best of all possible worlds for consumers. Indeed, when a monopoly "takes over" a previously competitive industry, great harm can be done to consumers and society in general. Monopolies, therefore, present a problem that nations around the world address with the very *visible* hand of government policy.

At the same time, a mythology has developed around monopolies. The media often portray their power as absolute and unlimited, and their behavior as capricious and unpredictable. As you are about to see, this characterization goes too far. A monopoly's power may be formidable, but it's far from unlimited. And monopoly behavior—far from capricious—is remarkably predictable.

This chapter will help you understand what monopolies are, how they arise, how they behave, and how they respond to changing market conditions. Our focus here will be on *understanding* monopolies and *predicting* their behavior. A fuller assessment of monopolies and policies for dealing with them will be provided in Chapters 15 and 16.

What Is a Monopoly?

In most of your purchases—a haircut, a meal at a restaurant, a car, a college education—more than one seller is competing for your dollars, and you can choose which one to buy from. But in some markets, you have no choice at all. If you want to mail a letter for normal delivery, you must use the U.S. Postal Service. In most American towns (at least for now), if you want cable television service, you must use the one cable television company in your neighborhood. Many cities have only a single local newspaper. And if you live in a very small town, you may have just one doctor, one gas station, or one movie theater to select from. These are all examples of *monopolies:*

A ***monopoly*** *is a market with just one seller.*

Monopoly The only seller in a market, or a market with just one seller.

The term *monopoly* is used for both the market and the firm that operates in that market.

Classifying a real-world firm or market as a monopoly can be tricky, because the number of sellers depends on how broadly a market is defined. Suppose, for example, that you live in a city or town with just one daily newspaper. Is that newspaper a monopoly?

That depends. If we define the market very broadly as "the market for current news," then there are many other providers in the market (including news from television, radio, and the Internet). In this case, the one newspaper in the town would *not* be a monopoly, because there are other providers in the market. But with a more narrow definition—"the market for printed news delivered to the home," then the only newspaper in town is, indeed, a monopoly.

In practice, we usually define a market to include all *close substitutes* for a product. If for most people, Internet news and daily newspapers are close substitutes, we should include them in the same market. If they are more distant substitutes, we should regard them as separate markets.

It makes sense, then, to view monopoly as a spectrum rather than a strict category. On one end of this spectrum is *pure monopoly,* where there is just one seller of a good for which very few buyers could find a substitute. The only doctor, attorney, or food market in a small town comes very close to being a pure monopoly. Further along the spectrum, we reach firms that sell a good for which reasonable substitutes do exist, but they are not very *close* substitutes for most buyers or most purposes. The sole local cable company is an example of this middle ground because the currently available substitutes (satellite, broadcast, or Internet video) are too different to be considered close substitutes. So most economists would extend the label "monopoly" (without the "pure") to this part of the spectrum as well.

But as we go further along the spectrum, we find goods for which so many buyers can find close substitutes that the term *monopoly* no longer makes sense. For example, Gap Inc. is the only company that sells Gap jeans. But most people regard other brands of jeans as close substitutes for Gap jeans. Therefore, we would not say that Gap Inc. is a monopoly in the market for jeans because we would not define the market so narrowly as to include only "Gap jeans." In the market for jeans in general, there are many competitors.

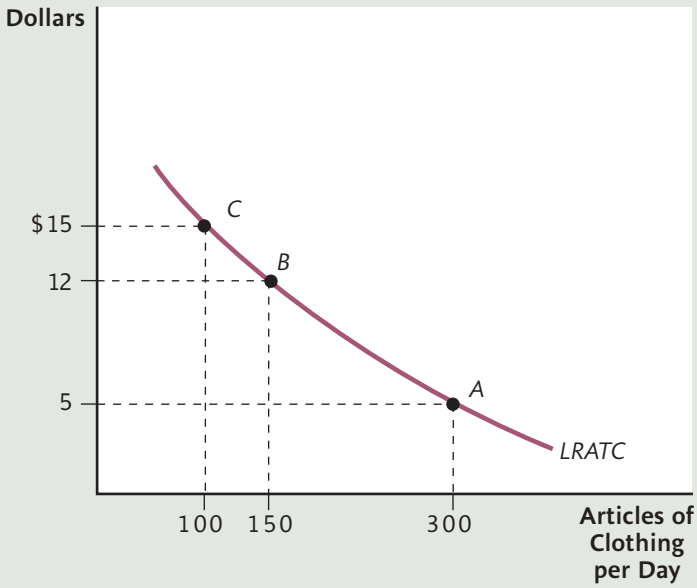
How Monopolies Arise

The mere existence of a monopoly means that *something* is causing other firms to stay out of the market rather than enter and compete with the one firm already there. Broadly speaking, there must be some *barrier to entry.* The question, "Why is the market a monopoly?" then becomes, "What *barrier* prevents other firms from entering the market?" There are several possible answers.

Economies of Scale

One barrier to entry—and thus one explanation for a monopoly—is economies of scale. Recall from Chapter 7 that economies of scale in production causes a firm's long-run average cost curve to slope downward. That is, the more output the firm produces, the lower will be its cost per unit. If economies of scale persists through a large-enough range of output, then a single firm can produce at lower cost than could two or more firms.

FIGURE 1 A Natural Monopoly



In the figure, the typical firm has an LRATC *curve as shown, with economies of scale through an output level of 300, which is assumed to be the total market quantity. A single firm could serve the market at a cost of $5 per unit, operating at point* A. *Two firms splitting this market would each produce 150 units, with each operating at point* B *on its* LRATC *curve. Cost per unit would be $12, higher than with just one firm. Cost per unit would be even higher with three firms. Each would produce 100 units (point* C*), at a cost of $15 per unit. Since a single firm could produce at lower cost than two or more firms, this market tends naturally toward monopoly.*

Figure 1 shows an example: the *LRATC* curve for a dry cleaner in a small town. We'll suppose the entire market for dry cleaning services in this town never exceeds 300 pieces of clothing per day. In the figure, the *LRATC* curve slopes downward, exhibiting economies of scale. Why might this be? A dry cleaning service uses a number of lumpy inputs: a parcel of land for the shop, a store clerk, a small dry cleaning machine if the clothes are cleaned on-site, or daily transportation to an off-site cleaning plant. In the figure, we assume that cleaning more clothes—by spreading these costs among more units—causes cost per unit to decline all the way to 300 units and beyond. As a result, one dry cleaner could achieve a lower cost per unit than could two or more dry cleaners. For example, the *LRATC* curve tells us that one firm could clean 300 pieces of clothing at a cost of $5 per piece (point *A*). But if two dry cleaners were to split this same output level (150 pieces of clothing each), each would have a higher cost per unit of $12 at point *B*. For three dry cleaners, cost per article cleaned would be $15 at point *C*, and so on. The first dry cleaner to locate in the town will have a cost advantage over any potential new entrants. This cost advantage will tend to keep newcomers out of the market.

A monopoly that arises because of economies of scale is called a **natural monopoly.** Local monopolies are often natural monopolies. In a very small town, there might be one gas station, one food market, one doctor, and so on. In all these cases, because there are sizable lumpy inputs and the market is small, the first firm to enter the market will likely be the last.

Natural monopoly A monopoly that arises when, due to economies of scale, a single firm can produce for the entire market at lower cost per unit than could two or more firms.

Legal Barriers

Many monopolies arise because of legal barriers. Of course, since laws are created by human beings, this immediately raises the question: Why would anyone want to create barriers that lead to monopoly? As you'll see, the answer varies depending on the type of

barrier being erected. Here, we'll consider two of the most important legal barriers that give rise to monopolies: protection of intellectual property and government franchise.

Protection of Intellectual Property

The words you are reading right now are an example of *intellectual property,* which includes literary, artistic, and musical works, as well as scientific inventions. The market for a specific intellectual property is a monopoly: One firm or individual owns the property and is the sole seller of the rights to use it. There is both good and bad in this. As you will learn in this chapter, prices tend to be higher under monopoly than under perfect competition, and monopolies often earn economic profit as a consequence. A higher price is good for the monopoly and bad for everyone else.

On the other hand, the promise of monopoly profit is what encourages the creation of original products and ideas in the first place. And this benefits the rest of us. Google's search engine, Apple's iPhone, Visex's cornea-shaping laser, and every film you've seen or novel you've read—all were launched by companies that took on considerable cost and risk for the prospect of future profit.

In dealing with intellectual property, government strikes a compromise: It allows the creators of intellectual property to enjoy a monopoly and earn economic profit, *but only for a limited period of time*. Once the time is up, other sellers are allowed to enter the market, and it is hoped that competition among them will, in the end, bring down the price.

Patent A temporary grant of monopoly rights over a new product or scientific discovery.

The two most important kinds of legal protection for intellectual property are *patents* and *copyrights*. New scientific discoveries and the products that result from them are protected by a **patent** obtained from the government. The patent prevents anyone else from selling the same discovery or product for about 20 years. If someone uses the discovery without obtaining (and paying for) permission from the patent owner, they can be sued. Every year, more than 2,000 patent infringement cases are brought before U.S. courts and thousands more in courts in other countries. Table 1 lists the largest awards granted by U.S. Courts (not including cases settled before final verdict).

TABLE 1

Largest Patent Infringement Awards

Year	Patent Holder	Patent Infringer	Award (in millions)
1990	Polaroid	Eastman Kodak	$909.5
1994	Alpex Computer	Nintendo	$253.6
1996	Haworthv	Steelcase	$211.5
1997	Proctor & Gamble	Paragon Trade Brands	$178.4
1998	Exxon, et al.	Mobil Oil and others	$171.0
2004	Eolas Technologies, et al.	Microsoft	$565.9
2007	Lucent Technologies	Microsoft	$769.0
2009	i4i	Microsoft	$200.0
2009	Rambus	Hynix	$397.0
2009	Uniloc	Microsoft	$338.0

Sources: "Industries Brace for Tough Battle over Patent Law," *Wall Street Journal*, p. 1, June 6, 2007; "Microsoft Told to pay $338 million over Piracy Patent," *Bloomberg.com,* April 8, 2009; "Microsoft Loses $200 Million Jury Verdict in i4i Patent Trial," *Bloomberg.com,* May 21, 2009; "Hynix Revises Down 2008 Earnings on Provisions for Rambus," *Bloomberg.com*, March 11, 2009. Some of the more recent awards are still being contested.

Literary, musical, and artistic works are protected by a **copyright**, which grants exclusive rights over the material for at least 70 years and often longer. For example, the copyright on this book is owned by South-Western/Cengage Learning. No other company or individual can print copies and sell them to the public, and no one can quote from the book at length without obtaining the company's permission.

Copyright A grant of exclusive rights to sell a literary, musical, or artistic work.

Copyrights and patents are often sold to another person or firm, but this does not change the monopoly status of the market, since there is still just one seller. For example, the song "Happy Birthday" was originally written about a century ago, but first received copyright protection in 1935. Since then, the copyright has changed hands numerous times and is currently owned by Warner Music Group. Of course you are free to sing this song at a private birthday party. But anyone who wants to use the song for profit—for example, by featuring it in a radio or television program—must pay a minimal royalty to Warner Music Group (at least until 2030, when the copyright expires).*

Government Franchise

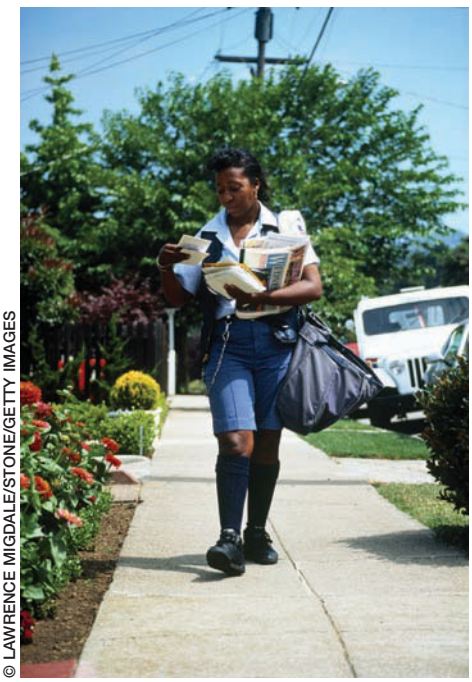
Some firms have their monopoly status guaranteed through **government franchise**, a grant of exclusive rights over a product. Here, the barrier to entry is quite simple: Any other firm that enters the market will be prosecuted!

Government franchise A government-granted right to be the sole seller of a product or service.

Governments often grant franchises when they think the market is a *natural monopoly*. In this case, a single large firm enjoying economies of scale would have a lower cost per unit than multiple smaller firms. Government tries to serve the public interest by *ensuring* that there are no competitors that would cause cost per unit to rise. In exchange for its monopoly status, the seller must submit to either government ownership and control or government regulation over its prices and profits. (We'll have more to say about the regulation of monopoly in Chapter 16.)

This is the logic behind the monopoly status of the U.S. Postal Service. Two postal companies would need many more carriers to deliver the same total number of letters each day, raising costs and, ultimately, the price of mailing a letter. The federal government has chosen to own and control this natural monopoly, rather than merely regulate. Federal law prohibits any other firm from offering normal letter delivery service.

Local governments, too, create monopolies by granting exclusive franchises in a variety of industries believed to be natural monopolies. These include utility companies that provide electricity, gas, and water, as well as garbage collection services.



Since ordinary letter delivery is a natural monopoly, the U.S. Postal Service has been granted an exclusive government franchise to deliver the mail.

Network Externalities

Imagine that you have created a new, superior operating system for personal computers. Compared to Microsoft Windows, your operating system is less vulnerable to viruses, works 10 percent faster, and uses 10 percent less memory. It even allows the user to turn off the caps-lock key, which most people use only by mistake.

*The copyright on Happy Birthday has a long and complex history, and may not even be valid. Although Warner Music Group collects about $2 million per year in royalties on the song, the charge for any one user is so low that no one has mounted a court challenge. For more information, see Robert Brauneis, "Copyright and the World's Most Popular Song," *GWU Legal Studies Research Paper No. 1111624* (2008).

Now all you need is a few million dollars to launch your new product. You manage to get appointments with several venture capital firms, specialists in funding new projects. But every time you make your pitch, and the venture capital people realize what you're proposing, you get the same reaction: hysterical laughter. "But really," you say. "It works better than Windows. I can prove it." "We believe you," they always respond. And they do. And then . . . they start laughing again.

Why? Because you're trying to enter a market with significant *network externalities*.

Network externalities Additional benefits enjoyed by all users of a good or service because others use it as well.

> ***Network externalities*** *are the added benefits for all users of a good or service that arise because other people are using it too.*

When network externalities are present, joining a large network is more beneficial than joining a small network, even if the product in the larger network is somewhat inferior to the product in the smaller one. Once a network reaches a certain size, additional consumers will want to join just because so many others already have. And if joining the network requires you to buy a product produced by only one firm, that firm can rapidly become the leading supplier in the market.[1]

All of this applies to the market for computer operating systems. When you buy a Windows computer, you benefit from the existence of so many other Windows users in a variety of ways. First, you have access to a large number of other computers—owned by friends and coworkers—that you can easily operate. Second, you have access to more software programs (because software developers know they can reach a bigger market when they write programs for Windows). Finally, there are more people around who can help you when you have a problem, saving you the time and trouble of calling a help desk.

In addition to the advantages of *joining* a larger network, there is also an advantage in not leaving it once you've joined: avoiding *switching costs*. It takes time to learn to operate and get used to a new operating system. Although the time you've spent mastering Windows is a sunk cost, the time you'd have to spend mastering a new system can be avoided by staying in the network you're in.

Windows, the first operating system to be used by tens of millions of people, has clearly benefited from network externalities, as well as switching costs. Although two competing operating systems—Apple's OS X and Linux—have managed to establish their own smaller networks, the Windows network remains dominant. And anyone coming up with a better system would have to overcome a vicious circle: Users don't want to switch to the new, superior product until the network grows sufficiently large, which doesn't happen because no one will make the switch.

Network externalities play a major role in the dominance of Facebook as the main social networking site for college students (the more students already in the network, the greater the benefits of joining) and for the continued use of the QWERTY arrangement of letters on keyboards (the more keyboards that already have it, the greater the benefits of having it on your keyboard as well).

[1] The term *externality* will be defined formally in Chapter 15. But if you're curious, an externality is a by-product of a transaction that affects someone other than the buyer or seller (someone external to the transaction). In the case of network externalities, by paying to join the network (e.g., buying a Windows computer), you make the network larger, benefiting others (everyone else with a Windows computer) who weren't involved in your transaction.

Monopoly Behavior

The goal of a monopoly, like that of any firm, is to earn the highest profit possible. And, like other firms, a monopolist faces constraints.

Reread that last sentence because it is important. It is tempting to think that a monopolist—because it faces no direct competitors in its market—is free of constraints or that its constraints are non-economic ones. For example, many people think that the only force preventing a monopolist from charging outrageously high prices is public outrage. In this view, a monopoly cable company would charge $500, $1,000, or even $10,000 per month if only it could "get away with it."

But with a little reflection, it is easy to see that a monopolist faces purely *economic* constraints that limit its behavior—constraints that are in some ways similar to those faced by other, nonmonopoly firms.

First, there is a constraint on the monopoly's *costs:* For any level of output the monopolist might produce, it must pay some total cost to produce it. This cost constraint is determined by the monopolist's production technology and by the prices it must pay for its inputs. In other words, the constraints on the monopolist's costs are the same as on any other type of firm, such as the perfectly competitive firm we studied in the previous chapter.

The monopolist also faces constraints on the price it can charge. This can be a bit confusing because a monopolist, unlike a competitive firm, is *not* a price taker: It does *not* take the market price as a given. But it does face a given demand curve for its product. Indeed, since a monopoly is the only firm in its market, its demand curve is the *market* demand curve. Thus, the monopoly faces a tradeoff: the more it charges for its product, the fewer units it will be able to sell.

Single Price versus Price Discrimination

As you'll see later in this chapter, some firms—including some monopolies—can charge different prices to different consumers, based on differences in the prices they are willing to pay. This kind of pricing is called *price discrimination*. Other firms—we'll call them *single-price firms*—must charge the same price for every unit they sell, regardless of any differences in willingness to pay among their customers. For the next several pages, we'll assume we are dealing with a **single-price monopoly.** (In general, when economists use the term "monopoly" without a modifier, it means "single-price monopoly.") We'll deal with price discrimination later in the chapter.

Single-price monopoly
A monopoly firm that is limited to charging the same price for each unit of output sold.

Monopoly Price or Output Decision

Notice that the title of this section reads "price *or* output decision," not "price *and* output decision." The reason is that a monopoly does not make two separate decisions about price and quantity, but rather *one* decision. Once the firm determines its output level, it has also determined its price (the maximum price it can charge and still sell that output level). Similarly, once the firm determines its price, it has also determined its output level (the maximum output the firm can sell at that price).

How does a monopoly determine its profit-maximizing output level (and therefore its profit-maximizing price)? The same way as any other firm: It considers how a change in output would affect revenue on the one hand and cost on the other. We'll start by exploring the relationship between output and revenue.

TABLE 2

Demand and Revenue at Patty's Pool

Q (swimmers per day)	*P* (admission fee)	*TR*	*MR*
0	$13	$ 0	
			$12
1	$12	$12	
			$10
2	$11	$22	
			$ 8
3	$10	$30	
			$ 6
4	$ 9	$36	
			$ 4
5	$ 8	$40	
			$ 2
6	$ 7	$42	
			$ 0
7	$ 6	$42	
			−$ 2
8	$ 5	$40	
			−$ 4
9	$ 4	$36	

Output and Revenue

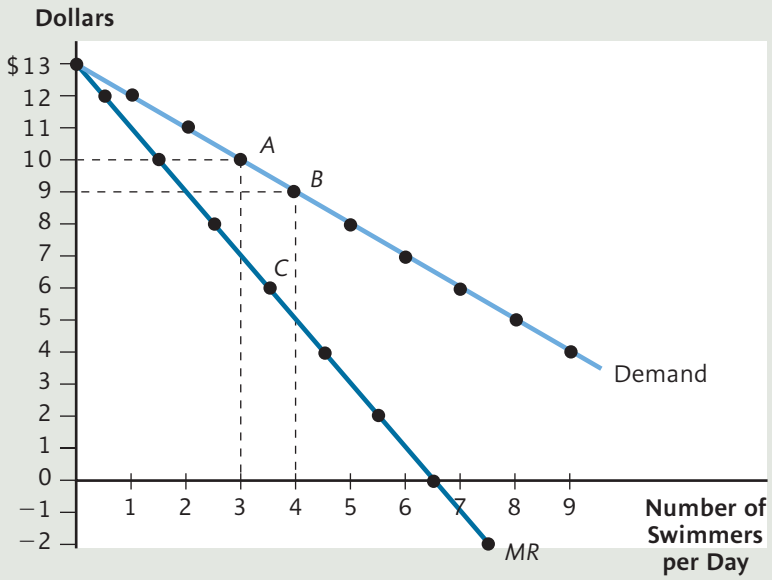
Table 2 shows some data for Patty's Pool, a firm that owns and operates the only swimming pool in a small town—a local monopoly. Patty earns revenue by charging an admission fee for using her pool.

The first two columns show various output levels (swimmers per day) and the highest price (admission fee) Patty could charge for each output level. These two columns tell us that Patty faces a downward-sloping demand curve: The lower the fee, the greater the number of people who will pay to swim each day. This demand curve is graphed in Figure 2 (as the upper curve).

The demand curve should look familiar to you. In Chapter 8, Ned's Beds had to lower its price in order to sell more bed frames. The firm faced a downward-sloping demand curve much like the one shown here. (Ned was not necessarily the only seller in his market, but as you'll see in the next chapter, monopolies are not the only firms that face downward-sloping demand curves.)

The third column of the table shows Patty's total revenue per day (quantity times price) at each output level. For example, at an output level of 3, her daily revenue will be $3 \times \$10 = \$30$. And the last column shows Patty's marginal revenue (*MR*), which is the increase in her revenue for a one-unit rise in output. For example, when Patty's output rises from 3 to 4, her total revenue rises from $30 to $36, so her marginal revenue for this change is $\$36 - \$30 = \$6$.

FIGURE 2 Demand and Marginal Revenue for Patty's Pool



When a firm faces a downward-sloping demand curve, marginal revenue (MR) *is less than price, and the* MR *curve lies below the demand curve. For example, moving from point* A *to point* B*, output rises from 3 to 4 units, while price falls from $10 to $9. For this move, total revenue rises from $30 to $36, so marginal revenue (plotted at point* C*) is only $6—less than the new price of $9.*

The marginal revenue column is graphed in Figure 2, *below* the demand curve. Why below? Mathematically, this is because when the firm's demand curve slopes downward, marginal revenue is less than the price for all increases in output (except the increase from zero to one unit). To see this, look at what happens when we move from point *A* to point *B* along the demand curve, and output rises from 3 to 4 units. The new price is $9, but the marginal revenue from producing the fourth unit is $6 (at point C), which is less than the new price.

When any firm, including a monopoly, faces a downward-sloping demand curve, marginal revenue is less than the price of output. Therefore, the marginal revenue curve will lie below the demand curve.

Why must marginal revenue be less than the price? Because when a firm faces a downward-sloping demand curve, it must lower the price in order to sell a greater quantity. The new, lower price applies to *all* units it sells, including those it was *previously* selling at some higher price.

For example, suppose Patty initially has 3 swimmers per day at $10 each. If she wants 4 swimmers, Table 1 tells us that she must lower her price to $9. Patty would *gain* $9 in revenue by admitting one more swimmer at that price. But she would also *lose* some revenue, because each of the first three swimmers that she *used* to charge $10 will now be charged $9—a *loss* of $3 in revenue. If we add the $9 gained on the fourth swimmer and subtract the $3 lost from lowering the price to the other three, the net impact on revenue is an increase of $6—less than the $9 price she is now charging.

Notice, too, that for increases in output beyond 7, marginal revenue turns negative. For these changes in output, Patty loses more in revenue from dropping the price on previous units than she gains by selling one new unit. No firm would ever want to operate where marginal revenue is negative, because it could then increase its revenue *and* have lower costs by *decreasing* output.

The Profit-Maximizing Output Level

Once we have a monopoly's marginal revenue curve, the profit-maximizing output level can be found by applying our (now familiar) rule from Chapter 8, which tells us how *any* firm can find its profit-maximizing output level:

> *To maximize profit, a monopoly—like any firm—should produce the quantity where* MC = MR *and the* MC *curve crosses the* MR *curve from below.*

Let's apply this rule to a different firm, Zillion-Channel Cable, a monopoly that sells cable television service to the residents of a small city. We'll assume that Zillion-Channel is free from government regulation and is free to set the profit-maximizing price.

In Figure 3, we've plotted the firm's demand curve, showing the number of cable subscribers at each monthly price. As with Patty's Pool, Zillion-Channel's marginal revenue curve lies below its demand curve. The figure also shows Zillion-Channel's marginal cost curve.

The greatest profit possible occurs at an output level of 16,000, where the *MC* curve crosses the *MR* curve from below. In order to sell this level of output, the firm will charge a price of $90, located at point *E* on its demand curve. You can see that for a monopoly, *price and output are not independent decisions, but different ways of expressing the same decision*. Once Zillion-Channel determines its profit-maximizing output level (16,000 units), it has also determined its profit-maximizing price ($90), and vice versa.

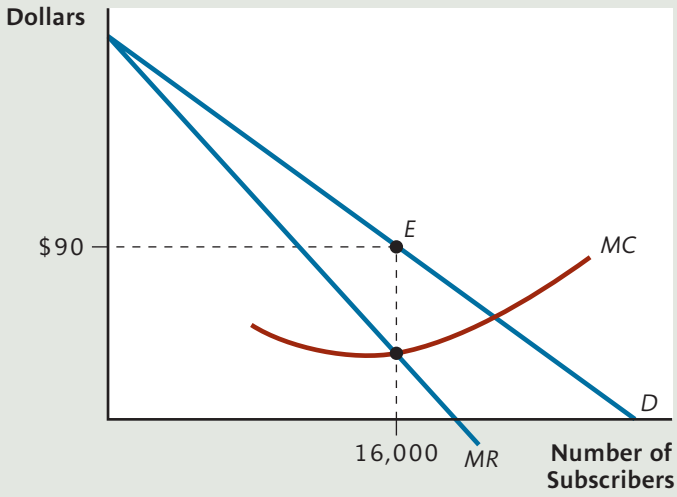
Monopoly and Market Power

Market power The ability of a seller to raise price without losing all demand for the product being sold.

A monopoly is an example of a firm with **market power**—the ability to raise price without causing quantity demanded to go to zero. Any firm facing a downward-sloping demand curve has market power: As it raises its price, quantity demanded

FIGURE 3 Monopoly Price and Output Determination

Like any firm, the monopolist maximizes profit by producing where MC *equals* MR. *Here, that quantity is 16,000 units. The price charged ($90) is read off the demand curve. It is the highest price at which the monopolist can sell the profit-maximizing level of output.*



falls, but some customers who value the firm's product will continue to buy it at the higher price. Only perfectly competitive firms, which face horizontal demand curves, have no market power at all. For a competitive firm, raising price even a tiny bit above the market price reduces quantity demanded to zero. This is why in Chapter 9 we referred to a competitive firm as a *price taker:* It must accept the market price as a given, so there is no decision about price.

By contrast, when a firm has market power, it is a **price setter**—it makes a choice about what price to charge. The choice is limited by constraints (such as the demand curve itself), but it is still a choice. Monopolies are one example of price-setting firms, but they are not the only example. In the next chapter, you'll learn about other market structures besides monopoly in which firms have market power and are therefore price setters.

Price setter A firm (with market power) that selects its price, rather than accepting the market price as a given.

Profit and Loss

In Figure 3, we've illustrated Zillion-Channel's price and output level, but we cannot yet see whether the firm is making an economic profit or loss. This will require one more addition to the diagram—the average cost curve. Remember that

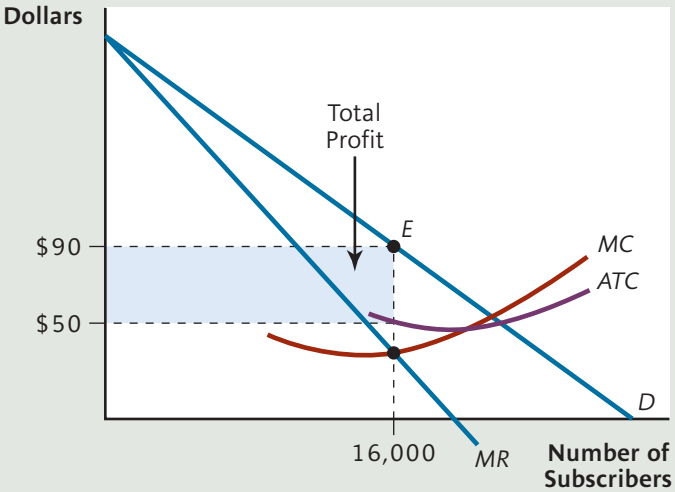
$$\text{Profit per unit} = P - ATC.$$

At any output level, the price is read off the demand curve. Profit per unit, then, is just the vertical distance between the firm's demand curve and its *ATC* curve.

Figure 4 is just like Figure 3 but adds Zillion-Channel's *ATC* curve. As you can see, at the profit-maximizing output level of 16,000, price is $90 and average total cost is $50, so profit per unit is $40.

Now look at the blue rectangle in the figure. The height of this rectangle is profit per unit ($40), and the width is the number of units produced (16,000). The *area* of the rectangle—height × width—equals Zillion-Channel's total profit, or $40 × 16,000 = $640,000.

FIGURE 4 A Monopoly Earning Profit



The monopoly in this figure is earning a profit. At the profit maximizing output level (16,000), profit per unit is equal to the difference between price ($90) and ATC ($50). Total profit is equal to profit per unit multiplied by the number of units, or $40 × 16,000 = $640,000, represented by the blue shaded rectangle.

A monopoly earns a profit whenever P > ATC. *Its total profit at the best output level equals the area of a rectangle with height equal to the distance between* P *and* ATC *and width equal to the level of output.*

dangerous curves

A Monopoly Supply Curve? A question may have occurred to you: Where is the monopoly's *supply curve*? The answer is that *there is no supply curve for a monopoly*. A firm's supply curve tells us how much output a firm will want to produce and sell when it is *presented* with different prices. This makes sense for a perfectly competitive firm that takes the market price as given and responds by deciding how much output to produce. A monopoly, by contrast, is *not* a price taker; It *chooses* its price. Since the monopolist is free to choose any price it wants—and it will always choose the *profit-maximizing* price and no other—the notion of a supply curve does not apply to a monopoly.

This should sound familiar: It is exactly how we represented the profit of a perfectly competitive firm (compare with Figure 3(a) in Chapter 9). The diagram looked different under perfect competition because the firm's demand curve was horizontal, whereas for a monopoly it is downward sloping.

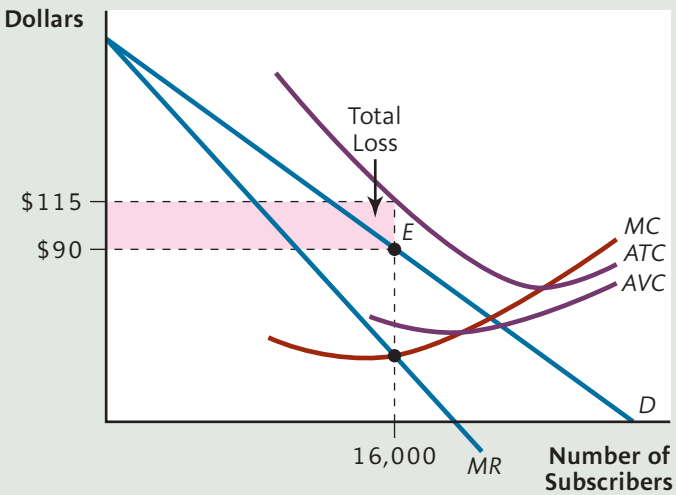
Figure 5 illustrates the case of a monopoly suffering a loss. Here, costs are higher than in Figure 4, and the *ATC* curve lies everywhere above the demand curve. As a result, the firm will suffer a loss at any level of output. At the best output level (where $MC = MR$), the loss will be smallest. In the figure, this occurs at 16,000 units, with $ATC = \$115$ and price = \$90, so the loss per unit is \$25. The total loss (\$400,000) is the area of the pink rectangle, whose height is the loss per unit (\$25) and width is the best output level (16,000).

As you can see, being a monopolist is no guarantee of profit. If costs are too high, or demand is insufficient, a monopolist may break even or suffer a loss.

A monopoly suffers a loss whenever P < ATC. *Its total loss at the best output level equals the area of a rectangle with height equal to the distance between* ATC *and* P *and width equal to the level of output.*

FIGURE 5 A Monopoly Suffering a Loss

The monopoly in this figure is suffering a loss. At the profit maximizing output level (16,000), the loss per unit is equal to the difference between price (\$90) and ATC (\$115). The total loss is equal to loss per unit multiplied by the number of units, or \$25 × 16,000 = \$400,000, represented by the pink shaded rectangle.



Equilibrium in Monopoly Markets

A monopoly market is in equilibrium when the only firm in the market, the monopoly firm, is maximizing its profit. After all, once the firm is producing the profit-maximizing quantity—and charging the highest price that will enable it to sell that quantity—it has no incentive to change either price or quantity, unless something in the market changes (which we'll explore later).

But for monopoly, as for perfect competition, we have different expectations about equilibrium in the short run and equilibrium in the long run.

Short-Run Equilibrium

In the short run, a monopoly may earn an economic profit or suffer an economic loss. (It may, of course, break even as well; see if you can draw this case on your own.) A monopoly that is earning an economic profit will, of course, continue to operate in the short run, charging the price and producing the output level at which $MR = MC$, as in Figure 4.

But what if a monopoly suffers a loss in the short run? Then it will have to make the same decision as any other firm: to shut down or not to shut down. The rule you learned in Chapter 8—that a firm should shut down if $TR < TVC$ at the output level where marginal revenue and marginal cost are equal—applies to any firm, including a monopoly. And (as you learned in Chapter 9), the statement "$TR < TVC$" is equivalent to the statement "$P < AVC$." Therefore,

> *any firm—including a monopoly—should shut down if* P < AVC *at the output level where* MR = MC.

In Figure 5, Zillion-Channel is suffering a loss. But since $P = \$90$ and AVC is less than \$90 at an output of 16,000, we have $P > AVC$: The firm should keep operating. On your own, draw in an alternative AVC curve in Figure 5 that would cause Zillion-Channel to shut down. (Hint: It will be higher than the existing AVC curve.)

The shutdown rule should accurately predict the behavior of most privately owned and operated monopolies. But if a monopoly operates under a government franchise or regulation and produces a vital service such as transportation, mail delivery, or mass transit, the government may not allow it to shut down. If, for example, the monopoly suddenly finds that $P < AVC$ at every output level (perhaps because the cost of a variable input suddenly rises or because the demand curve suddenly shifts leftward), the government might order the firm to continue operating, and use tax revenue to cover the loss.

Long-Run Equilibrium

One of the most important insights of the previous chapter was that perfectly competitive firms will *not* earn a profit in long-run equilibrium. Profit attracts new firms into the market, and market production increases. This, in turn, causes the market price to fall, eliminating any temporary profit earned by a competitive firm.

But there is no such process at work in a monopoly market, where barriers *prevent* the entry of other firms into the market. Outsiders will *want* to enter an

industry when a monopoly is earning positive economic profit, but they will be *unable to do so*. Thus, the market provides no mechanism to eliminate monopoly profit.

Unlike perfectly competitive firms, monopolies may earn economic profit in the long run.

What about economic loss? If a monopoly is franchised or regulated by the government, and it faces the prospect of long-run loss, the government may decide to subsidize it in order to keep it running. But if the monopoly is privately owned and controlled, it will not tolerate long-run losses. A monopoly suffering an economic loss that it expects to continue indefinitely should always exit the industry, just like any other firm.

A privately owned, unregulated monopoly suffering an economic loss in the long run will exit the industry, just as would any other business firm. In the long run, therefore, we should not find such monopolies suffering economic losses.

Comparing Monopoly to Perfect Competition

We have already seen one important difference between monopoly and perfectly competitive markets: In perfect competition, economic profit is relentlessly reduced to zero by the entry of other firms; in monopoly, economic profit can continue indefinitely.

But monopoly also differs from perfect competition in another way:

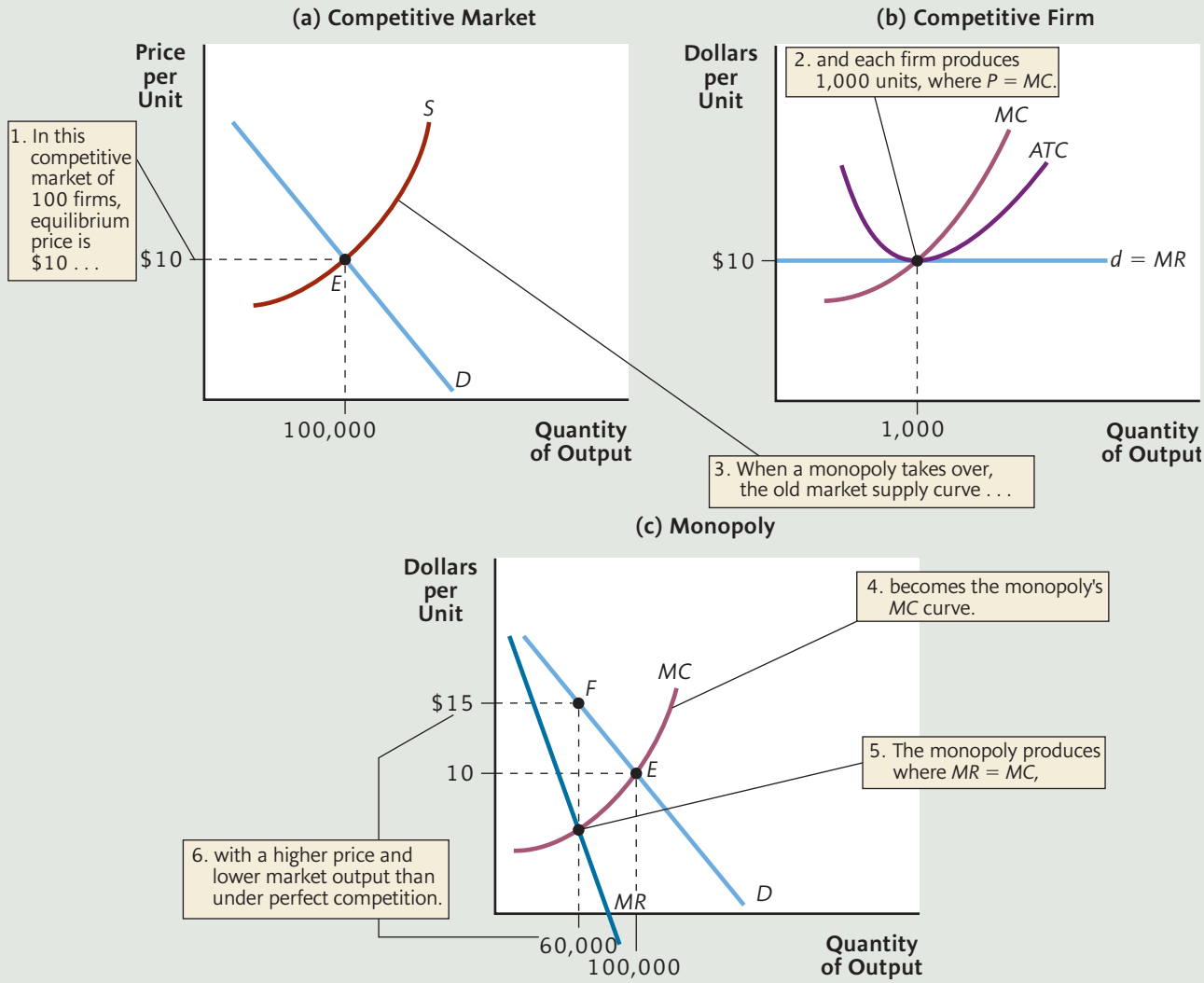
All else equal, a monopoly market will have a higher price and lower output than a perfectly competitive market.

To see why this is so, let's explore what would happen if a single firm took over a perfectly competitive market, changing the market to a monopoly. Panel (a) of Figure 6 illustrates a competitive market consisting of 100 identical firms. The market is in long-run equilibrium at point *E*, with a market price of $10 and market output of 100,000 units. In panel (b), the typical firm faces a horizontal demand curve at $10, produces output of 1,000 units, and earns zero economic profit.

Now, imagine that a single company buys all 100 firms, to form a monopoly. The new monopoly market is illustrated in panel (c). Under monopoly, the horizontal demand curve facing each firm becomes irrelevant. Now, the demand curve facing the monopoly is the downward-sloping *market* demand curve *D*—the same as the market demand curve in panel (a). Since the demand curve slopes downward, marginal revenue will be less than price, and the *MR* curve will lie everywhere below the demand curve. To maximize profit, the monopoly will want to find the output level at which $MC = MR$. But what is the new monopoly's *MC* curve?

We'll assume that the monopoly doesn't change the way output is produced. (This is part of the "all else equal" assumption in the highlighted statement above.) That is, each previously competitive firm will continue to produce its output with

FIGURE 6 Comparing Monopoly and Perfect Competition



the same technology as before, only now it operates as one of 100 different plants that the monopoly controls. With this assumption, *the monopoly's marginal cost curve will be the same as the market supply curve in panel (a)*. Why? First, remember that in a perfectly competitive industry the market supply curve is obtained by adding up each individual firm's supply curve, that is, each individual firm's marginal cost curve. Therefore, the market supply curve tells us the marginal cost—at each firm—of producing another unit of output for the market. When the monopoly takes over each of these individual firms, the market supply curve tells us how much it will cost the monopoly to produce another unit of output at each of its plants.

For example, point *E* on the market supply curve tells us that, when total supply is 100,000, with each plant producing 1,000 units, increasing output by one more unit will cost the monopoly $10 because that is the marginal cost at each of its plants. The same is true at every other point along the old competitive market

supply curve: It will always tell us the new monopoly's cost of producing one more unit at each of the plants it now owns. In other words, the upward-sloping curve in panel (c), which is the market supply curve when the market is competitive, becomes the marginal cost curve for a single firm when the market is monopolized.

Now we have all the information we need to find the monopoly's choice of price and quantity. In panel (c), the monopoly's *MC* curve crosses the *MR* curve from below at 60,000 units of output. This will be the monopoly's profit-maximizing output level. To sell this much output, the monopoly will charge $15 per unit—point *F* on its demand curve.

Notice what has happened in our example: After the monopoly takes over, the price rises from $10 to $15, and market quantity drops from 100,000 to 60,000. The monopoly, compared to a competitive market, *charges more and produces less*.

Why does this happen? When the market was perfectly competitive, each firm could sell all the output it wanted at the given market price of $10, and each firm knew it could earn an additional $10 in revenue for each additional unit it sold. The best option for the firm was to increase output until marginal cost rose to $10.

But the new monopoly does *not* treat price, or marginal revenue, as given values. Instead, it knows that raising its own output *lowers* the market price. So if the monopoly goes all the way to the competitive output level (100,000 units in the figure), it will be producing units for which $MR < MC$ (all units beyond 60,000). This will reduce its profit. To maximize profit, the monopoly has to stop short of the competitive output—producing 60,000 rather than 100,000. Of course, since the monopoly sells a lower market quantity, it will charge a higher market price.

Now let's see who gains and who loses from the takeover. By raising price and restricting output, the new monopoly earns economic profit. We know this because if the firm were to charge $10—the old competitive price—each of its plants would break even, giving it zero economic profit. But we've just seen that $10 is *not* the profit-maximizing price—$15 is. So, the firm must make higher profit at $15 than at $10, ensuring it will earn more than zero economic profit.

Consumers, however, lose in two ways: They pay more for the output they buy, and, due to higher prices, they buy less output. The changeover from perfect competition to monopoly thus benefits the owners of the monopoly and harms consumers of the product.

An Important Proviso

Keep in mind, though, an important proviso concerning this result: Comparing monopoly and perfect competition, we see that price is higher and output is lower under monopoly *if all else is equal*. In particular, we have assumed that after the market is monopolized, the technology of production remains unchanged at each of the monopoly's "plants" (i.e., at each of the previously competitive firms).

But a monopoly may be able to *change* the technology of production, so that all else would *not* remain equal. For example, a monopoly may have each of its new plants *specialize* in some part of the production process, or it may be able to achieve efficiencies in product planning, employee supervision, bookkeeping, or customer relations. These cost savings might shift down the monopoly's marginal cost curve. If you draw a new, lower *MC* curve in panel (c), you'll see that this works to *decrease* the monopoly's price and *increase* its output level—exactly the reverse of the effects discussed earlier. If the cost savings are great enough, and the *MC* curve drops low enough, a profit-maximizing monopoly could even charge a lower price and produce more output than would a competitive market. (See if you can draw a diagram to demonstrate this case.)

The general conclusion is this:

The monopolization of a competitive industry leads to two opposing effects. First, for any given technology of production, monopolization leads to higher prices and lower output. Second, changes in the technology of production made possible under monopoly may lead to lower prices and higher output. The ultimate effect on price and quantity depends on which effect is stronger.

Government and Monopoly Profit

Monopolies, as you learned earlier, often exist with government permission. When we bring the government into our analysis, the monopoly's total profit may be less than that predicted by the analysis we've done so far. Government involvement reduces monopoly profit in two ways.

Government Regulation

As discussed earlier, in many cases of natural monopoly, a firm is granted a government franchise to be the sole seller in a market. This has been true of monopolies that provide water service, electricity, and natural gas. In exchange for its franchise, the monopoly must accept government regulation, often including the requirement that it submit its prices to a public commission for approval. The government will often want to keep prices high enough to keep the monopoly in business, but no higher. Since the monopoly will stay in business unless it suffers a long-run loss, the ideal pricing strategy for the regulatory commission would be to keep the monopoly's economic profit at zero.

Remember, though, that economic profit includes the opportunity cost of the funds invested by the monopoly's owners. If the public commission succeeds, the monopoly's *accounting* profit will be just enough to match what the owners could earn by investing their funds elsewhere—that is, the monopoly will earn zero economic profit. Government regulation of monopoly will be discussed further in Chapter 15.

Rent-Seeking Activity

Another factor that reduces a monopoly's profit comes from the interplay between politics and economics. As we've seen, many monopolies achieve and maintain their monopoly status due to legal barriers to entry. And many of these monopolies are completely unregulated. For example, a movie theater or miniature golf course may enjoy a monopoly in an area because zoning regulations prevent entry by competitors. Or, especially in less developed countries, a single firm may be granted the exclusive right to sell or produce a particular good even though it is not a natural monopoly. In all of these cases, the monopoly is left free to set its price as it wishes.

But legal barriers to entry—for example, zoning laws—are often controversial. After all, as you've learned, a monopoly may charge a higher price and produce less output than would a competitive market. Thus, government will be tempted to pull the plug on a monopoly's exclusive status and allow competitors into the market. The monopoly, in turn, will often take action to *preserve* legal barriers to entry. Economists call such actions *rent-seeking activity*.

Any costly action a firm undertakes to establish or maintain its monopoly status is called ***rent-seeking activity.***

Rent-seeking activity Any costly action a firm undertakes to establish or maintain its monopoly status.

In economics, the term *economic rent* refers to any earnings beyond the minimum needed in order for a good or service to be produced. For example, the minimum price to get *land* "produced" is zero, since it's a gift of nature. This is why all the earnings of landowners are called "rent." A monopoly's economic profit is another example of rent, since it represents earnings above the minimum needed to keep the monopoly in business.

In countries with the most corrupt bureaucracies, rent-seeking activity typically takes the form of outright bribes to government officials. But rent seeking occurs in virtually all countries. It includes the time and money spent lobbying legislators and the public for favorable policies. The costs of such activities can reduce a monopoly's profit below what the simple monopoly model would suggest.

What Happens When Things Change?

Once a monopoly is maximizing profit, it has no incentive to change its price or its level of output . . . unless something that affects these decisions changes. In this section, we'll consider two such events: a change in demand for the monopolist's product, and a change in its costs.

A Change in Demand

Back in Chapter 9 we saw how a competitive market adjusted to a change in demand. In particular, we saw that an increase in demand caused an increase in both market price and market quantity.[2] Does the same general conclusion hold for a monopolist? Let's see.

Panel (a) of Figure 7 shows Zillion-Channel Cable earning a positive profit in the short run. As before, it is producing 16,000 units per month, charging $90 per unit, and earning a monthly profit of $640,000 (not shown). The fact that Zillion-Channel is a monopolist, however, does not mean that it is immune to shifts in demand.

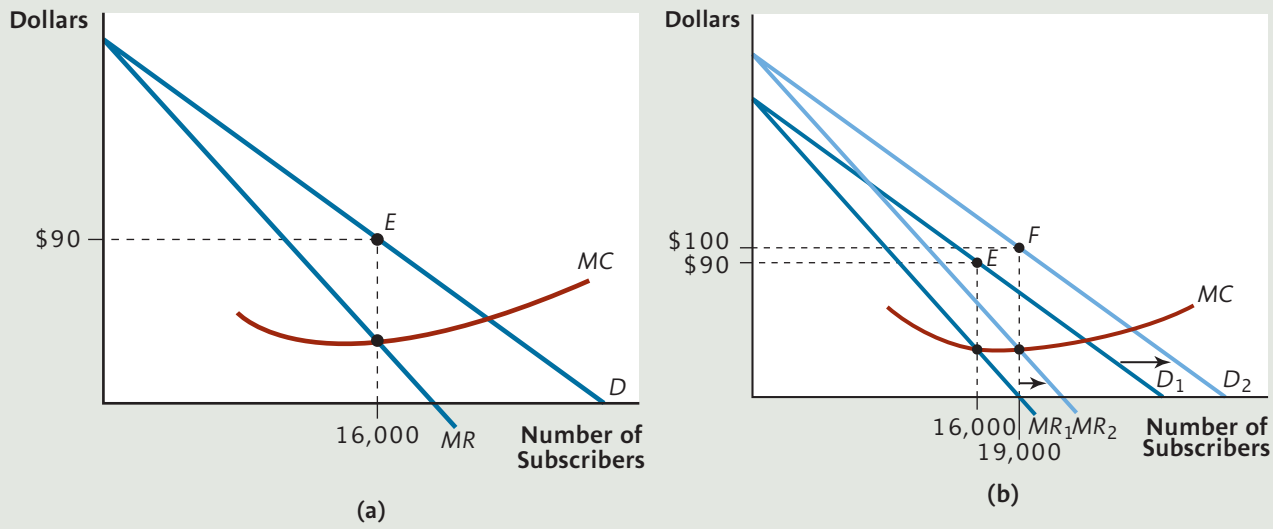
What might cause a monopolist to experience a shift in demand? The list of possible causes is the same as for perfect competition. If you need a reminder of these causes, look back at Figure 4 in Chapter 3. For example, an increase in consumer tastes for the monopolist's good will shift its demand curve rightward, and a decrease in consumer incomes can shift it leftward.

Suppose that the demand for local cable service increases because a sitcom shown on one of Zillion-Channel's premium services attracts an enthusiastic following (an increase in tastes for cable services). In panel (b) of Figure 7, this is shown as a rightward shift of the demand curve from D_1 to D_2.

Notice that the marginal revenue curve shifts as well, from MR_1 to MR_2. Why is this? As you can see in the figure, a rightward shift in demand is also an *upward* shift in demand. At each quantity, the firm can charge a greater price than before. With a higher price, the rise in revenue (MR) for each increase in quantity will be greater as well. So the MR curve shifts upward (rightward), just like the demand curve. (If you want to demonstrate this with numbers, use Table 2. Cross out the price associated with each quantity and write in a price that's $2 greater instead.

[2] One partial exception: in a *decreasing-cost* competitive industry, an increase in demand does raise both price and output in the short run. But in the long run, while output rises further, price drops below its initial value.

FIGURE 7 A Change in Demand



Panel (a) shows Zillion-Channel in equilibrium. It is providing 16,000 units of cable TV service at a price of $90 per month. Panel (b) shows the same firm following an increase in demand from D_1 *to* D_2*. With the increased demand,* MR *is higher at each level of output. In the new equilibrium, Zillion-Channel is charging a higher price ($100), providing more TV service (19,000 units), and earning a larger profit.*

Then, recalculate the *TR* and *MR* columns. You'll see that marginal revenue for each increase in quantity is greater than before.)

With an unchanged cost structure, the new short-run equilibrium will occur where MR_2 intersects the unchanged *MC* curve. As you can see, the result is an increase in quantity from 16,000 to 19,000 and a higher price: $100 per month rather than the original $90. In this sense, monopoly markets behave very much like competitive markets (although the *extent* of the rise in price and quantity will generally *not* be the same as in a competitive market).

What about the monopolist's profit, though? With both price and quantity now higher, total revenue has clearly increased. But total cost is higher as well. (Total cost always rises with greater output.) So it seems as if profit could either rise or fall.

It turns out, however, that profit *must* be higher in the new equilibrium at point *F*. We know that because Zillion-Channel has the option of continuing to sell its original quantity, 16,000, at a price higher than before. If, as we assume, it started out earning a profit at that output level, then the higher price would certainly give it an even *higher* profit. But the logic of $MR = MC$ tells us that the greatest profit of all occurs at 19,000 units. So profit is certainly greater after the increase in demand.

We can conclude that:

A monopolist will generally react to an increase in demand by producing more output, charging a higher price, and earning a larger profit. It will react to a decrease in demand by reducing output, lowering price, and suffering a reduction in profit.

A Cost-Saving Technological Advance

In Chapter 9, you learned that in a perfectly competitive market, all cost savings from a technological advance are passed along to consumers in the form of lower prices. Is the same true of monopoly? Let's see.

Suppose a new type of cable box becomes available that breaks down less often, requiring fewer service calls. When Zillion-Channel Cable begins using this equipment, it finds that it gets fewer service calls, so its labor costs decrease by $15 per customer.

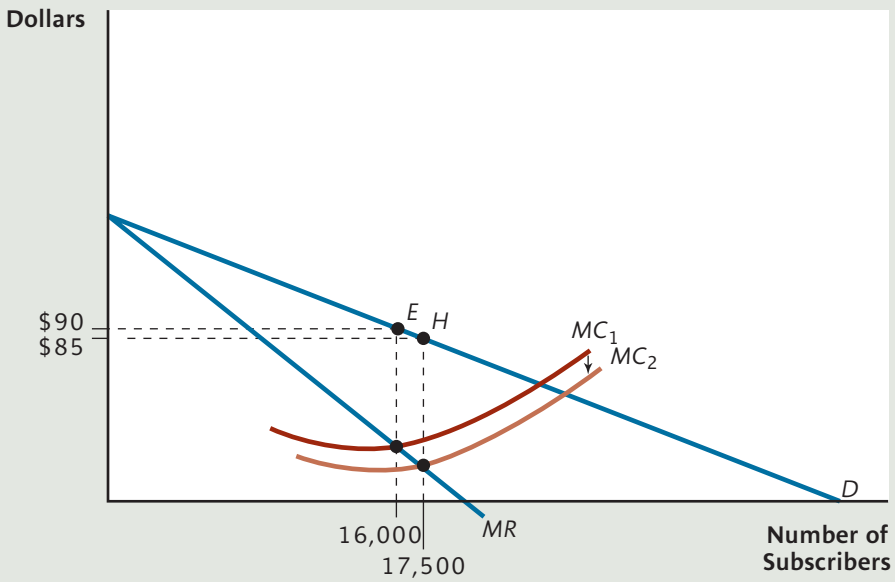
Figure 8 shows the result. Before the new equipment is used, Zillion-Channel is charging $90 and producing output of 16,000, where its *MR* and *MC* curves cross. The technological advance, when it's distributed to all of Zillion's customers, will lower not only the monthly cost per *current* customer (shifting the *ATC* curve down by $15, which isn't shown), but also the monthly cost of servicing each *additional* subscriber. That is, Zillion's *marginal cost* curve will shift down by $15, from MC_1 to MC_2 (which *is* shown).

Zillion-Channel will now want to add subscribers. This is because, after the downward shift in the *MC* curve, *MR* exceeds *MC* at the original output of 16,000. An opportunity to raise profit by increasing output has been created. In the figure, the new intersection point between *MC* and *MR* occurs at an output level of 17,500, so that's Zillion-Channel's new profit-maximizing output level. The demand curve tells us that when output is 17,500, Zillion-Channel will charge a price of $85.

Furthermore, we know that Zillion-Channel's profits have increased. How? If Zillion-Channel had left its output unchanged, the downward shift in its *ATC* curve (not shown) would have raised its profit. Increasing output from 16,000 to 17,500 increased profit further (because $MR > MC$ for that move). Thus, profit must be greater than before.

FIGURE 8 A Cost-Saving Technological Change

A cost-saving technological advance shifts the monopolist's marginal cost curve down, from MC_1 to MC_2. Consumers gain because the price falls, but the drop in price is less than the drop in marginal cost. The monopoly gains because its profit is greater.



Let's summarize what's happened: Zillion-Channel's cost per subscriber decreased by $15, but its price decreased by only $5 (from $90 to $85), and its profit increased. It appears that while consumers do get some benefit from the technological change, they don't get all the benefit. Zillion-Channel keeps a chunk of the benefits for itself (the biggest chunk, in our specific example).

We can summarize our results this way:

In general, a monopoly will pass to consumers only part *of the benefits from a cost-saving technological change. After the change in technology, the monopoly's profits will be higher.*

This stands in sharp contrast to the impact of technological change in perfectly competitive markets, where—as stated earlier—all of the cost saving is passed along to consumers in the long run.

But there's a silver lining for consumers. Suppose that Zillion-Channel's monthly costs *increased* by $15 per subscriber, say, because of a rise in the wage rate needed to maintain its workforce. Figure 8 could be used to analyze this case as well. This time, the *MC* curve would shift *upward* by $15, so we'd view MC_2 as the initial curve and MC_1 as the new one. And while the price of cable service would rise, it would rise by *less* than the $15 increase in cost per unit (in our example, price would rise by only $5). Zillion-Channel would bear part of the burden of the increase in costs, and its profits would fall.

In general, a monopoly will pass only part *of a cost increase on to consumers in the form of a higher price. After the cost increase, the monopoly's profits will be lower.*

Price Discrimination

So far, we've analyzed the decisions of a single-price monopoly—one that charges the same price on every unit that it sells. But not all monopolies operate this way. For example, local utilities typically charge different rates per kilowatt-hour, depending on whether the energy is used in a home or business. Telephone companies charge different rates for calls made by people on different calling plans. Nor is this multiprice policy limited to monopolies: Movie theaters charge lower prices to senior citizens, airlines charge lower prices to those who book their flights in advance, and supermarkets and food companies charge lower prices to customers who clip coupons from their local newspaper.

In some cases, the different prices are due to differences in the firm's costs of production. For example, it may be more expensive to deliver a product a great distance from the factory, so a firm may charge a higher price to customers in outlying areas. But in other cases, the different prices arise not from cost differences but from the firm's recognition that *some customers are willing to pay more than others:*

Price discrimination *occurs when a firm charges different prices to different customers for reasons other than differences in costs.*

Price discrimination Charging different prices to different customers for reasons other than differences in cost.

The term *discrimination* in this context requires some getting used to. In everyday language, *discrimination* carries a negative connotation: We think immediately of discrimination against someone because of his or her race, sex, or age. But a price-discriminating monopoly does not discriminate based on prejudice, stereotypes, or ill will toward any person or group; rather, it divides its customers into different categories based on their *willingness to pay* for the good—nothing more and nothing less. By doing so, a monopoly can squeeze even more profit out of the market. Why, then, doesn't *every* firm practice price discrimination?

Requirements for Price Discrimination

Although every firm would *like* to practice price discrimination, not all of them can. To successfully price discriminate, three conditions must be satisfied:

Market Power. To price discriminate, a firm must have *market power.* That is, the firm must face a downward-sloping demand curve so that it behaves as a price setter. To see why, think about a perfectly competitive firm that faces a horizontal demand curve and has no market power. If such a firm tried to charge some customers a higher price than others, the high-price customers would simply buy from other firms that are selling the same product at the market price. By contrast, all monopolies face a downward sloping demand curve, so they meet the market power requirements.

Identifying Willingness to Pay. In order to determine which prices to charge to which customers, a firm must be able to identify how much different customers or groups of customers are willing to pay. But this is often difficult. Suppose your barber or hairstylist wanted to price discriminate. How would he determine how much you are willing to pay for a haircut? He could *ask* you, but . . . let's be real: You wouldn't tell him the truth, since you know he would only use the information to charge you more than you've been paying. Price-discriminating firms—in most cases—must be a bit sneaky, relying on more indirect methods to gauge their customers' willingness to pay.

For example, airlines know that business travelers, who must get to their destination quickly, are willing to pay a higher price for air travel than are tourists or vacationers, who can more easily travel by train, bus, or car. Of course, if airlines merely *announced* a higher price for business travel, then no one would admit to being a business traveler when buying a ticket. So the airlines must find some way to identify business travelers without actually asking. Their method is crude but reasonably effective: Business travelers typically plan their trips at the last minute and don't stay over Saturday night, while tourists and vacationers generally plan long in advance and do stay over Saturday. Thus, the airlines give a discount to any customer who books a flight several weeks in advance and stays over, and they charge a higher price to those who book at the last minute and don't stay over.[3]

[3] It is sometimes argued that airlines' pricing behavior is based entirely on a cost difference to the airline. For example, it is probably more costly for an airline to keep seats available until the last minute because there is a risk that they will go unsold. The higher price for last-minute bookings would then compensate the airline for the unsold seats. (See, for example, the article by John R. Lott, Jr., and Russell D. Roberts in *Economic Inquiry,* January 1991.) But we know that cost differences are not the only reason for the price differential, or else the airlines would not have added the Saturday stayover requirement, which has nothing to do with their costs.

Catalog retailers—such as Victoria's Secret—have an easily available clue for determining who is willing to pay more: the customer's address. People who live in high-income zip codes are mailed catalogs with higher prices than people who live in lower-income areas. Some Internet retailers have even used software to track customers' past purchases to gauge whether each is a free spender or a careful shopper. Only the careful shoppers get the low prices.[4]

Prevention of Resale. To price discriminate, a firm must be able to prevent low-price customers from reselling its product to high-price customers. This can be a vexing problem for many would-be discriminators. For example, when airlines began price discriminating, a resale market developed: Business travelers could buy tickets at the last minute from intermediaries, who had booked in advance at the lower price and then advertised their tickets for sale. To counter this, the airlines imposed the additional requirement of a Saturday stayover for the lower price. By adding this restriction, the airlines were able to substantially reduce the reselling of low-price tickets to business travelers.

It is often easy to prevent resale of a *service* because of its personal nature. A hairstylist can charge different prices to different customers without fearing that one customer will sell her haircut to another. The same is true of the services provided by physicians, attorneys, and music teachers. Resale of *goods,* however, is much harder to prevent, since goods can be easily transferred from person to person without losing their usefulness.

Effects of Price Discrimination

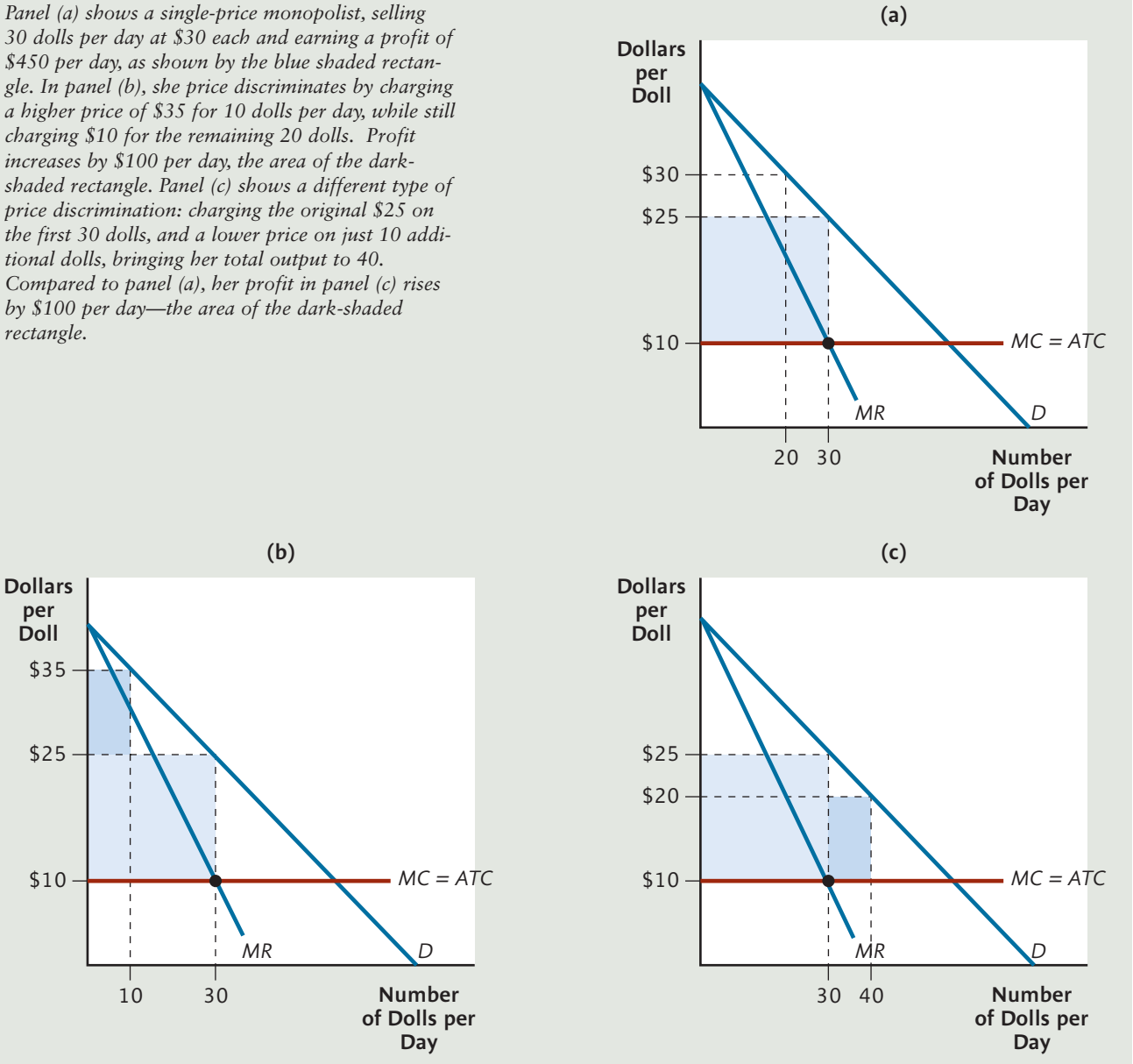
Price discrimination always benefits the owners of a firm: When the firm can charge different prices to different consumers, it can use this ability to increase its profit. But the effects on consumers can vary. To understand how price discrimination affects the firm and the consumers of its product, consider Larisa, a monopolist who produces and sells Elvis dolls at flea-markets. To keep our analysis as simple as possible, we'll assume that Larisa has no fixed costs, and that each doll costs \$10 to make no matter how many she produces. Thus, Larisa's cost per doll (*ATC*) is \$10 at every output level. Furthermore, because each additional doll costs \$10 to make, her marginal cost (*MC*) is also \$10 at any output level. This is why, in Figure 9(a), both the *MC* and *ATC* curves are represented by the same horizontal line at \$10.

Let's first suppose that Larisa is a single-price monopolist, charging a pre-announced price on every doll she sells. The figure shows the demand and marginal revenue curves she would face on a typical day. Using the $MR = MC$ rule, Larisa would earn maximum profit by selling 30 dolls per day, and charging \$25 per doll. Her profit per unit would be \$25 − \$10 = \$15, which is the vertical distance between the *ATC* curve and the demand curve at 30 units. Her total profit would be \$15 × 30 dolls = \$450 per day, which is equal to the area of the shaded rectangle.

[4] Anita Ramasastry, "Websites That Charge Different Customers Different Prices: Is Their 'Price Customization' Illegal? Should It Be?" *FindLaw Legal News and Commentary*, June 20, 2005.

FIGURE 9 Two Kinds of Price Discrimination

Panel (a) shows a single-price monopolist, selling 30 dolls per day at $30 each and earning a profit of $450 per day, as shown by the blue shaded rectangle. In panel (b), she price discriminates by charging a higher price of $35 for 10 dolls per day, while still charging $10 for the remaining 20 dolls. Profit increases by $100 per day, the area of the dark-shaded rectangle. Panel (c) shows a different type of price discrimination: charging the original $25 on the first 30 dolls, and a lower price on just 10 additional dolls, bringing her total output to 40. Compared to panel (a), her profit in panel (c) rises by $100 per day—the area of the dark-shaded rectangle.



Price Discrimination That Harms Consumers

Now suppose that Larisa discovers that on a typical day, some of her dolls are sold to particularly eager customers who show up first thing in the morning to buy them. Larisa figures she has found a way to identify those willing to pay more, so she charges the early-morning buyers $35 each and continues charging $25 to everyone else. The result, seen in panel (b), is that the first 10 dolls are now sold

to those willing to pay $35 or more for them, and the remaining 20 dolls continue to sell for $25.

What will happen to Larisa's profit? Because she's selling the same total number of dolls each day, her costs are unchanged. In effect, she has merely raised the price of the first 10 dolls by $10 each (from $25 to $35), increasing her revenue by $10 × 10 dolls = $100. Thus, Larisa's profit must rise by $100. This *increase* in profit is represented by the darker shaded rectangle in Figure 9(b). (Make sure you see why.) Larisa's total profit is now equal to the areas of *both* shaded rectangles together—the lightly-shaded one (her original profit of $450) and the darker one (her additional profit of $100). Thus, her total profit has risen to $550.

What about her customers? Those who pay a higher price than they otherwise would are harmed by her price discrimination (compared to the single-price outcome). Her other customers, who continue to pay $25, are unaffected. Thus, the increase in Larisa's profit is equal to the additional payments by the customer who pay more.

More generally,

Price discrimination can raise the price for some consumers above the price they would pay under a single-price policy. The additional profit for the firm comes at the expense of the consumers who pay more.

Price Discrimination That Benefits Consumers

Let's go back to the initial, single-price policy and suppose that Larisa had discovered a different way to price discriminate. Observing that those who come late in the day are rushed and tend not to buy any dolls at $25, she decides to lower the price to $20 during her last hour of business. Sure enough, she sells an additional 10 dolls that way. The result is shown in panel (c), where Larisa charges $25 for the first 30 dolls, and $20 for an additional 10 beyond those. Her total output is now 40 dolls per day.

But wait: By pushing her output all the way to 40 dolls per day, isn't Larisa violating the $MC = MR$ rule and decreasing her profit? Not really. The MR curve in the figure was drawn under the assumption that Larisa would have to lower her price on *all* dolls in order to sell more of them. But this is no longer the case. With price discrimination, the MR curve no longer tells us what will happen to Larisa's revenue when she increases her output.

In fact, we know that each doll she sells for $20 will now add a full $20 to her revenue. At the same time, each one adds only $10 to her cost. So she earns profit of $20 − $10 = $10 on each additional doll. She should sell as many of these additional dolls at $20 as people will buy. According to the demand curve, that is 10 dolls beyond the previous 30. Selling these 10 additional dolls increases her total profit by $10 × 10 = $100. This *increase* in profit is represented by the darker shaded rectangle in Figure 9(c). (Make sure you see why.) Larisa's total profit is once again equal to the areas of *both* shaded rectangles together—the lightly-shaded one (her original profit of $450) and the darker one (her additional profit of $100). Her total profit has risen to $550.

In this case, Larisa's customers are better off, too. The first 30 customers are unaffected—they continue to pay $25. But with price discrimination, an additional 10 people are able to buy dolls at a price they are willing to pay.

More generally,

Price discrimination can lower the price for some consumers below the price they would pay under a single-price policy. Those consumers benefit, while the firm earns higher profit.

Of course, it is possible for a firm to combine *both* types of price discrimination. That is, it could raise the price above what it would charge as a single-price monopoly for some consumers, and lower it for others. This would increase the firm's profit, while benefiting some consumers and harming others. In theory, a firm can go even further, as the next section discusses.

Perfect Price Discrimination

Suppose a firm could somehow find out the maximum price customers would be willing to pay for *each* unit of output it sells. Then it could increase its profits even further by practicing *perfect price discrimination:*

Perfect price discrimination
Charging each customer the most he or she would be willing to pay for each unit purchased.

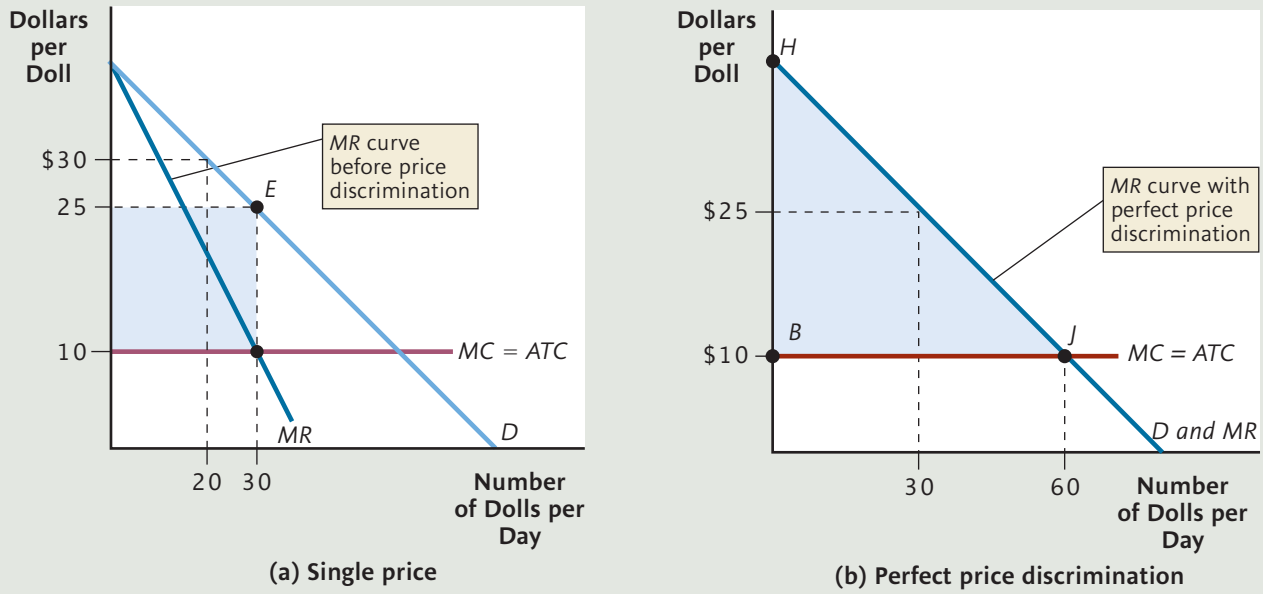
*Under **perfect price discrimination**, a firm charges each customer the most the customer would be willing to pay for each unit he or she buys.*

Perfect price discrimination is very difficult to practice in the real world, since it would require the firm to read its customers' minds. However, many real-world situations come rather close to perfect price discrimination. Used-car dealers routinely post a sticker price far higher than the price they think they can actually get. They then size up each customer to determine the discount needed to complete the sale. The dealer may look at the customer's clothes and the car the customer is currently driving, inquire about the customer's job, and observe how sophisticated the customer is about cars. The aim is to determine the maximum price he or she would be willing to pay. A similar sizing up takes place in flea markets, yard sales, and many other situations in which the final price is *negotiated* rather than fixed in advance.

Suppose that Larisa becomes especially good at sizing up her customers. She learns how to distinguish true Elvis fanatics (a white, sequined jumpsuit is a dead giveaway) from people who merely want the doll as a gag gift. Moreover, by observing the way people handle the doll and listening to their conversations with their companions, Larisa can discern the exact maximum price each customer would pay. In effect, she knows exactly where on the demand curve each customer would be located. With her new skills, she can increase her profit by becoming a *perfect price discriminator:* For each unit along the horizontal axis, she will charge the price indicated by the vertical height of the demand curve.

How many dolls should Larisa sell now? To answer this question, we need to find the new output level at which $MR = MC$. But once again, the MR curve in the figure is no longer valid: It was based on the assumption that Larisa had to lower the price on *all* units each time she wanted to sell another one. Now, as a perfect price discriminator, she needs to lower the price only on the *additional* unit she sells, and her revenue will rise by the price of that additional unit. For example, if she is currently selling 30 dolls and wants to sell 31, she would lower the price just on the additional doll by a tiny bit—say, to $24.50—and in that case, her revenue would rise by $24.50.

FIGURE 10 **Perfect Price Discrimination**



The single-price monopolist sells 30 dolls per day at $25 each. With a constant ATC *of $10, she earns a profit of $450 per day, as shown by the blue rectangle in panel (a). However, if she can charge each customer the maximum the customer is willing to pay, shown by the height of the demand curve, then her* MR *curve is the demand curve she faces. In panel (b), she would sell 60 dolls, where* MC = P *at point* J. *Her profit would increase to the area of triangle* HBJ.

For a perfect price discriminator, marginal revenue is equal to the price of the additional unit sold. Thus, the firm's MR *curve is the same as its demand curve.*

Now it is easy to see what Larisa should do: Since our requirement for profit maximization is that $MC = MR$, and for a perfect price discriminator, MR is the same as price (P), Larisa should produce where $MC = P$. In Figure 10(b), this occurs at point J, where the MC curve intersects the demand curve—at 60 units of output. At that point, the only way to increase sales would be to lower the price on an additional doll below $10, but since the marginal cost of a doll is always $10, we would have $P < MC$, and Larisa's profit would decline.

What is Larisa's profit-maximizing price? Think for a moment. Then see Footnote 5 for the answer.

What about Larisa's total profit? On each unit of output, she charges a price given by the demand curve and bears a cost of $10. Adding up the profit on *all* units gives us the area under the demand curve and above $10, or the area of triangle HBJ (not shaded).

Now we can determine who gains and who loses when Larisa transforms herself from a single-price monopolist to a perfect price discriminator. Larisa clearly gains: Her profit increases, from the rectangle in Figure 10(a) to the larger, shaded triangle

[5] Sorry, that's a trick question: There *is* no profit-maximizing price. As a perfect price discriminator, Larisa earns the highest profit by charging *different* prices to different customers.

in Figure 10(b). Consumers of the product are the clear losers: Since they all pay the most they would willingly pay, no one gets to buy a doll at a price he or she would regard as a "good deal."

> *A* perfect price discriminator *increases profit at the expense of consumers, charging each customer the most he or she would willingly pay for the product.*

How Firms Choose Multiple Prices

In our discussion of price discrimination, we showed that a firm can generally increase profit by charging different prices to different groups of customers. But we did not discuss how a firm would actually *choose* these prices. The marginal approach to profit can help us see how this is done.

Consider No-Choice Airlines, the only airline that flies direct between two small cities. No-Choice offers one flight each day, and serves two kinds of customers: business travelers and college students. Business travelers want to minimize their travel time, so are less sensitive to price. Students, by contrast, are more willing to travel by train or take a road trip and are generally more price-sensitive.

Suppose No-Choice has determined that it can separate these two markets by requiring a Saturday stay-over and advanced booking for a special "student price," while charging everyone else (for example, business travelers) the "normal price." It will charge a single (different) price in *each* market. But what should those prices be?

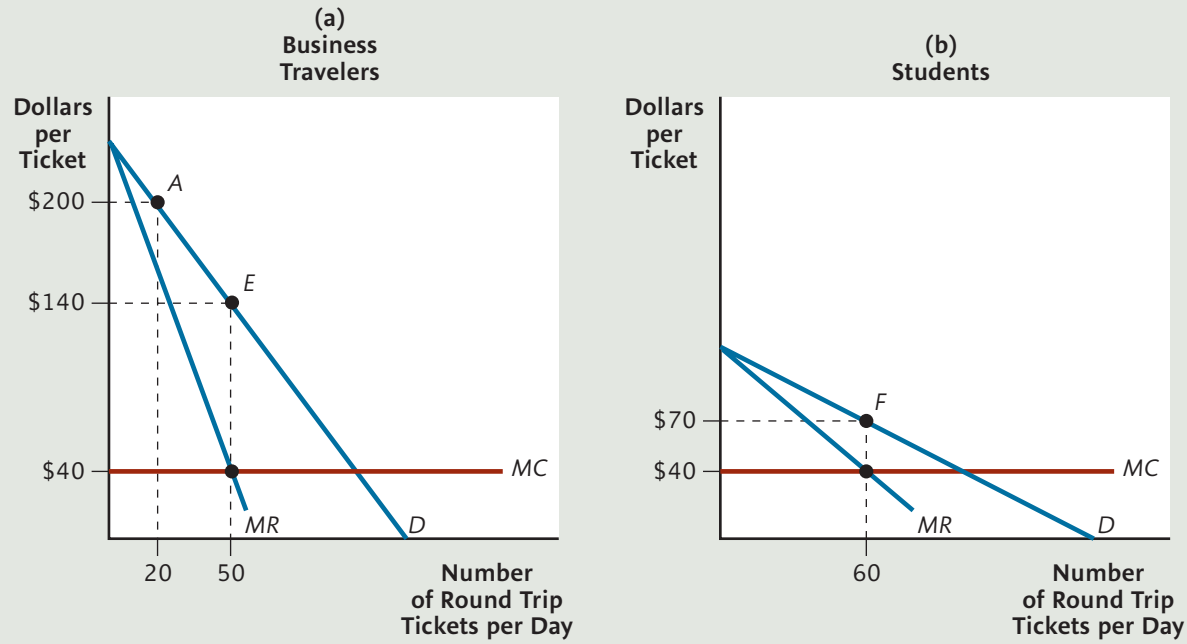
The guiding rule is:

> *A price discriminating firm facing separate market demand curves in different markets* (A, B, C, etc. . .) *should choose its prices and output levels so that marginal revenue in each market is equal to its marginal cost of production:* $MR_A = MR_B = MR_C = \ldots = MC$.

To see how this works, look at Figure 11, where we (once again) keep it simple by assuming that the marginal cost is constant at $40. (For No-Choice, marginal cost would include the additional fuel for an additional passenger, additional in-flight snacks, and perhaps the additional labor hours of ticket takers and baggage handlers). Panel (a) shows the demand and marginal revenue curves of the business travelers, while panel (b) shows the same curves for students.

Imagine, first, that No-Choice was charging $200 for business travelers, and selling 20 tickets, at point *A*. Could this be the best price to charge? The answer is no. If you draw a vertical line from point *A* down to the *MR* curve, you will find that marginal revenue at 20 tickets is about $160—substantially above the $40 marginal cost. So the airline could increase profit by lowering its price for business travelers and selling more tickets to them. In fact, the airline should continue lowering its price to business travelers until its marginal revenue decreases to $40, which occurs at 50 tickets per day. At that number of tickets, the *MR* curve intersects the *MC* curve, and the demand curve (at point *E*) tells us that price will be $140. Any further increase in ticket sales in the business market—say, to 60 tickets per day—would cause *MR* to drop below *MC*, and profit would decrease.

FIGURE 11 How a Price-Discriminating Monopoly sets Prices in Multiple Markets



No Choice Airlines is able to separate travelers into two different types: Business travelers in panel (a) and students in panel (b). In each market, it sells the profit-maximizing number of tickets at which marginal revenue is equal to its marginal cost, and charges the price on the demand curve for that number of tickets. Because the demand curves are different, so are the marginal revenue curves, so price and output will differ in each market. In panel (a), for business travelers, MR = MC *at 50 tickets, and No Choice charges $140. In panel (b), for students,* MR = MC *at 60 tickets, and No Choice charges $70.*

No-Choice should follow the same procedure in the student market in panel (b). There, the *MR* curve intersects the *MC* curve at 60 tickets, and the demand curve (at point *F*) gives us a price of $70. So No-Choice will charge $70 in the student market. (Convince yourself that charging more than $70 in the student market would leave profit opportunities unexploited.)

We cannot use this diagram to determine No-Choices total profit or loss, because No-Choice—like any airline—has other costs that are not part of its horizontal *MC* curve. (This includes the cost for its air terminal, salaries for pilot and copilot, and plane servicing costs.) But we do know that, assuming No-Choice operates at all, it maximizes profit by selling 50 tickets to business travelers for $140 each, and 60 tickets to students at $70 each. At any other prices (or output levels), profit would be smaller.

Price Discrimination in Everyday Life

Price discrimination is not limited to monopolies. It can be practiced by *any* firm that satisfies the three requirements discussed earlier. As a result, price discrimination is more prevalent than you might think.

Rebates on electronic goods are an example. If you've recently purchased a printer or computer, chances are your receipt included a coupon for a rebate from the store or the manufacturer—in effect, offering you a lower price. But to pay this

lower price, you must go through the time and trouble to read all the directions, cut the UPC code from the box, mail it in, wait several weeks or months for your check to arrive, and then deposit the check.

Many people complain about all this time and trouble, and wonder why the manufacturer or store doesn't just lower the sticker price. The answer, in large part, is price discrimination. By adding time, trouble, and delay for the discount, the store can separate those who are very price sensitive (they will go through the trouble) from those who are not (they will forget about the rebate). In effect, each group is charged a different price.

Discount coupons for the supermarket or drugstore work much the same way. You only get the discount if you happen to have the coupon with you at the store. Only the most price-sensitive customers will go through the trouble of clipping, saving, and organizing their coupons so that they have them when they need them.

When retailers put items "on sale" (for a reduced price) after a delay of weeks or months, it is in part an effort to price discriminate. Those who feel they must have the latest fashions, video games, or DVDs immediately after they arrive at the store, and have the income to buy them at higher prices, will make their purchases soon after the goods arrive. Weeks or months later, when the goods go on sale, everyone else pays a lower price.

Finally, colleges and universities are extensive practitioners of price discrimination. The college announces a high "sticker price" (official tuition), then gives discounts to students it believes would not attend without the discount. Colleges have various ways of determining willingness to pay. Financial disclosure forms for financial aid indicate a family's ability to bear a high price. And academic qualifications indicate how many other offers a student is likely to get and, therefore, how much of a discount the college must offer to tilt the odds in its favor.

Using the Theory

MONOPOLY PRICING AND PARALLEL TRADE IN PHARMACEUTICALS

© SGM SGM/AGE FOTOSTOCK/PHOTOLIBRARY

The pharmaceutical industry is always embroiled in controversy. On the one hand, we want the industry to provide us with new and better drugs, and to help cure or manage more diseases, with fewer side effects. The industry has done this remarkably well, helping hundreds of millions of people live longer and better lives. But we also want drugs to be inexpensive.

The problem is that new drugs are costly to develop. It takes more than 10 years to get a profitable drug to market, and only a fraction of those that make it most of the way down the road are deemed safe and effective enough to get government approval for sale. For this reason, the total research and development (R&D) cost for each *marketable* new drug is in the hundreds of millions of dollars.

To create incentives to pay for these high R&D costs, governments give internationally recognized patents to the companies that discover new drugs. In effect, the world's governments create a temporary monopoly for the firm, usually lasting for several years after the drug is first brought to market. During this time, a pharmaceutical

company—at least *in theory*—is free to exploit its market power and charge a high price. This allows it to cover its R&D costs and earn additional profit to compensate its owners for risking their investments. Once the patent expires, competitors rush in to produce generic versions of the drug, driving the price down toward marginal cost, thus ending the company's monopoly status. If a company cannot cover R&D costs and earn sufficient profit on the average drug while its patent remains in effect, it will not be able to stay in the business of discovering and selling new drugs.

Monopoly pricing during the patent period is how the industry works *in theory*. In actual practice, things are a bit different. Almost all countries honor pharmaceutical patents. But only one country—the United States—ends up funding the bulk of the world's R&D expenditures for new drugs. The reason: drugs are sold for significantly higher prices in the United States than in other countries. American buyers pay these higher prices either directly (if they don't have health insurance, and must pay the full price themselves) or indirectly (through higher premiums charged by insurance companies to cover their higher costs).

Many Americans—including members of Congress—have objected to this arrangement. They argue that buyers in other countries enjoy the benefits of newly discovered drugs without paying their fair share of the development costs. This might be acceptable for very poor countries that need life-saving medications and cannot afford to pay the same prices as residents of developed countries. But the list of countries where drug prices are substantially lower than in the United States includes Canada, Spain, France, Germany, Greece, Italy, and dozens of other high-income developed countries.

In the last decade, some U.S. consumers have taken matters into their own hands, purchasing medications at lower prices from other countries—especially Canada—even though this violates American law. This *parallel trade*—the resale of a product in another country without the manufacturer's consent—has increased every year. Some members of Congress want to encourage it further by overturning the 1987 law that prohibits U.S. residents from importing drugs once they are sold abroad. The pharmaceutical industry wants to keep the prohibition.

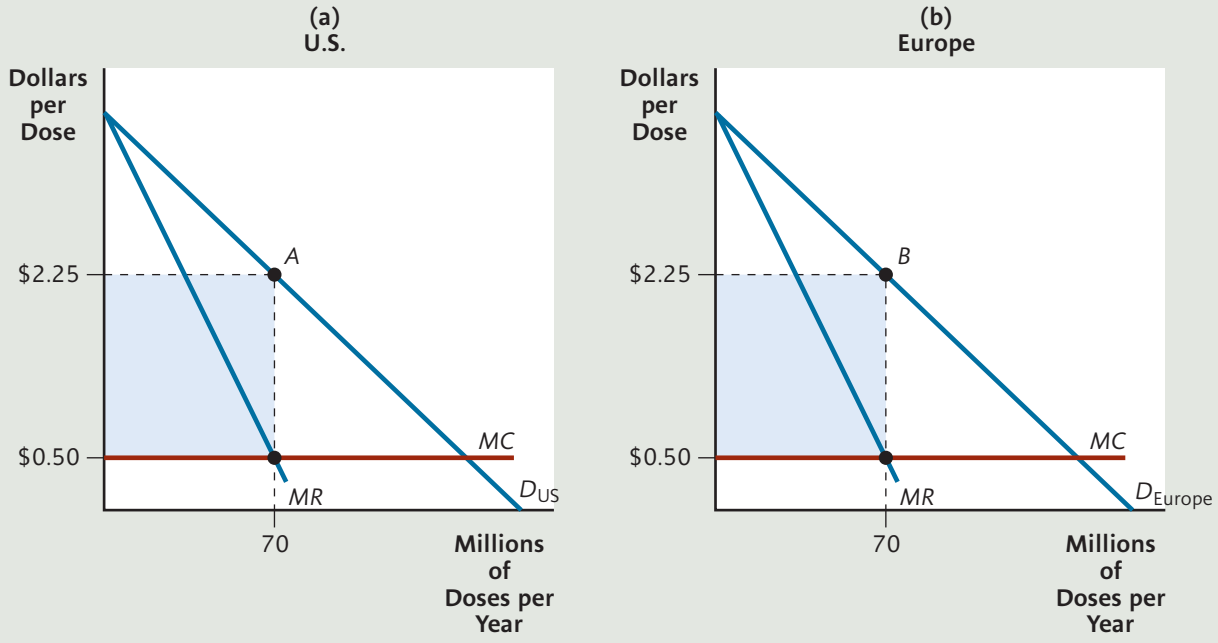
In this section, we explore two issues: (1) How has the lopsided pricing system come about? and (2) What are some possible solutions? Before we answer these questions, let's start our analysis by applying what you've learned about monopoly pricing to a hypothetical case.

Monopoly Pricing in Similar Markets

Let's suppose that a pharmaceutical company, Pfycor, Inc. holds the patent on a new drug that has been approved for sale by governments around the world. Figure 12 shows two markets for this drug: the U.S. and Europe. The R&D costs for the drug have already been paid, and we'll assume that the marginal cost and average cost of producing each dose is a constant $0.50.

Let's suppose, for now, that Pfycor can price discriminate and charge different prices in each market . . . if it wants to. But in the figure, we've assumed that the demand curves in the two markets are identical. By applying the pricing guideline to this example ($MR_{US} = MR_{EUROPE} = MC$), you'll see that Pfycor will maximize profits by first finding the output level at which the *MR* curve in each market crosses the *MC* curve: 70 million doses. Then, on the respective demand curves at Points *A* and *B*, it will find the price it can charge for that output level. As you can see, with identical demand curves, the profit-maximizing price is the same in each market: $2.25.

FIGURE 12 Monopoly Pricing in Two Identical Markets



The two panels illustrate the situation for a monopoly selling a drug in two markets with identical demand curves: The U.S. in panel (a), and Europe in panel (b). In each market, the monopoly sells the profit-maximizing output level at which marginal revenue is equal to marginal cost, and charges a price given by the demand curve at that output level. With identical demand and marginal revenue curves in the U.S. and Europe, the monopoly will sell the same output (70 million doses) and charge the same price ($2.25 per dose) in each market. Production cost is $0.50 per dose, so total revenue exceeds production cost by ($2.25 − $0.50) × 70 million = $122.5 million in each market (the area of the blue shaded rectangles).

Let's look more closely at the U.S. market on the left side. Notice that Pfycor charges $2.25 for each dose that costs $0.50 to produce. The difference of $1.75 per dose contributes to Pfycor's R&D expenses and profit. The U.S. market's contribution is therefore $1.75 × 70 million = $122.5 million per year, represented by the shaded rectangle.

When we do the same analysis for the European market on the right side, we see an identical shaded rectangle, representing another $122.5 million contribution for R&D from Europe. Combining these contributions, Pfycor earns $122.5 million + $122.5 million = $245 million beyond its production costs each year, for the duration of its patent.

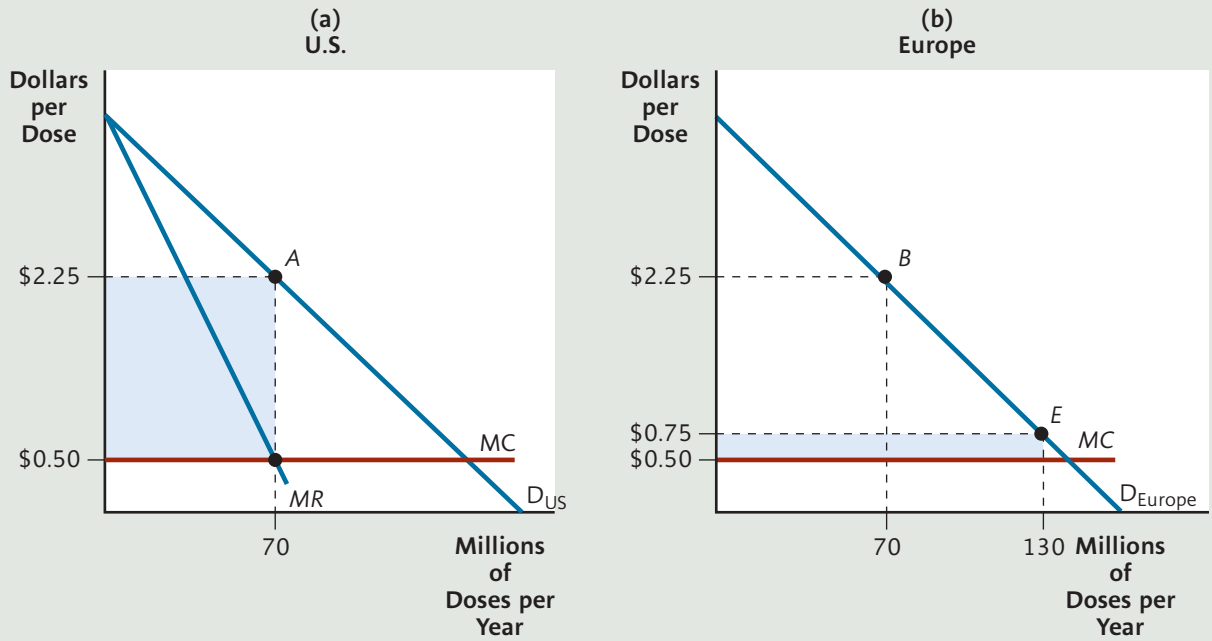
Clearly, the example in Figure 12—with Americans and Europeans paying the same price—does not explain the pricing controversies we discussed earlier. So let's bring in a fact we haven't yet addressed.

Asymmetrical Pricing from Government Bargaining

In almost all European countries (and most other developed countries as well), pharmaceutical companies must negotiate a price with the government before the drug can be sold there. These governments are in a strong bargaining position for two reasons: (1) they can refuse access to their markets entirely; and (2) they know that as long as the price is greater than $0.50 per dose (Pfycor's production cost), Pfycor will still want to sell the drug in their country.

But Pfycor has some bargaining power as well. If it refuses to sell at a low price, European patients will not be able to get the drug unless they buy it in the

FIGURE 13 Monopoly Pricing with Bargaining in One Market



As in the previous figure, a monopolist sells a drug in the U.S.in panel (a), and Europe in panel (b), and both markets have the same demand and marginal revenue curves. But in panel (b), European governments negotiate a lower price per dose of $0.75. Additional revenue from each dose sold in Europe is now $0.75, so that is the new (constant) marginal revenue in Europe. With marginal revenue greater than production cost, the monopoly will sell all that Europeans will buy at $0.75 per dose, which is 130 million doses. Production cost is still $0.50 per dose, so in Europe, total revenue exceeds production cost by ($0.75 - $0.50) × 130 million = $32.5 million. Meanwhile, in panel (a) for the U.S., nothing has changed, so price remains at $2.25. In the U.S., total revenue continues to exceed production cost by ($2.25 - $0.50) × 70 million = $122.5 million. Europe now contributes less to R&D expenses and profit ($32.5 million) than does the U.S. ($122.5 million).

U.S. market at the high monopoly price. When Europeans find out about the drug (with Pfycor's help, of course), they may clamor for their governments to approve it. But the governments can counter with the threat of a nasty public relations campaign, blaming Pfycor's extravagant pricing for any impasse. As you can see, bargaining over the price can be complicated.

Let's suppose that European governments drive a hard bargain, and the price ends up at $0.75 per dose. (In the real world, the price would be different in each European country, but we'll assume the price is the same for all of them.) This outcome is shown in the right panel of Figure 13, at point *E* on the demand curve. Europeans now buy 130 million doses per year—more than they did in Figure 12. (Note that the European *MR* curve does not appear, because it's no longer relevant. With a negotiated price, Pfycor does not have to lower its price to sell more; it simply gets $0.75 for each dose, no matter how much it sells. Its marginal revenue is, therefore, a constant $0.75.)

The shaded area in the figure shows how much the European market now contributes to Pfycor's R&D expenses and profit: $0.25 per dose (after deducting production costs of $0.50 from the price of $0.75). With 130 million doses, the total European contribution is now $32.5 million. This is substantially less than the $122.5 million Europe contributed before, when Pfycor charged the monopoly price there. Meanwhile, in the left panel, nothing has changed for the U.S. Americans still

pay the monopoly price of $2.25, and contribute the same $122.5 million each year. Thus, Pfycor's total recovery of funds for R&D expenses and profit is now $122.5 million + $32.5 million = $155 million.

Is this situation sustainable? Yes, if two conditions are satisfied: (1) $155 million per year for the life of the patent is enough, on average, to keep Pfycor developing new drugs; and (2) Pfycor can prevent parallel trade in the U.S.—that is, it can prevent low-priced drugs in the European market from being resold at similar low prices to Americans.

The Impact of Parallel Trade

Preventing parallel trade in the U.S. has become increasingly difficult. Through the Internet Americans can find out about the low European prices in seconds, and buy drugs to be mailed to their homes. In the case of Canada—whose government often negotiates even lower prices than Europe—Americans can just drive across the border to pick up their medications. The foreign pharmacies are not even breaking the law. It is U.S. law that prohibits the *import* of pharmaceuticals originally sold in other countries; these other countries do not prohibit their pharmacies from exporting drugs. And the U.S. government—its hands full with other problems—has largely ignored the problem of illegal pharmaceutical imports by individuals.

You can see that the widespread entry of low-price foreign drugs into the U.S. market creates a problem for Pfycor. If the price differential is large enough, everyone in the U.S. will buy their drugs in Europe. In that extreme case, the price in the U.S. market—and Pfycor's U.S. revenue—will fall to that in Europe, because Pfycor will not be able to sell at any higher price in the United States. This outcome—the convergence of prices to those in the low-price country when parallel trade is allowed—is sometimes called "importing another country's price controls." If this occurs, the U.S. market's contribution to R&D and profit (the revenue beyond production cost) would fall to that in Europe—$32.5 million. And the total contribution from both markets would fall to $65 million.

In general,

> *Parallel trade that causes convergence to a given lower price means less revenue for pharmaceutical companies to cover* R&D *expenses and profit, and reduces companies' willingness to bear high* R&D *costs and risks in the future.*

The U.S. drug-import ban, passed by Congress in 1987, was justified to the public as necessary to guarantee the purity and safety of drugs that Americans buy. But this cannot explain continuing to ban imports from countries with similar quality standards to the U.S., such as Canada or Germany. Rather, the pressure to continue the ban has come mostly from the pharmaceutical companies as a way to prevent parallel trade in the U.S. and preserve their high U.S. revenue in the face of heavy bargaining by foreign governments.[6]

[6] Some economists have questioned the view that parallel trade automatically reduces pharmaceutical revenue and funding for R&D. They argue that the negotiated price is itself affected by parallel trade. For example, without parallel trade, pharmaceutical companies won't bargain as hard: They know they can cover their expenses with the higher U.S. prices. Similarly, without parallel trade, foreign governments bargain harder: They know that high U.S. prices will maintain the pace of new drug discovery. Parallel trade changes these bargaining incentives, leading to a higher negotiated price and possibly greater funding for R&D than without parallel trade. See, for example, Gene M. Grossman and Edwin L. C. Lai, "Parallel Imports and Price Controls," *RAND Journal of Economics*, Vol. 39, No. 2, Summer 2008, pp. 378–402.

SUMMARY

A *monopoly firm* is the only seller of a good or service in a market, where the market is defined broadly enough to include any close substitutes. Monopoly arises because of some barrier to entry: economies of scale, legal barriers, or network externalities. As the only seller, the monopoly faces the market demand curve and must decide what price (or prices) to charge in order to maximize profit.

Like other firms, a single-price monopolist will produce where $MR = MC$ and set the maximum price consumers are willing to pay for that quantity. Monopoly profit ($P - ATC$ multiplied by the quantity produced) can persist in the long run because of barriers to entry. However, government regulation and rent-seeking activity can reduce monopoly profit.

All else equal, a monopoly charges a higher price and produces less output than a perfectly competitive market. When demand for a monopoly's product increases, it will raise prices and increase production. When a monopoly's marginal costs decrease, it will pass only part of the cost savings on to consumers.

Some monopolies can practice *price discrimination* by charging different prices to different customers. Doing so requires the ability to identify customers who are willing to pay more and to prevent low-price customers from reselling to high-price customers. Price discrimination always benefits the monopolist (otherwise, it would charge a single price), but it can either benefit or harm consumers, depending on whether they face higher or lower prices after the discrimination. With *perfect* price discrimination, every consumer is charged the highest price they are willing to pay.

When a price discriminating firm faces more than one market, it maximizes its profits by equating the marginal revenue in each market to its marginal cost of production. This leads to a higher price in markets where buyers are less price-sensitive, and lower prices in markets where buyers are more price sensitive.

PROBLEM SET

Answers to even-numbered Questions and Problems can be found on the text website at www.cengage.com/economics/hall.

1. In a certain large city, hot dog vendors are *perfectly competitive,* and face a market price of $1.00 per hot dog. Each hot dog vendor has the following total cost schedule:

Number of Hot Dogs per Day	Total Cost
0	$ 63
25	73
50	78
75	88
100	103
125	125
150	153
175	188
200	233

   a. Add a *marginal cost* column to the right of the total cost column. (Hint: Don't forget to divide by the *change* in quantity when calculating *MC*.)
   b. What is the profit-maximizing quantity of hot dogs for the typical vendor, and what profit (loss) will he earn (suffer)? Give your answer to the nearest 25 hot dogs.

   One day, Zeke, a typical vendor, figures out that if he were the only seller in town, he would no longer have to sell his hot dogs at the market price of $1.00. Instead, he'd face the following demand schedule:

Price per Hot Dog	Number of Hot Dogs per Day
> $6.00	0
6.00	25
5.00	50
4.00	75
3.25	100
2.75	125
2.25	150
1.75	175
1.25	200

   c. Add *total revenue* and *marginal revenue* columns to the table above. (Hint: Once again, don't forget to divide by the *change* in quantity when calculating MR.)
   d. As a monopolist with the cost schedule given in the first table, how many hot dogs would Zeke choose to sell each day? What price would he charge?
   e. A lobbyist has approached Zeke, proposing to form a new organization called "Citizens to Eliminate Chaos in Hot Dog Sales." The organization will lobby the city council to grant Zeke the only hot dog license in town, and it is guaranteed to succeed. The only problem is, the lobbyist is asking for a payment that amounts to $200 per business day as long as Zeke stays in business.

On purely economic grounds, should Zeke go for it? (Hint: If you're stumped, re-read the section on rent-seeking activity.)

2. Draw demand, *MR,* and *ATC* curves that show a monopoly that is just breaking even.

3. Below is demand and cost information for Warmfuzzy Press, which holds the copyright on the new best-seller, *Burping Your Inner Child*.

Q (No. of Copies)	*P* (per Book)	*ATC* (per Book)
100,000	\$100	\$20
200,000	\$ 80	\$15
300,000	\$ 60	\$16⅔
400,000	\$ 40	\$22½
500,000	\$ 20	\$31

   a. Determine what quantity of the book Warmfuzzy should print, and what price it should charge in order to maximize profit.
   b. What is Warmfuzzy's maximum profit?
   c. Prior to publication, the book's author renegotiates his contract with Warmfuzzy. He will receive a great big hug from the CEO, along with a one-time bonus of \$1,000,000, payable when the book is published. This payment was not part of Warmfuzzy's original cost calculations. How many copies should Warmfuzzy publish now? Explain your reasoning.

4. Regarding Figure 11(b) in the chapter, one of your fellow students says, "I think the airline is making a mistake by charging students \$70. It should drop the price further, so it could sell even *more* tickets than the 60 tickets in the figure. After all, as long as it charges even a little bit more than \$40, its price will still be above its marginal cost, so it will still make a profit on each ticket." What is wrong with your fellow students' argument?

5. A doctor in a rural area faces the following demand schedule:

Price per Office Visit	Number of Office Visits per Day
\$200	2
\$175	3
\$150	5
\$125	8
\$100	12
\$ 75	18
\$ 50	23
\$ 25	25

   The doctor's marginal cost of seeing patients is a constant \$50 per patient.
   a. If the doctor must charge all patients the same price, what price will she charge, and how many patients will she see each day?
   b. If the doctor can perfectly price discriminate, how many patients will she see each day?

6. You are thinking about tutoring students in economics, and your research has convinced you that you face the following demand curve for your services:

Price per Hour of Tutoring	Number of Students Tutored per Week
> \$50	0
\$40	1
\$35	2
\$27	3
\$26	4
\$20	5
\$15	6
< \$15	6

   Each student who hires you gets one hour of tutoring per week. You have decided that your time and effort is worth \$25 per hour and that you will not tutor anyone for less than that.
   a. Suppose you are wary that your students might talk to each other about the price you charge, so you decide to charge them all the same price. Determine (1) how many students you will tutor; (2) what price you will charge; and (3) your weekly earnings from tutoring.
   b. Now suppose you discover that your students don't know each other, and you decide to perfectly price discriminate. Once again, determine (1) how many students you will tutor; (2) what price you will charge; and (3) your weekly earnings from tutoring.

   Now suppose that your city requires all tutors to get a license, at a cost of \$1,300 per year (\$25 per week).
   c. Does it make sense for you to buy this license and be a tutor if you must charge each student the same price? Explain.
   d. Does it make sense for you to buy the license and be a tutor if you can perfectly price discriminate? Explain.

7. Draw demand, *MR*, *MC*, *AVC*, and *ATC* curves that show a monopolist operating at a loss that would cause it to *stay open* in the short run, but *exit* the industry in the long run. Then, show how a technological advance that lowers *only* the monopolist's *fixed costs* could cause a change in its long-run exit decision.

8. Answer the following:
   a. Complete the following table and use it to find this monopolist's short-run profit-maximizing

level of output. How much profit will this firm earn?

b. Redo the table to show what will happen to the short-run profit-maximizing level of output if the monopolist's marginal costs rise by $1 at each level of output. How much profit will the firm earn now?

c. Redo the original table to show what will happen to the short-run profit-maximizing level of output if the monopolist's marginal cost at each level of output is $0.40 less than before. How much profit would the firm earn in this case?

Output	Price	Total Revenue	Marginal Revenue	Total Cost	Marginal Cost	Profit
0	$5.60			$ 0.50		
1	$5.50			$ 3.50		
2	$5.40			$ 5.45		
3	$5.30			$ 6.45		
4	$5.20			$ 6.90		
5	$5.10			$ 8.90		
6	$5.05			$13.40		
7	$4.90			$20.40		

9. In the short run, a monopoly uses both fixed and variable inputs to produce its output. Draw a diagram illustrating a monopoly breaking even. Then alter your graph to show why, if the price of using a fixed input rises, there will be no change in the monopoly's short-run equilibrium price or quantity. (Hint: Which curves shift if the price of a fixed input rises?)

10. Suppose that Patty's Pool has the demand data given in Table 2 in the chapter. Further, suppose that Patty has just two types of costs: (1) rent of $25 per day and (2) towel service costs equal to 50 cents per swimmer. Over the short run, rent is a fixed cost (Patty has a lease she can't get out of), but towel service is a variable cost (it varies with the number of swimmers). Patty's marginal cost is therefore constant at 50 cents.

a. Under these cost conditions, what are Patty's short-run profit-maximizing output and price? What is her profit or loss per day?

b. Now suppose that, in addition to the costs just described, the town imposes a "swimming excise tax" on Patty's Pool equal to $2 per swimmer. What are Patty's new short-run profit-maximizing output and price? What is her new profit per day? (Hint: First decide whether or not the excise tax affects Patty's marginal cost.)

c. In addition to the costs just described (including the swimming excise tax of $2 per swimmer), suppose the town imposes a "fixed swimming tax" requiring Patty to pay $2 per day for operating her pool, regardless of the number of swimmers. What are Patty's new short-run profit-maximizing output and price? What is her new profit per day? (Hint: First decide whether or not this new fixed swimming tax affects Patty's marginal cost.)

d. Now suppose that costs are as in (c), except that the fixed swimming tax is $5 per day instead of $2 per day. What are Patty's new short-run profit-maximizing output and price? What is her new profit per day?

e. With the $5 per day fixed swimming tax, what should Patty do in the short run? If Patty's long-run costs are the same as her short-run costs, what should she do in the long run? (Hint: Think about shutdown for short run; exit for long run.)

f. Based on your answers to *b, c, d,* and *e,* assess the following statement: "When an excise (variable) tax is imposed on a monopoly, it will pass part, but not all, of the tax on to consumers in the form of a higher price. But a fixed tax has no effect on monopoly behavior over any time horizon." Are both of these sentences true? Explain briefly.

More Challenging

11. In Figure 13, the equation for each country's demand curve is $Q^D = 160 - 40\,P$, where Q^D represents millions of doses. Use this demand equation to answer the following questions:

a. Suppose the price negotiated for Europe is $1 per dose (rather than $0.75 as in the chapter), and there is no parallel trade. How many doses would Europe buy, and what would be its contribution to R&D expenditures and profit? What would be the total contribution from the U.S. and Europe?

b. Suppose that parallel trade occurs, and the price in both countries ends up converging to $1.00 per dose. How much would Pfycor collect in total to cover its R&D expenses and profit? How does this amount compare to what it collected without parallel trade, in part (a)?

c. Suppose as in part (b) that parallel trade occurs. But this time, because they know parallel trade will occur, the pharmaceutical companies bargain harder, and end up with a negotiated price of $1.50 per dose in Europe. Once again, the price converges to the negotiated price in both countries. How much would Pfycor collect in total to cover its R&D expenses and profit? How does this amount compare to the total collected in part (a), without parallel trade and the lower negotiated price? How does this amount

compare to the total collected in the chapter, when Pfycor charged the monopoly price in both countries?

12. Suppose that Patty's Pool has the demand data given in Table 2. Further, suppose that Patty has just two types of costs: (1) rent of $24 per day and (2) towel and other service costs equal to $5 per swimmer. Over the short run, rent is a fixed cost, but towel and other service costs are variable costs. Patty's marginal cost is therefore constant at $5.
    a. Under these cost conditions, and assuming first that Patty is a *single-price monopolist*, what are Patty's short-run profit-maximizing output and price? What is her profit or loss per day? (Hint: Be sure to check the shutdown rule if you determine that she is suffering a loss.) In the long run, if Patty's rent remains the same as in the short run, should she stay in this business?
    b. Now suppose that Patty figures out a way to price discriminate by dividing her swimmers into two groups: those willing to pay the price in part *a* and those who would not be willing to pay that price but *would* swim if the price were $5. What is Patty's short-run profit-maximizing output now? What is her profit or loss per day? In the long run, if Patty's rent remains the same as in the short run, should she stay in this business? (Hint: Be sure to recalculate the *TR* and *MR* numbers to answer this question. Also, if Patty is indifferent between two output levels, choose the higher one.)
    c. Let's say that Patty figures out a way to price discriminate by charging *three* different prices: a high price of $10 to those willing to pay that much to swim; a medium price equal to the price you found in part *a;* and a lower price of $5 to those who would not pay the price in part *a,* but would pay $5. What is Patty's short-run profit-maximizing output now? What is her profit or loss per day? In the long run, if Patty's rent remains the same as in the short run, should she stay in this business?
    d. Now suppose that Patty figures out a way to *perfectly* price discriminate, still facing the same demand curve given in Table 2 and Figure 2. What is Patty's short-run profit-maximizing output now? What is her profit or loss per day? In the long run, if Patty's rent remains the same as in the short run, should she stay in this business?

13. Suppose a single-price monopoly's demand curve is given by $P = 20 - 4Q$, where P is price and Q is quantity demanded. Marginal revenue is $MR = 20 - 8Q$. Marginal cost is $MC = Q^2$. How much should this firm produce in order to maximize profit?

CHAPTER 11

Monopolistic Competition and Oligopoly

On any given day, you are exposed to hundreds of advertisements. On the way to class, you might see billboards suggesting that you stay at the Holiday Inn, eat at Burger King, or buy the latest, lightest, thinnest iPod. You will likely spend more time watching television advertisements for breakfast cereals than you will spend eating them. And as you search for information on the Internet, ads for video cameras, credit cards, and vitamins flash before your eyes. In all of these cases, firms are trying to convince you that their product is better or cheaper than those of its competitors.

Yet so far in this book, not much has been said about this kind of competitive advertising. That's because in the market structures we've studied so far, there is no reason for it. A perfectly competitive firm would not advertise quality or price, because each firm produces the same product as every other, and each sells at the same market price. Monopolies often *do* advertise, to make their product more appealing and to shift the market demand curve rightward. But a *pure* monopoly does not advertise that its product is *better* or *cheaper* than the alternatives, because there *are* no reasonable alternatives.

Where, then, is all of this advertising about price and quality coming from? To answer this question, we must look beyond the market structures we've studied so far and consider firms that are neither perfect competitors nor monopolists. That is what we will do in this chapter. While advertising is one interesting feature we will explore, there are many others as well.

The Concept of Imperfect Competition

When thinking of market structure, perfect competition and monopoly can be viewed as the two extremes. In perfect competition, there are so many firms producing the same product that each takes the market price as given. In monopoly, there is only one firm in the market, producing a product without close substitutes. It can set its price without worrying about other firms that are selling a similar product.

Most goods and services, however, are sold in markets that are neither perfectly competitive nor monopolies. Instead, they lie somewhere *between* these two extremes. In these markets, there is more than one firm, but each firm has some market power—some ability to set price.

Consider, for example, the market for wireless phone service in the United States. It is certainly not a monopoly, because there is more than one firm. But neither does it resemble perfect competition. For one thing, there are only four large firms (Verizon Wireless, AT&T, Sprint, and T-Mobile) that provide service to more than 90 percent of those with cell phones. Further, their service differs in important ways

that matter to consumers: coverage, reliability, customer service, available phone models, and more. Thus, in terms of the number of firms and differences in the product, the market for wireless phone service falls somewhere between the extremes of monopoly and perfect competition.

Or consider restaurants. Even a modest-size city such as Cincinnati has thousands of different restaurants. This is certainly a large number of competitors. But they are not *perfect* competitors, because each one sells a product that is differentiated in important ways—in the type of food served, the recipes used, the atmosphere, the location, and even the friendliness of the staff. The markets for wireless phone service and restaurant meals in most cities are examples of **imperfect competition.** Imperfectly competitive markets have more than one firm (so they are not monopolies), but they violate one or more of the requirements of perfect competition.[1]

Imperfect competition A market structure in which there is more than one firm but one or more of the requirements of perfect competition is violated.

In this chapter, we study two types of imperfectly competitive markets: *monopolistic competition* and *oligopoly.*

Monopolistic Competition

Suppose you live in a midsize or large city, and you're having a night out with friends when you notice the gas tank is empty. You pass several stations, looking for one with a decent convenience store, because one of your friends needs to get some hair gel. After filling up the tank, a lively discussion ensues because everyone wants to go to a different pizza place. Finally, after some cajoling and compromising, you head across town to your group's choice. Over dinner, you all decide to see a movie. But that leads to another discussion: which multiplex to go to. One has the advantage of being closest, but the popcorn is stale. A second has great popcorn but it's impossible to park. And a third has great parking and great popcorn, but it's a 20-minute drive.

Although most people would have no reason to notice, all of your purchases that night would have something in common. The gas station, the pizza place, and the movie theaters all sell their products under a market structure called **monopolistic competition.**

Monopolistic competition A market structure in which there are many firms selling products that are differentiated, and in which there is easy entry and exit.

As the name suggests, monopolistic competition is a hybrid of perfect competition and monopoly, sharing some of the features of each. Specifically,

> *a monopolistically competitive market has three fundamental characteristics:*
>
> 1. *many buyers and sellers;*
> 2. *sellers offer a differentiated product; and*
> 3. *sellers can easily enter or exit the market.*

If you compare this list of characteristics with the list for perfect competition (in Chapter 9), you'll notice that the first and last items are shared by both market structures. The second item is new, but it leads to a feature shared with monopoly. Let's examine each of these characteristics in turn.

Many Buyers and Sellers

In *perfect* competition, the existence of many buyers and sellers played an important role: ensuring that no individual buyer or seller could influence the market price. In monopolistic competition, the "many buyers and sellers" assumption plays the same

[1] Imperfect competition is sometimes defined as *any* market structure other than perfect competition, which would include monopoly as well.

role on the buying side: an individual *buyer* has no influence on the price he pays. But an individual seller, in spite of having many competitors, has market power and acts as a *price setter*.

Our assumption of many sellers, however, has another purpose: it rules out strategic interaction among firms in the market. That is, when a firm under monopolistic competition makes a decision (about price, advertising, product guarantees, etc.), it does not take into account how it will affect other firms, or how they might respond. There are simply too many other firms, each supplying such a small part of the market, that no one of them can have much impact on the others.

Restaurants in most cities satisfy this requirement. With so many other restaurants, when one decides whether to offer an early-bird special or advertise in the local paper or put flyers under windshields, it usually doesn't worry how the other restaurants in the city will react.[2]

Sellers Offer a Differentiated Product

In perfect competition, sellers offer a standardized product. In *monopolistic competition,* by contrast, each seller produces a somewhat different product from the others. No two coffeehouses, photocopy shops, or food markets are exactly the same. For this reason, a monopolistic competitor can raise its price (up to a point) and lose only *some* of its customers. The others will stay with the firm because they like its product, even when it charges somewhat more than its competitors.

Thus, a monopolistic competitor faces a *downward-sloping demand curve,* so it has market power. In this sense, it is more like a monopolist than a perfect competitor:

> *Because it produces a differentiated product, a monopolistic competitor faces a downward-sloping demand curve: It can sell more by charging less, or raise its price without losing all of its customers.*

What makes a product differentiated? Sometimes, it is the *quality* of the product. By many objective standards—longevity, performance, frequency of repair—a Mercedes is a better car than a Hyundai. Similarly, based on room size and service, the Hilton has better hotel rooms than Motel 6. But the difference can also be a matter of taste. Objectively speaking, Colgate toothpaste may be neither better nor worse than Crest. But each has its own flavor and texture, and each appeals to different people.

Another type of differentiation arises from differences in *location*. Two bookstores may be identical in every respect—range of selection, atmosphere, service—but you will often prefer the one closer to your home or office.

Ultimately, though, product differentiation is subjective: A product is different whenever people *think* that it is, whether their perception is accurate or not. You may know, for example, that all bottles of bleach have identical ingredients—5.25 percent sodium hypochlorite and 94.75 percent water. But if *some* buyers think that Clorox bleach is different and would pay a bit more for it, then Clorox bleach is a differentiated product.

Because a monopolistic competitor faces a downward-sloping demand curve, the firm *chooses* its price. Like a monopoly, it is a *price setter*.

[2] Monopolistic competitors do imitate the successful practices of others in the market, as you'll see in a few pages. But a monopolistic competitor does not take into account the *potential* for imitation when making a decision. In the second half of this chapter, when we study oligopoly, we'll examine what happens when firms *do* take into account the potential reactions of their rivals.

Easy Entry and Exit

This feature is shared by monopolistic competition and perfect competition, and—as you'll see—it plays the same role in both: ensuring that firms earn zero economic profit in the long run. Remember that "easy entry" does *not* mean that entry is effortless or inexpensive. Rather, it means that you can open up, say, a pizza place if you're willing to bear the same costs that existing pizza places must bear. There are no significant *barriers* to entry that keep out newcomers—no law, for example, that new pizza places must pay higher annual license fees than established ones.

In monopolistic competition, however, our assumption about easy entry extends to business practices as well: Any firm can copy the successful practices of other firms. If one movie theater finds that offering lower prices for Wednesday-afternoon showings generates economic profit, any other movie theater can do the same. If Best Buy's multi-year replacement plan for electronic goods brings it profit, then other electronic goods retailers can choose to offer the same plan. Although it may take time, success will eventually lead to imitation. You'll see that this extended view of entry—specific to monopolistic competition—plays a role in ensuring zero economic profit in the long run.

Monopolistic Competition in the Short Run

The individual monopolistic competitor behaves very much like a monopoly. Its goal is to maximize profit by producing where $MR = MC$. The result may be economic profit or loss in the short run.

The key difference is this: While a monopoly is the *only* seller in its market, a monopolistic competitor is one of many sellers. When a *monopoly* raises its price, its customers must pay up or buy less of the product. When a *monopolistic competitor* raises its price, its customers have one additional option: They can buy a similar (though not identical) good from some other firm. Thus, all else equal, the demand curve facing a firm will be more elastic under monopolistic competition than under monopoly.

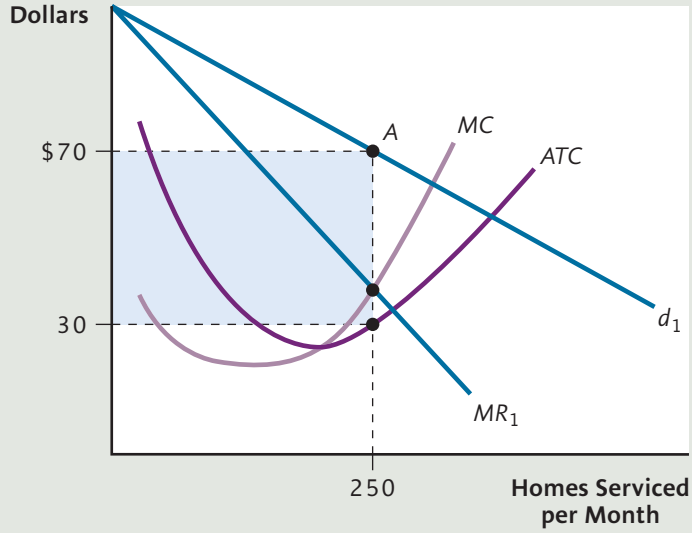
Figure 1 illustrates the situation of a monopolistic competitor, Kafka Exterminators. The figure shows the demand curve, d_1, that the firm faces, as well as the marginal revenue, marginal cost, and average total cost curves. As a monopolistic competitor, Kafka Exterminators competes with many other extermination services in its local area. Thus, if it raises its price, it will lose some of its customers to the competition. If Kafka had a *monopoly* on the local extermination business, we would expect less price elasticity than in the figure, because customers would have to buy from Kafka or get rid of their bugs on their own.

Like any other firm, Kafka Exterminators will produce where $MR = MC$. As you can see in Figure 1, when Kafka faces demand curve d_1 and the associated marginal revenue curve MR_1, its profit-maximizing output level is 250 homes serviced per month, and its profit-maximizing price is \$70 per home. In the short run, the firm may earn an economic profit or an economic loss, or it may break even. In the figure, Kafka is earning an economic profit: Profit per unit is $P - ATC = \$70 - \$30 = \$40$, and total monthly profit—the area of the blue rectangle—is $\$40 \times 250 = \$10{,}000$.

Monopolistic Competition in the Long Run

If Kafka Exterminators were a monopoly, Figure 1 might be the end of our story. The firm could continue to earn economic profit forever, since barriers to entry would keep out any potential competitors.

FIGURE 1 A Monopolistically Competitive Firm in the Short Run



Like any other firm, a monopolistic competitor maximizes profit by producing the level of output where its MR *and* MC *curves intersect. Kafka exterminators maximizes profit by servicing 250 homes per month. The profit-maximizing price ($70) is found on the demand curve at an output level of 250 (point* A*). Profit per unit of $40 is the difference between the price ($70) and average total cost ($30) at output of 250. Total profit is profit per unit times output ($40 × 250 = $10,000), equal to the area of the shaded rectangle.*

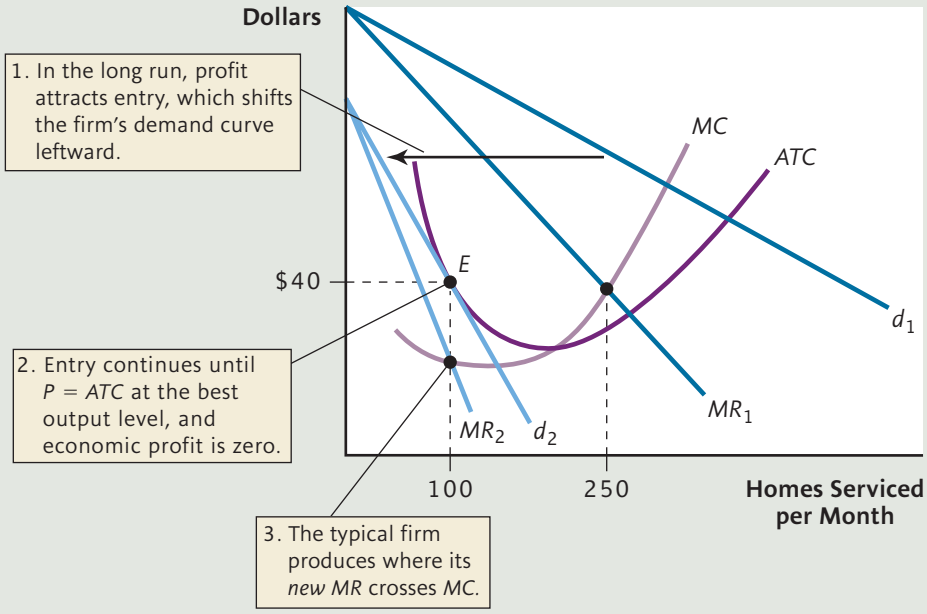
But under monopolistic competition—in which there are no barriers to entry and exit—the firm will not enjoy its profit for long. New sellers will enter the market, attracted by the profits that can be earned there. Kafka will lose some of its customers to the new entrants. At any given price, Kafka will find itself servicing fewer homes than before, so the demand curve it faces will shift leftward. Entry will continue to occur, and the demand curve will continue to shift leftward, until Kafka and other firms are earning zero economic profit.

This process of adjustment is shown in Figure 2. The demand curve shifts leftward (from d_1 to d_2). The marginal revenue curve shifts left as well (from MR_1 to MR_2). Kafka's new profit-maximizing output level, 100, is found at the intersection point between its marginal cost curve and its *new* marginal revenue curve MR_2. Kafka's new price—found on its demand curve d_2 at 100 units—is $40. Finally, since *ATC* is also $40 at that output level, Kafka is earning zero economic profit—the best it can do in the long run.[3] In long-run equilibrium, the profit-maximizing price, $40, will always equal the average total cost of production.

We can also reverse these steps. If the typical firm is suffering an economic loss (draw this diagram on your own), *exit* will occur. With fewer competitors, those firms that remain in the market will gain customers, so their demand curves will shift *rightward*. Exit will cease only when the typical firm is earning zero economic profit, where the demand curve just touches the *ATC* curve point like *E* in

[3] In the figure, we've assumed that as long-run adjustment takes place, the *ATC* and *MC* curves remain as drawn. That is, we assume that Kafka's optimal plant-size and the prices it pays for its inputs remain unchanged in the long run. If these assumptions don't hold, the *ATC* and *MC* curves will shift during long-run adjustment, but our conclusion remains the same: Entry will occur until the demand curve shifts leftward by just enough to touch the *new ATC* curve, so that $P = ATC$ at the new profit maximizing output level.

FIGURE 2 A Monopolistically Competitive Firm in the Long Run



dangerous curves

Contradictory curves A common error when drawing a monopolistic competitor in long-run equilibrium is to have the demand curve *cross* the *ATC* curve, rather than be *tangent* to the *ATC* curve as in Figure 2. An example of this mistake is drawn here, with a (supposed) long-run equilibrium output of Q_1. It is true that at Q_1, $P = ATC$, so economic profit would indeed be zero. But Q_1 is not the profit maximizing output level. Why? Notice that there are *other* output levels (e.g., Q_2) where the demand curve lies *above* the *ATC* curve. Because $P > ATC$ at those output levels, the firm could earn an economic profit by producing them. Because Q_1 is *not* the profit-maximizing output level (since other output levels have positive profit), then Q_1 *cannot* be the output level at which the *MR* and *MC* curves intersect. One or more of these curves must be drawn incorrectly.

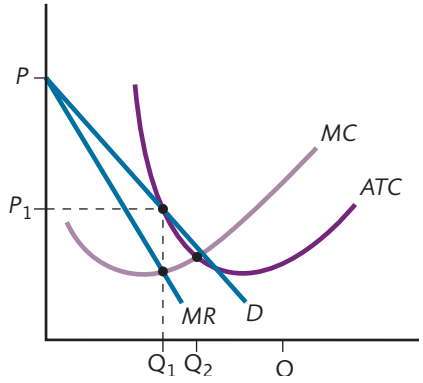


Figure 2. Thus, the typical firm will earn zero economic profit in the long run, whether we start from a position of economic profit or economic loss:

> *Under monopolistic competition, firms can earn positive or negative economic profit in the short run. But in the long run, free entry and exit ensure that each firm earns zero economic profit, just as under perfect competition.*

Is this prediction of our model realistic? Indeed it is: In the real world, monopolistic competitors often earn economic profit or loss in the short run. But, given enough time, profits attract new entrants while losses result in an "industry shakeout" as firms exit. In the long run, restaurants, retail stores, hair salons, and other monopolistically competitive firms earn zero economic profit for their owners. That is, there is just enough accounting profit to cover the implicit costs of doing business—just enough to keep the owners from shifting their time and money to some alternative enterprise.

Excess Capacity Under Monopolistic Competition

Look again at Figure 2, which shows Kafka's long-run equilibrium, at point *E*, after entry has eliminated its profits.

At that point, the *ATC* curve has the same slope as the demand curve, a *negative* slope. Thus, in the long run, a monopolistic competitor always produces on the *downward-sloping* portion of its *ATC* curve and therefore *never produces at minimum average cost*. Indeed, its output level is always *too small* to minimize cost per unit. The firm operates with *excess capacity*. (The output level at which cost per unit is minimized is often called capacity output.) In Figure 2, Kafka Exterminators *would* reach minimum cost per unit by servicing about 200 homes per month (the firm's capacity output), but in the long run, it will service only 100 homes per month.

In the long run, a monopolistic competitor will operate with excess capacity—that is, it will not sell enough output to achieve minimum cost per unit.

To see why a monopolistic competitor *cannot* minimize average cost in the long run, imagine that Kafka Exterminators wanted to do so, by servicing 200 homes per month. With its current demand curve, it would suffer a loss, since $P < ATC$ at that output level. It would quickly return to its profit-maximizing output of 100 homes, where at least it breaks even.

Excess capacity suggests that monopolistic competition is costly to consumers, and indeed it is. Recall that under perfect competition, $P =$ minimum *ATC* in long-run equilibrium. (Look back at Figure 9 in Chapter 9.) But under monopolistic competition, $P >$ minimum *ATC* in the long run. Thus, if the *ATC* curves were the same, price would always be greater under monopolistic competition.

This reasoning may tempt you to leap to a conclusion: Consumers are better off under perfect competition. But don't leap so fast. Remember that in order to get the beneficial results of perfect competition, all firms must produce identical output. It is precisely because monopolistic competitors produce *differentiated* output—and therefore have downward-sloping demand curves—that $P >$ minimum *ATC* in the long run.

And consumers usually *benefit* from product differentiation. (If you don't think so, imagine how you would feel if every restaurant in your town served an identical menu, or if everyone had to wear the same type of clothing, or if every rock group in the country performed the same tunes in exactly the same way.) Seen in this light, we can regard the higher costs and prices under monopolistic competition as the price we pay for product variety. Some may argue that there is too much variety in a market economy—how many different brands of toothpaste do we really need?—but few would want to transform all monopolistically competitive industries into perfectly competitive ones.

Nonprice Competition

If a monopolistic competitor wants to increase its output, one way is to cut its price. That is, it can move *along* its demand curve. But a price cut is not the only way to increase output. Since the firm produces a differentiated product, it can also sell more by convincing people that its own output is better than that of competing firms. Such efforts, if successful, will *shift* the firm's demand curve rightward.

Any action a firm takes to shift the demand curve for its output to the right is called ***nonprice competition.***

Nonprice competition Any action a firm takes to shift its demand curve rightward.

Better service, product guarantees, free home delivery, more attractive packaging, better locations, as well as advertising to inform customers about these things, are all examples of nonprice competition.

This type of competition is another reason why monopolistic competitors earn zero economic profit in the long run. If an innovative firm discovers a way to shift its demand curve rightward—say, by offering better service or more clever advertising—then in the *short run,* it may be able to earn a profit.

But not for long. Remember that in monopolistic competition, the "free entry" assumption includes the ability of new entrants, as well as existing firms, to replicate the successful business practices of others. If product guarantees are enabling some firms to earn economic profit, then *all* firms will offer product guarantees. If advertising is doing the trick, then *all* firms will start ad campaigns. In the long run, imitation by others will reverse any advantage that the initiators hoped to achieve, and will begin shifting their demand curves back again. At the same time, the costs of the nonprice competition shifts up each firm's *ATC* curve. After all, firms have to *pay* for advertising, product guarantees, or better staff training.

As you can see, even if nonprice competition leads to profits for the early adopters in the short run, we can identify *two* forces that shrink profit back to zero in the long run: (1) imitation by others reverses the initial rightward shift in demand; and (2) the costs of nonprice competition shift the *ATC* curve upward. In the end, each firm will once again earn zero economic profit, with its demand curve tangent to the new, higher *ATC* curve. We will take a closer look at one form of nonprice competition, advertising, in the Using the Theory section at the end of the chapter.

Oligopoly

A monopolistic competitor enjoys a certain amount of independence. There are so many *other* firms selling in the market—each one such a small fish in such a large pond—that each of them can make decisions without worrying about how the others will react. For example, if a single pharmacy in a large city cuts its prices or begins advertising, it can safely assume that any other pharmacy that could benefit from price cutting or advertising has already done so, or will shortly do so, *regardless of its own actions*. Thus, there is no reason for the first pharmacy to take the reactions of other pharmacies into account when making its own decisions.

But in some markets, most of the output is sold by just a few firms. These markets are not monopolies (there is more than one seller), but they are not monopolistically competitive either. There are so few firms that the actions taken by any one will *very much* affect the others and will likely generate a direct response. Before the management team makes a decision, it must reason as follows: "If we take action *A,* our competitors will do *B,* and then we would do *C,* and they would respond with *D*. . . ," and so on. This kind of strategic interaction among firms is the hallmark of the market structure we call *oligopoly:*

Oligopoly A market structure with a small number of strategically interacting firms.

*An **oligopoly** is a market dominated by a small number of strategically interacting firms.*

There are many different types of oligopolies. The products may be more or less identical among firms, such as copper wire, or differentiated, such as laptop computers. An oligopoly market may be international, as in the market for automobile

tires; mostly national, as in the U.S. market for breakfast cereals; or local, as in the market for some daily newspapers. One firm may be significantly larger than any of its rivals (such as Nike in the U.S. market for athletic shoes). Or there may be two or more large firms of roughly similar size (like Boeing and Airbus in the global market for large passenger aircraft). You can see that oligopoly markets can have different characteristics, but in all cases, *a small number of strategically interacting firms produce the dominant share of output in the market.*

Oligopoly in the Real World

While defining an oligopoly in theory is straightforward, *applying* the definition to real-world markets is sometimes difficult. This is not just a matter of semantics. The extent to which a market follows the oligopoly model—with market dominance by a few firms—is at the heart of public policy toward market structure, a subject we'll examine more closely in Chapter 15.

Whether we view a market as an oligopoly depends on how the market is defined. With a narrow-enough definition, we can find oligopoly everywhere. For example, in a large city there will be thousands of restaurants, so we would properly consider the market for "restaurant meals" in that city to be monopolistically competitive. But if we define the market as "Thai restaurants within a half-mile from the civic center," there may be only two or three such firms, and voilá—we have an oligopoly!

Similarly, with a broad-enough definition, we can make any oligopoly disappear. In the U.S. market for breakfast cereal, just four firms control about 90 percent of the market—an oligopoly. But if we widen our market definition to include "all food," then these four firms compete with thousands of other firms. The market no longer satisfies the definition of oligopoly.

How, then, should we define a market in trying to identify oligopoly? The approach taken by economists (including those who work in government agencies that have the power to approve or prohibit mergers between large firms) is to define the market *just* broadly enough so that it includes all "reasonably close" substitutes. In many cases, common sense can guide us in applying this principle. Thus, we refer to the market for "breakfast cereals," because one breakfast cereal is a close substitute for another. The market for "food" would be too broad because it would include too many products that are not close substitutes. And the market for "cornflakes" would be too narrow, because it would leave out other types of breakfast cereal, which are close substitutes.

But in some cases, common sense isn't definitive. Consider what happened in 2007, when Whole Foods—the largest natural food chain in the U.S.—announced it would acquire Wild Oats, a smaller natural foods chain. The Federal Trade Commission (FTC) tried to block the takeover, claiming that it would reduce competition in localities where the two companies owned the only natural food stores in town. The FTC was defining the relevant market in each town as "the market for natural foods sold in natural food stores." Whole Foods argued that *all* natural food sellers should be considered part of the market, including regular supermarkets and Wal-Mart branches with natural food sections. Using that broader measure, Whole Foods and Wild Oats together controlled a much smaller fraction of total sales in each town.

In 2009, the two sides reached a settlement, with Whole Foods agreeing to give up about half of the Wild Oats stores it had wanted to acquire. In effect, the FTC's narrower definition of the market prevailed. But one could make a reasonable argument for the broader definition advocated by Whole Foods.

How Oligopolies Arise

If a market has just a few sellers, we should naturally ask: Why aren't there more? Especially because in some oligopoly markets firms earn economic profit year after year. Why doesn't such profit attract entry, as it does in perfect competition and monopolistic competition? What *barriers to entry* keep out new competitors, leaving just a few firms with the market all to themselves?

Economies of Scale

One familiar barrier to entry is economies of scale. In Chapter 7, you learned that when a firm has economies of scale over a wide range of output, a large firm will have lower cost per unit than would a small firm. This can create a natural monopoly if a firm's minimum efficient scale (MES) occurs when it produces for the entire market. Or it can create a **natural oligopoly** if the MES occurs when a firm produces for a large fraction of the market. Since small firms can't compete, only a few large firms survive, and the market becomes an oligopoly. Airlines, college textbook publishers, and passenger jet manufacturers are all examples of oligopolies in which economies of scale play a large role.

Natural oligopoly A market that tends naturally toward oligopoly because the minimum efficient scale of the typical firm is a large fraction of the market.

Reputation as a Barrier

A new entrant may suffer just from being new. Established oligopolists are likely to have favorable reputations. In many oligopolies—like the markets for soft drinks and breakfast cereals—heavy advertising expenditure has also helped to build and maintain brand loyalty. A new entrant might be able to catch up to those already in the industry, but this may require a substantial period of high advertising costs and low revenues. This puts new entrants at a disadvantage compared to the firms already in the industry.

Strategic Barriers

Oligopoly firms often pursue strategies designed to keep out potential competitors. They can maintain excess production capacity as a signal to a potential entrant that, with little advance notice, they could easily saturate the market and leave the new entrant with little or no revenue. They can make special deals with distributors to receive the best shelf space in retail stores or make long-term arrangements with customers to ensure that their products are not displaced quickly by those of a new entrant.

Legal Barriers

Patents and copyrights, which can be responsible for monopoly, can also create oligopolies. For example, only four medications have received government approval for treatment of mild to moderate Alzheimer's disease, and all four are still protected by patents. Until these patents expire, or several new drugs are developed, this market will continue to be an oligopoly in which just four large pharmaceutical companies are the sellers.

Like monopolies, oligopolies are not shy about lobbying the government to preserve their market domination. One of the easiest targets is foreign competition. U.S. steel companies are relentless in their efforts to limit the amount of foreign—especially Japanese—steel sold in the U.S. market. In the past, they have succeeded in getting special taxes on imported steel and financial penalties imposed upon successful foreign steel companies. Other U.S. industries, including automobiles, textiles, and lumber, have had similar successes.

In local markets, zoning regulations may prohibit the building of a new supermarket, movie theater, or auto repair shop, thereby preserving the oligopoly status of the few firms already established there. Lobbying by established firms is often the source of these restrictive practices.

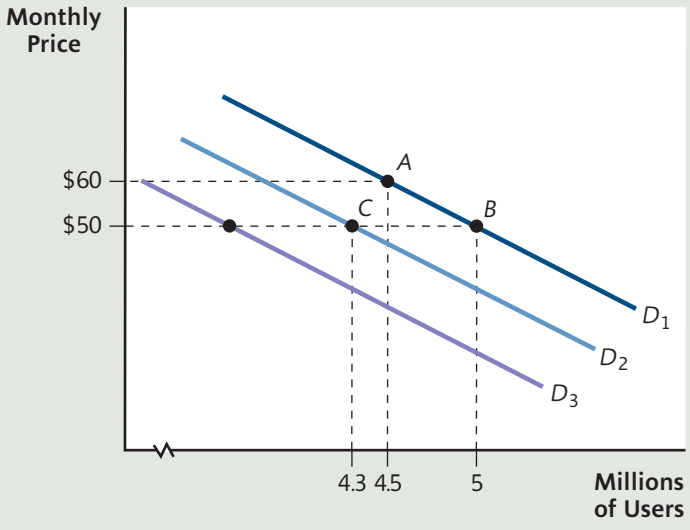
Oligopoly versus Other Market Structures

Of the market structures you have studied in this book, oligopoly presents the greatest challenge to economists.

To see why, look at Figure 3. It shows some demand curves for an unlimited voice-and-text wireless plan that Sprint (through its subsidiary, Boost Mobile) was planning to introduce in early 2009. Let's first consider the curve labeled D_1. It shows (hypothetically) the number of customers Sprint would have at each price—*if* the other three major carriers continued to charge $99 for their own, similar plans. For example, with a price of $60, the diagram shows that Sprint would have 4.5 million customers (point *A*). With a lower price of $50—and the other carriers continuing to charge $99—Sprint's service would attract 5 million customers (point *B*).

If Sprint were a monopolistic competitor (like Kafka Exterminators in our earlier example), it could do the usual: find the marginal revenue curve associated with demand curve D_1, find its marginal cost curve, and then find the profit maximizing number of customers where $MR = MC$. Finally, it would find the price for that output level on its demand curve. It wouldn't have to worry about how the other carriers would respond to that price, because Sprint would be one of many small firms—too small for its decisions to influence the others and elicit a reaction from them.

FIGURE 3 An Oligopolist's Demand Curve Depends on its Competitors' Responses



For an oligopoly firm, such as Sprint, the quantity demanded at each price depends on the response of its competitors. Demand curve D_1 *is drawn assuming that Sprint's competitors maintain their current prices (say, $99). In that case, Sprint can charge $60 and attract 4.5 million users (point* A*) or $50 and attract 5 million users (point* B*). However, if Sprint decides to charge $50, and one of its competitors matches Sprint's price, Sprint's demand curve will shift leftward, perhaps to* D_2*. In that case, at a price of $50, Sprint will attract only 4.3 million users (point* C*). If more than one competitor drops its price to match Sprint's $50 price, the demand curve will shift further leftward, as in the move to* D_3*.*

But because this market is an oligopoly, Sprint could *not* assume that the other wireless carriers would continue to charge $99. On the contrary, Sprint would have to *anticipate* how the other carriers would respond after it priced its own service. Suppose, for example, that Sprint decided to charge $50, and T-Mobile lowered its own price to $60 in response. Then Sprint's demand curve would shift leftward, to a curve like D_2. After all, D_1 was drawn under the assumption that all the other carriers would continue to charge $99; if one of them (T-Mobile) lowers its price, Sprint will have fewer customers at each price than before. Along the new demand curve D_2, Sprint would now have 4.3 million customers at $50 (point *C*), rather than the 5 million customers we found earlier (point *B*).

But wait—what if T-mobile lowered its price all the way to $50? Or what if not just T-mobile, but Verizon and AT&T cut their prices to $50 as well? Then the demand curve Sprint faces would shift even farther to the left, to a curve like D_3. Clearly, which demand curve (and therefore, which associated *MR* curve) Sprint believes it will face depends on what it thinks its rivals will do. This kind of strategic interaction makes it impossible to use the simple "$MR = MC$" approach to find the profit-maximizing price and output of an oligopolist.[4]

Game theory An approach to modeling the strategic interaction of oligopolists in terms of moves and countermoves.

You can see why oligopoly presents such a challenge, not only to the firms themselves, but also to economists studying them. However, one approach, called **game theory,** has yielded rich insights into oligopoly behavior.

The Game Theory Approach

The word *game* applied to oligopoly decision making might seem out of place. Games—like poker, basketball, or chess—are usually played for fun, and even when money is at stake, the sums are usually small. What do games have in common with important business decisions, where hundreds of millions of dollars and thousands of jobs may be at stake?

In fact, quite a bit. In all games—except those of pure chance, such as roulette—a player's strategy must take account of the strategies followed by other players. This is precisely the situation of the oligopolist. Game theory analyzes oligopoly decisions as if they were games by looking at the rules players must follow, the payoffs they are trying to achieve, and the strategies they can use to achieve them.

The Prisoner's Dilemma

The easiest way to understand how game theory works is to start with a simple, noneconomic example—the prisoner's dilemma. This game explains why a technique for obtaining confessions, commonly used by police, is so often successful.

Imagine that two partners in crime (let's call them Rose and Colin) have committed a serious offense (say, murder) but have been arrested for a lesser offense (say, robbery). The police have enough evidence to ensure a robbery conviction, but their evidence for murder cannot be used in court. Their only hope for a murder conviction is to get one or both partners to incriminate the other.

The police typically separate the partners and explain the following to each one: "Look, you're already facing a five-year sentence for robbery. But we'll offer you a deal: If you confess to the murder and implicate your partner, and your partner does *not* confess, we'll make sure that the D.A. goes easy on you. You'll get three years,

[4] In fact, on February 19, 2009—just a month after Sprint priced its new service at $50, T-Mobile lowered the price of its own, similar service from $99 to $50. Verizon and AT&T were expected to follow.

tops. If you and your partner *both* confess, we'll send you each away for 20 years. But if your partner confesses, and you do *not*, we'll send *you* away for 30 years."

We can regard each partner in this situation as a *player* in a *game*. Figure 4 shows the **payoff matrix** for this game, a listing of the payoffs that each player will receive for each possible combination of strategies the two might select. The payoff matrix presents a lot of information at once, so let's take it step-by-step.

Payoff matrix A table showing the payoffs to each of two players for each pair of strategies they choose.

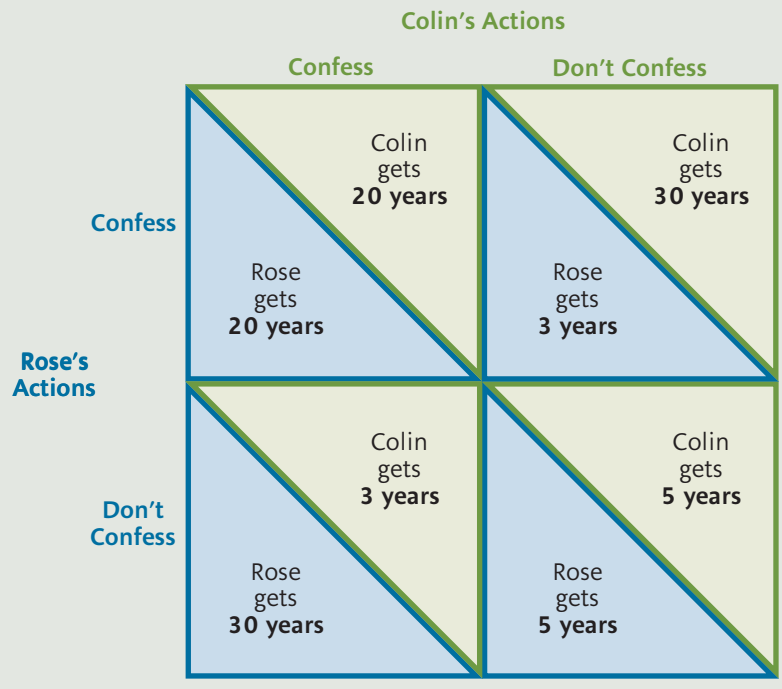
First, notice that each *column* represents a strategy that Colin might choose: confess or not confess. Second, each *row* represents a strategy that Rose might select: confess or not confess. Thus, each of the four boxes in the payoff matrix represents one of four possible strategy combinations that might be selected in this game:

1. Upper left box: Both Rose and Colin confess.
2. Lower left box: Colin confesses and Rose doesn't.
3. Upper right box: Rose confesses and Colin doesn't.
4. Lower right box: Neither Rose nor Colin confesses.

Let's now look at the game from Colin's point of view. The green-shaded entries in each box are Colin's possible *payoffs:* jail sentences. (Ignore the blue-shaded entries for now.) For example, the lower left square shows that when Colin confesses and Rose does not, Colin will receive just a three-year sentence.

Colin wants the best possible deal for himself, but he is not sure what his partner will do. (Remember, they are in separate rooms.) So Colin first asks himself which strategy would be best *if* his partner were to confess. The *top row* of the matrix guides us through his reasoning: "If Rose decides to confess, my best choice would be to confess, too, because then I'd get 20 years rather than 30." Next, Colin determines the best strategy if Rose does *not* confess. As the *bottom row* shows, he'll

FIGURE 4 **The Prisoner's Dilemma**



reason as follows: "If Rose does not confess, my best choice would be to confess, because then I'd get three years rather than five."

Let's recap: If Rose confesses, Colin's best choice is to confess; if Rose does *not* confess, Colin's best choice is—once again—to confess. Thus, regardless of Rose's strategy, Colin's best choice is to confess. In this game, the strategy "confess" is an example of a *dominant strategy:*

Dominant strategy A strategy that is best for a player no matter what strategy the other player chooses.

> *A **dominant strategy** is a strategy that is best for a player regardless of the strategy of the other player.*

If a player has a dominant strategy in a game, we can safely assume that he will follow it.

What about Rose? In another room, she is presented with the *same* set of options and payoffs as her partner, as shown by the blue entries in the payoff matrix. When Rose looks down each *column,* she can see *her* possible payoffs for each strategy that Colin might follow. As you can see (and make sure that you can, by going through all the possibilities), Rose has the same dominant strategy as Colin: confess. We can now predict that *both* players will follow the strategy of confessing. The outcome of the game—the upper left-hand corner—is a confession from both partners, with each receiving a 20-year sentence.

The outcome of this game is an example of a *Nash equilibrium,* appropriately named after the mathematician John Nash, who originated the concept. (Nash won the Nobel Prize in economics in 1994, and was the subject of the film *A Beautiful Mind*.)

Nash equilibrium A situation in which every player of a game is taking the best action for themselves, given the actions taken by all other players.

> *A **Nash equilibrium** is a combination of strategies in which each player is making the best choice for him- or herself, given the choices of all other players.*

We use the term *equilibrium* for this situation because, once the players are in it, neither one would have an incentive to change his or her behavior. In a Nash equilibrium, if we took each player aside, and offered each one the opportunity to make a change, each would turn down the offer.

When there are two players and both have a dominant strategy, that outcome will always be the only Nash equilibrium in the game. After all, when both players have a dominant strategy, they *always* do best by using it, no matter what the other player is doing. And once each player has chosen his or her dominant strategy, neither would change it if offered the chance.

Note, however, that in Figure 4, both Rose and Colin could do even better than the Nash equilibrium if both choose *not* to confess (the lower right corner). But this would require the players to coordinate their decisions and come to an agreement not to confess. Each would have to trust that the other would stick by that agreement. When we try to predict the outcome of a game, however, we take a more narrow view of players' attitudes. We'll assume for now that each player acts in his or her own self-interest, always taking the option that is best for him or her. And in the prisoner's dilemma game, the best option for either player would be to break any such agreement and confess. This is why we end up in the upper left corner in the prisoner's dilemma game. We'll discuss other situations, in which coordinating decisions is a more realistic possibility, in a few pages.

Simple Oligopoly Games

The same method used to understand the behavior of Rose and Colin in the prisoner's dilemma can be applied to a simple oligopoly market. Imagine a town with just two

gas stations: Gus's Gas and Filip's Fillup. This is an example of an oligopoly with just two firms, called a **duopoly.** We'll assume for now that Gus and Filip, like Rose and Colin in the prisoner's dilemma, must make their decisions independently, without knowing in advance what the other will do. We'll consider three types of situations these duopolists might face: (1) both players have dominant strategies, (2) only one player has a dominant strategy, and (3) neither player has a dominant strategy.

Duopoly An oligopoly market with only two sellers.

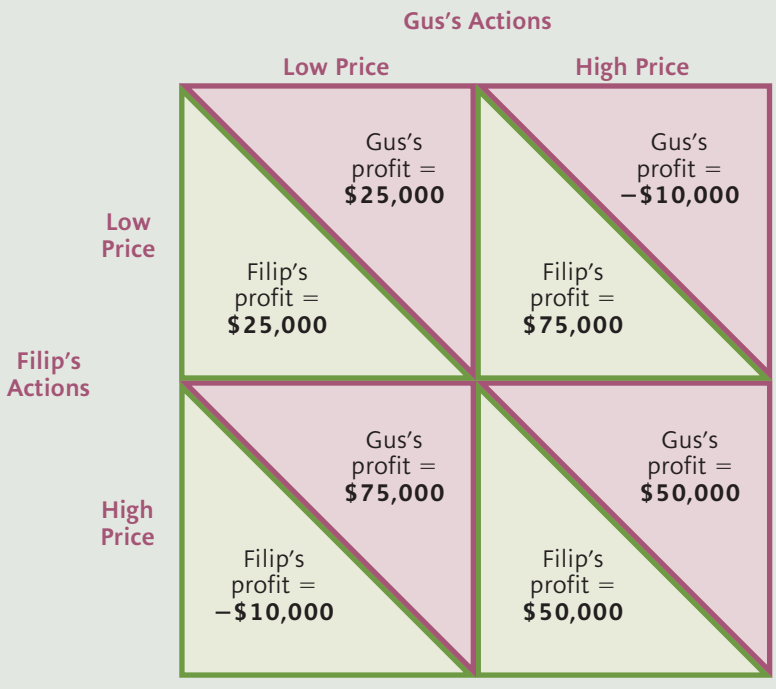
Both Players Have Dominant Strategies

Figure 5 shows an example of a payoff matrix in which (as you're about to see) both players have a dominant strategy. To keep it simple, we've limited each player to two possible actions: charging a high price or a low price for gas. The columns of the matrix represent Gus's possible strategies, while the rows represent Filip's strategies. Each square shows a possible payoff, yearly profit, for Gus (shaded purple) and Filip (shaded green). (Make sure you can see, for example, that if Gus sets a high price and Filip sets a low price, then Gus will suffer a loss of $10,000 while Filip will enjoy a profit of $75,000.)

The payoffs in the figure follow a logic that we find in many oligopoly markets: Each firm will make greater profit if all firms charge a higher price. But the best situation for any one firm is to have its rivals charge a high price, while *it alone* charges a low price and lures customers from the competition. The worst situation for any one firm is to charge a high price while its rivals charge a low one, for then it will lose much of its business to its rivals.

The entries in the payoff matrix in Figure 5 reflect this situation: Profits are higher ($50,000) for both Gus and Filip when they both charge a high price and lower ($25,000) when they both charge a low price. But when the two follow

FIGURE 5 **A Duopoly Game: Both Players Have Dominant Strategies**



different strategies, the low-price firm gets the best possible payoff ($75,000), while the high-price firm gets the worst possible payoff (−$10,000).

Let's look at the game from Gus's point of view, using the purple-shaded entries in the payoff matrix. If Filip chooses a low price (the top row), then Gus should choose a low price, too, since this will get him a $25,000 profit instead of a $10,000 loss. If Filip selects a high price (the bottom row), then, once again, Gus should choose a low price, since this will get him a profit of $75,000 rather than $50,000. Thus, no matter what Filip does, Gus's best move is to charge a low price—his *dominant* strategy.

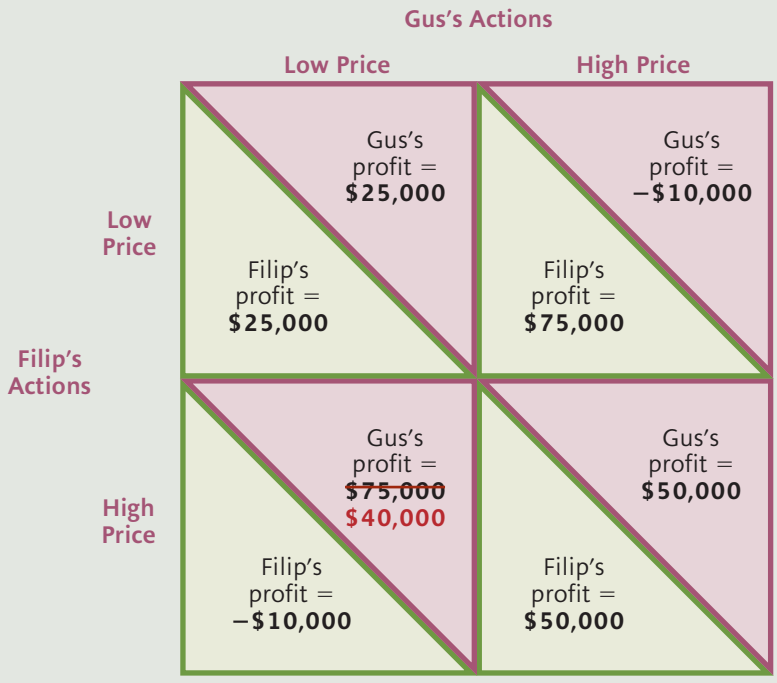
A similar analysis from Filip's point of view, using the green-shaded entries, tells us that his dominant strategy is the same: a low price. Thus, the outcome of this game is the box in the upper left-hand corner, where both players charge a low price and each earns a profit of $25,000.

Notice that our outcome—like the outcome of the prisoner's dilemma—is a Nash equilibrium. Once Gus and Filip reach the upper left-hand corner, each is doing the best that he can do, given what the other is doing. Neither has any incentive to change.

Only One Player Has a Dominant Strategy

Figure 6 shows a payoff matrix that is exactly like the one in Figure 5 except for one alteration: Gus's payoff in the lower left-hand cell has been changed from $75,000 to $40,000. In this new game, Gus no longer has a dominant strategy. If Filip charges a low price, Gus should charge a low price; if Filip charges a high price, Gus

FIGURE 6 A Duopoly Game: Only One Player Has a Dominant Strategy



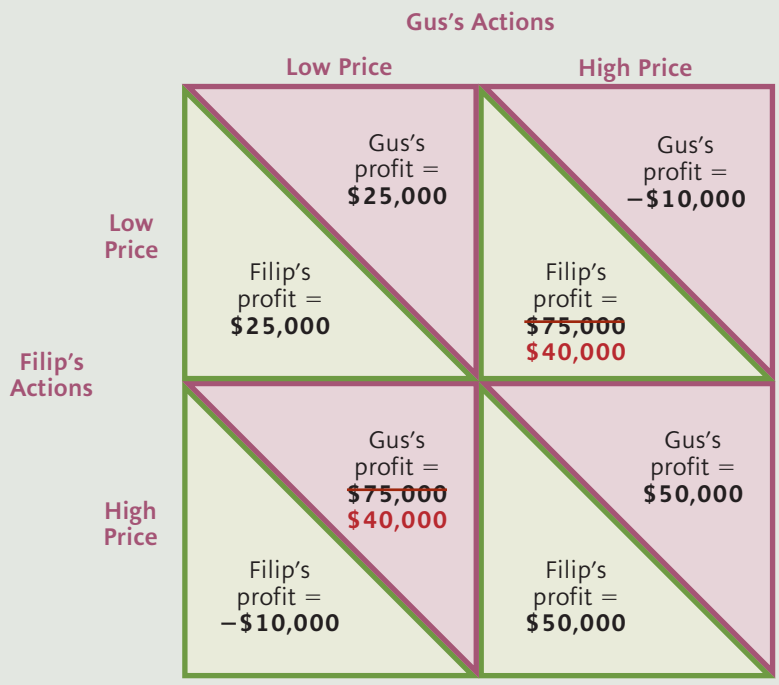
should charge a high price. (Take a moment to verify this before reading on.) Thus, Gus's best choice depends on Filip's choice. However, since we haven't changed any of Filip's payoffs, he still has a dominant strategy: to charge a low price. Since Gus *knows* that Filip will select the low price, Gus will always select a low price too. We know, therefore, exactly where these two station owners will end up: both charging the low price.

No Player Has a Dominant Strategy

What happens in a game in which *neither* player has a dominant strategy? Figure 7 illustrates this situation. Now we've changed *Filip's* $75,000 payoff to $40,000, just as we did for Gus in our previous example. We've already shown in our previous example that Gus does not have a dominant strategy with his payoffs. You can easily prove to yourself that Filip has no dominant strategy either. In fact, in Figure 7, the best strategy for either player (low or high price) depends on the choice of the other. Therefore, it is more difficult to predict an outcome to this game.

Economists are very interested in situations like these, and have come up with some more advanced techniques to help predict how players are likely to behave in such situations. If you continue your study of microeconomics, you'll learn about some of these techniques.[5]

FIGURE 7 A Duopoly Game: Neither Player Has a Dominant Strategy



[5] For those interested, an excellent introduction to game theory, by Don Ross, is available online at *http://plato.stanford.edu/entries/game-theory/*.

Complications

While our simple example helps us understand the basic ideas of game theory, real-world oligopoly situations are seldom so simple. First, there will typically be more than two strategies from which to choose (for example, a variety of different prices or several different amounts to spend on *nonprice* competition such as advertising). Also, there will usually be more than two players, so a two-dimensional payoff matrices like the ones we've been using would not suffice.

Finally, in our examples, we've limited the players to *one* play of the game. While this might make sense in the prisoner's dilemma—where the players get only one chance to make a decision—it is not realistic for most oligopoly markets. In reality, for gas stations and almost all other oligopolies, there is **repeated play,** where both players select a strategy, observe the outcome of that trial, and play the game again and again, as long as they remain rivals. Repeated play can fundamentally change the way players view a game and lead to new strategies based on long-run considerations. One possible result of repeated trials is *cooperative behavior,* to which we now turn.

Repeated play A situation in which strategically interdependent sellers compete over many time periods.

Cooperative Behavior in Oligopoly

In the real world, oligopolists will usually get more than one chance to choose their prices. Pepsi and Coca-Cola have been rivals in the soft drink market for decades, as have Kellogg, Post (now owned by Ralcorp), Quaker (now owned by PepsiCo), and General Mills in the breakfast cereal market. These firms can change their strategies after observing their rivals' strategies.

The equilibrium in a game with repeated plays may be very different from the equilibrium in a game played only once. Often, firms will evolve some form of *cooperation* in the long run.

For example, look again at Figure 5. If this game were played only once, we would expect each player to pursue its dominant strategy, select a low price, and end up with $25,000 in yearly profit. But there is a better outcome for both players. If each were to charge a high price, each would make a profit of $50,000 per year. If Gus and Filip remain competitors year after year, we might expect them to realize that by cooperating, they would both be better off. And there are many ways for the two to cooperate.

Explicit Collusion

The simplest form of cooperation is **explicit collusion,** in which managers meet face-to-face to decide how to set prices. These arrangements are commonly called price-fixing agreements. In our example, Gus and Filip might strike an agreement that each will charge a high price, moving the outcome of the game to the lower right-hand corner in Figure 5, where each earns $50,000 in yearly profit instead of $25,000.

Explicit collusion Cooperation involving direct communication between competing firms about setting prices.

The most extreme form of explicit collusion is the creation of a **cartel**—a group of firms that tries to maximize the total profits of the group as a whole. To do this, the group of firms behaves as if it were a monopoly, treating the market demand curve as the "monopoly's" demand curve. Then, it finds the point on the demand curve—the price and quantity of output—that maximizes total profit. Each member is instructed to charge the agreed-upon price (cartels are often called *price-fixing* agreements), and each is allotted a share of the cartel's total output. This last step is crucial: If any member produces and sells more than its allotted portion, then the group's total *output* rises. The extra output would cause the price to fall below the agreed-upon profit-maximizing price.

Cartel A group of firms that selects a common price that maximizes total industry profits.

The most famous cartel in recent years has been OPEC—the Organization of Petroleum Exporting Countries—which meets periodically to influence the price of oil by setting the amount that each of its members can produce. In the mid-1970s, OPEC quadrupled its price per barrel in just two years, leading to a huge increase in profits for the cartel's members. In the late 1990s, OPEC exerted its muscle once again, doubling the price of oil over a period of 18 months.

If explicit collusion to raise prices is such a good thing for oligopolists, why don't they all do it? A major reason is that it's usually *illegal*. OPEC was not considered illegal by any of the oil-producing nations, but cartels are against the law in the United States, the European Union, and most of the developed nations.[6] In these countries, explicit collusion must be conducted with the utmost secrecy. And the penalties, if the oligopolists are caught, can be severe.

Interestingly, authorities in both the United States and Europe now use a strategy based on the prisoner's dilemma game to uncover price-fixing agreements: The first manager to confess is given automatic amnesty, while those who don't confess (or confess too late) are treated harshly. The U.S. Department of Justice has reported that since this policy went into effect in 1993, the number of corporate confessions and applications for amnesty has increased dramatically. Table 1 lists some of the largest fines imposed by the U.S. and other governments in recent years, after price-fixing arrangements were exposed and prosecuted.

The chances of getting caught, and the severe penalties at stake, often lead oligopolists to other forms of collusion that are harder to detect.

TABLE 1

Some Recent U.S. Price-Fixing Cases

Date of Fine	Company	Product	Fine
Various dates in 2008 and 2009	• Japan Airlines • British Airways • Korean Airlines • Qantas Airlines • 12 other airlines	Air freight	$1.6 billion combined
October 2008	• Sasol • Exxon Mobil • Total SA • Six other firms	Paraffin Wax	$470 million combined
October 2008	• Dole Food • Fresh Del Monte Produce	Bananas	$82 million combined
November 2008	• Asahi Glass • Nippon Sheet Glass • Saint-Gobain	Car Windows	$1.7 billion combined
November 2008–March 2009	• Sharp Corp • LG Display Co. • Chungwa Picture Tubes • Hitachi	LCD displays	$621 million combined

[6] OPEC is not prosecuted only because its members are sovereign governments, rather than private firms.

Tacit Collusion

Tacit collusion Any form of oligopolistic cooperation that does not involve an explicit agreement.

Any time firms cooperate *without* an explicit agreement, they are engaging in **tacit collusion.** Typically, players adopt strategies along the following lines: "In general, I will set a high price. If my rival also sets a high price, I will go on setting a high price. If my rival sets a low price this time, I will punish him by setting a low price next time." You can see that if both players stick to this strategy, they will both likely set the high price. Each is waiting for the other to go first in setting a low price, so it may never happen.

Tit-for-tat A game-theoretic strategy of doing to another player this period what he has done to you in the previous period.

This type of strategy is often called **tit-for-tat**, defined as doing to the other player what he has just done to you. In our gas station duopoly, for example, Gus will pick the high price whenever Filip has set the high price in the previous play, and Gus will pick the low price if that is what Filip did in the previous play. With enough plays of the game, Filip may eventually catch on that he can get Gus to set the desired high price by setting the high price himself and that he should not exploit the situation by setting the low price, because that will cause Gus to set the low price next time. The result of every play will then be a *cooperative outcome:* The players move to the lower right-hand corner of Figure 5, with each firm earning the higher $50,000 in profit.

Tit-for-tat strategies are prominent in the airline industry. When one major airline announces special discounted fares, its rivals almost always announce identical fares the next day. The response from the rivals not only helps them remain competitive, but also provides a signal to the price-cutting airline that it will not be able to offer discounts that are unmatched by its rivals.

Price leadership A form of tacit collusion in which one firm sets a price that other firms copy.

Another form of tacit collusion is **price leadership**, in which one firm, the *price leader,* sets its price, and other sellers copy that price. The leader may be the dominant firm in the industry (the one with the greatest market share, for example), or the position of leader may rotate from firm to firm. In recent decades, American Airlines has behaved as a price leader in many air-travel markets. American's price increases have often been matched within days by Delta, United, and other major airlines.

With price leadership, there is no formal agreement. Rather, the choice of the leader, the criteria it uses to set its price, and the willingness of other firms to follow come about because the firms realize—without formal discussion—that the system benefits all of them.

The Limits to Collusion

It is tempting to think that collusion—whether explicit or tacit—gives oligopolies absolute power over their markets, leaving them free to jack up prices and exploit the public without limit. But oligopoly power, even with collusion, has its limits.

First, even colluding firms are constrained by the market demand curve: A rise in price will always reduce the quantity demanded from all firms together. There is one price—the cartel monopoly price—that maximizes the total profits of all firms in the market, and it will never serve the group's interest to charge any price higher than this.

Second, oligopolies are often weakened—and sometimes destroyed—by new technologies. This is especially true of local oligopolies. A small town, for example, might be able to support only a few stores selling luggage, office equipment, or books. But the Internet has enabled residents in small towns everywhere to choose among dozens or more online sellers of the same merchandise.

Third, collusion is limited by powerful incentives to cheat on any agreement. In Figure 5, for example, suppose Gus and Filip collude to end up in the lower right-hand corner ($50,000 in profit for each). Will they stay there? Perhaps not, because each player has an incentive to cheat by switching back to the low price. The other player may punish the cheater by lowering his own price, and cooperation may be restored. But periodic cheating often plagues oligopolies.

Even the explicit collusion practiced by the OPEC cartel periodically falls apart, because so many members cheat. One member may think that it can secretly sell a bit more than its allotted quantity of oil and earn more revenue. (While the *market* demand for oil is inelastic, the demand for any *one* member's oil is highly elastic.) But when *several* OPEC members use this reasoning and cheat, market quantity rises significantly and the market price falls. At that point, no one wants to stick to a revenue-sacrificing agreement that no one else seems to be following, and cooperation breaks down.

Finally, there is government.

Anti-Trust Legislation and Enforcement

We've already discussed that explicit price-fixing agreements among firms violate the law in most countries. But even tacit collusion can attract the watchful eye of government. Antitrust policies—which are designed to protect the interests of consumers and preserve adequate competition—in the United States and many other countries often prevent oligopolies from forming, or police them when they do.

In practice, antitrust enforcement has focused on three types of actions: (1) preventing collusive agreements among firms, such as price-fixing agreements; (2) breaking up large firms or limiting their activities when market dominance harms consumers; and (3) preventing mergers that would lead to harmful market domination.

The impact of these antitrust actions goes far beyond the specific companies called into the courtroom. Managers of firms even considering anticompetitive moves have to think long and hard about the consequences of acts that might violate the antitrust laws. For example, many economists believe that in the late 1940s and early 1950s, General Motors would have driven Ford and Chrysler out of business or bought them out were it not for fear of antitrust action.

Still, antitrust and other government policies toward business are a part of our *political* system. Although the thrust of these policies is always to preserve competition, the type of competition preserved—and the zeal with which the policies are applied—can shift.

For an example of how antitrust policy can shift, consider Section 2 of the Sherman Act of 1890—a major foundation of current anti-trust policy. The act declares it a felony to "monopolize, or attempt to monopolize, or combine or conspire with any other person or persons, to monopolize any part of the trade or commerce among the several States, or with foreign nations. . . ."

In 2007, the Bush administration issued a report with some recommended guidelines regarding Section 2. It advised against overzealous enforcement, and stressed that there were risks—not just benefits—in hindering the growth of large firms in order to encourage competition, especially when larger firms could achieve efficiencies that smaller firms could not.

In 2009, President Obama's new antitrust chief in the U.S. Department of Justice left little doubt that antitrust policy would take a different direction, declaring:

> *. . . the Section 2 Report lost sight of an ultimate goal of antitrust laws—the protection of consumer welfare. . . . We must change course and take a new tack. . . . For these reasons, I have withdrawn the Section 2 Report by the [Bush administration's] Department of Justice.*[7]

Shortly after this declaration, the Justice Department announced it would investigate several business practices that the previous administration had tacitly accepted.

[7] Christine Varney, "Vigorous Antitrust Enforcement in this Challenging Era," Remarks as Prepared for the United States Chamber of Commerce, May 12, 2009 (http://www.usdoj.gov/atr/public/speeches/245777.htm).

Using the Theory

ADVERTISING IN MONOPOLISTIC COMPETITION AND OLIGOPOLY

© TRAVEL INK/GALLO IMAGES/GETTY IMAGES

We began this chapter by noting that perfect competitors never advertise and monopolies advertise relatively little. But advertising is almost always found under monopolistic competition and very often in oligopoly. Why? All monopolistic competitors, and many oligopolists, produce differentiated products. In these types of markets, the firm gains customers by convincing them that its product is different and better in some way than that of its competitors. Advertising, whether it merely informs customers about the product ("The new Toyota Corolla gets 45 miles per gallon on the highway") or attempts to influence them more subtly and psychologically ("Our exotic perfume will fill your life with mystery and intrigue"), is one way to sharply differentiate a product in the minds of consumers. Since other firms will take advantage of the opportunity to advertise, any firm that *doesn't* advertise will be lost in the shuffle. In this section, we use the tools we've learned in this chapter to look at some aspects of the economics of advertising.

Advertising and Market Equilibrium Under Monopolistic Competition

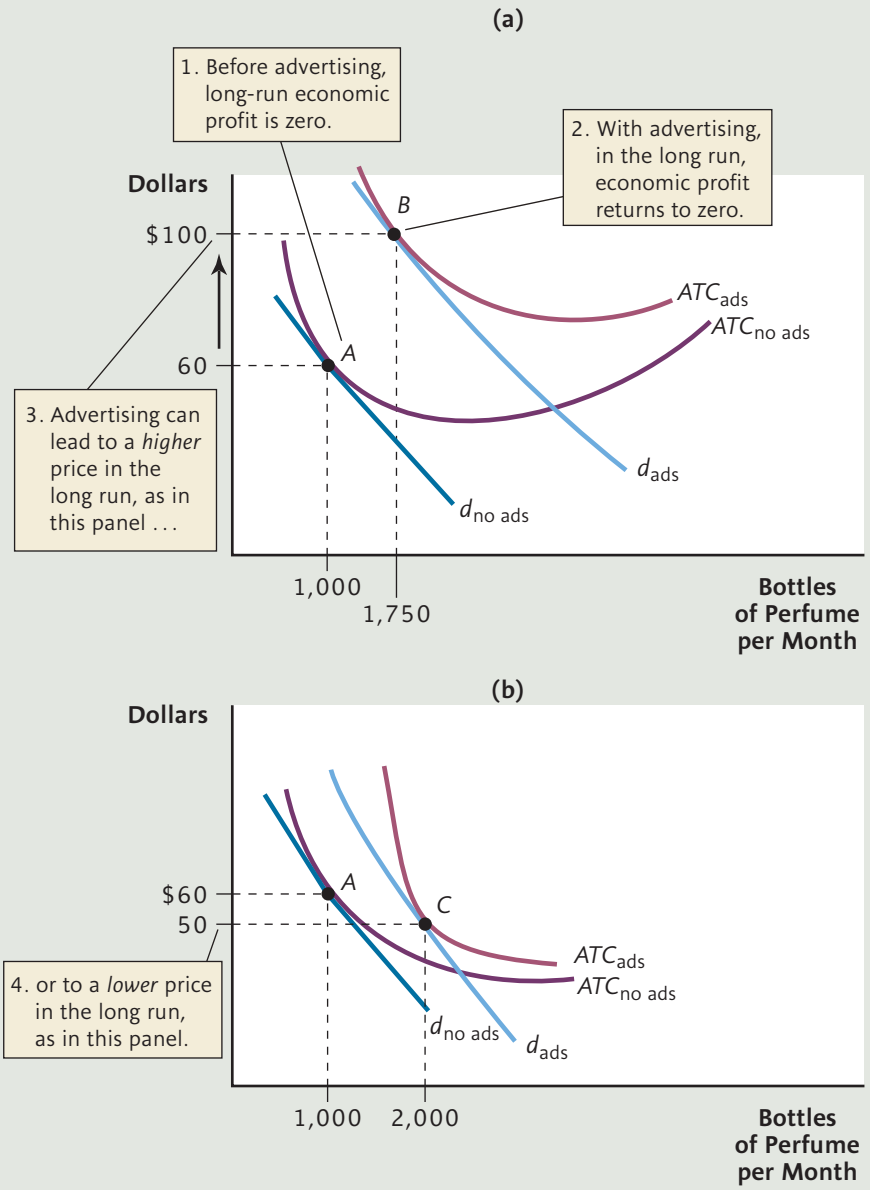
A monopolistic competitor advertises for two reasons: to shift its demand curve rightward (greater quantity demanded at each price) and to make demand for its output *less* elastic (so it can raise price and suffer a smaller decrease in quantity demanded). Advertising costs money, so in addition to its impact on the demand curve, it will also affect the firm's *ATC* curve. What is the ultimate impact of advertising on the typical firm?

Figure 8(a) shows demand and *ATC* curves for a company, Narcissus Fragrance, that manufactures and sells perfume. Initially, there is no advertising at all in the industry. Narcissus is in long-run equilibrium at point *A*, in panel (a), where its demand curve ($d_{\text{no ads}}$) and *ATC* curve ($ATC_{\text{no ads}}$) touch, so economic profit is zero. The firm charges $60 per bottle and sells at the profit-maximizing output level of 1,000 bottles per month. This is the output level where its marginal revenue and marginal cost curves (not shown) intersect.

Now suppose that we introduce advertising into this market. Initially, the first few firms that discover advertising may have a temporary advantage over firms that don't advertise. But remember that in monopolistic competition any successful form of nonprice competition will be automatically replicated by *all* firms (otherwise they would be at a competitive disadvantage). So let's skip over the temporary situation in which only some firms advertise, and examine our new long-run equilibrium when *all* firms advertise. In the long run, how will advertising change the situation of a typical monopolistic competitor in this market?

One change is that, with each firm paying additional costs for advertising, cost per unit will be greater at every output level. So the typical firm's *ATC* curve will shift upward. In panel (a), we show that Narcissus's *ATC* curve shifts upward

FIGURE 8 Advertising in Monopolistic Competition



to ATC_{ads}. Notice, however, that the upward shift is smaller at higher output levels, where the cost of any given ad campaign is spread over a larger number of units.

In addition to the shift in *ATC*, we can expect that with all firms advertising, the demand for the product *in general* will increase. (More people are aware of the product, or have had their appetites stimulated.) And this, in turn, means that *each* firm should be able to sell more units at any given price than before: The demand curve facing each firm shifts rightward.

How much will the typical firm's demand curve shift? We know that, in the long run, the combination of a rightward shift in demand and an upward shift in *ATC* must eventually lead to a new equilibrium in which economic profit is zero. To see why, remember that if advertising creates economic profit in the short run, entry will occur, and every firm's demand curve will then shift leftward. If advertising creates economic loss in the short run, exit will occur, and the remaining firms' demand curves will shift rightward. In the end, long-run equilibrium (a situation of neither entry nor exit) requires that the typical firm earn zero economic profit. And in monopolistic competition, as you've learned, this can only occur when the demand curve touches but does not cross the *ATC* curve, with $P = ATC$ at the profit-maximizing output level.

In panel (a), the new long-run equilibrium for our typical firm, Narcissus, occurs at point *B*. Narcissus sells 1,750 bottles of perfume and charges consumers a higher price ($100) than before. But because it has to pay for advertising, it is breaking even, just as it was in the initial long-run equilibrium without advertising.

In panel (a), the impact of advertising is to *raise* prices for consumers. When consumers buy perfume, they are now paying for the advertising as well as all the inputs they paid for before. But you may be surprised that advertising can also have the opposite result: It can actually *lower* prices for consumers.

Panel (b) illustrates this case. As before, we begin in a long-run equilibrium with no advertising in the market, and Narcissus operating at point *A*. When we introduce advertising to all firms, each firm (including Narcissus) sees its *ATC* curve shift upward, to ATC_{ads}. But this time, when long-run equilibrium is restored with zero economic profit (point *C*), Narcissus is charging only $50—less than the initial $60. Advertising has brought down the price of perfume.

How can this be? By advertising, each firm is able to produce and sell more output. This remains true even when *all* firms advertise because total market demand has increased. Since the firm was originally on the downward-sloping portion of its *ATC* curve, we know that its *non*advertising costs per unit will decline as output expands. If this decline is great enough—as in panel (b)—then costs per unit will drop, even when the cost of advertising is included. In other words, because you and I and everyone else is buying more perfume, each producer can operate with lower costs per unit. In the long run, entry will force each firm to pass the cost savings on to us.

Our analysis suggests the following conclusion:

Under monopolistic competition, advertising may increase the size of the market so that more units are sold. But in the long run, each firm earns zero economic profit, just as it would if no firm were advertising. The price to the consumer may either rise or fall.

Advertising and Collusion in Oligopoly

In this chapter, you've learned that oligopolists have a strong incentive to engage in tacit collusion. But such collusion is difficult to detect. When one firm raises its prices and others follow, that may be evidence of price leadership, or it may be that costs in the industry have risen, and *all* firms—affected in the same way—have decided independently to raise their prices. But in some cases, such as strategic decisions about advertising, we can use a simple game theory model to show that collusion is almost certainly taking place.

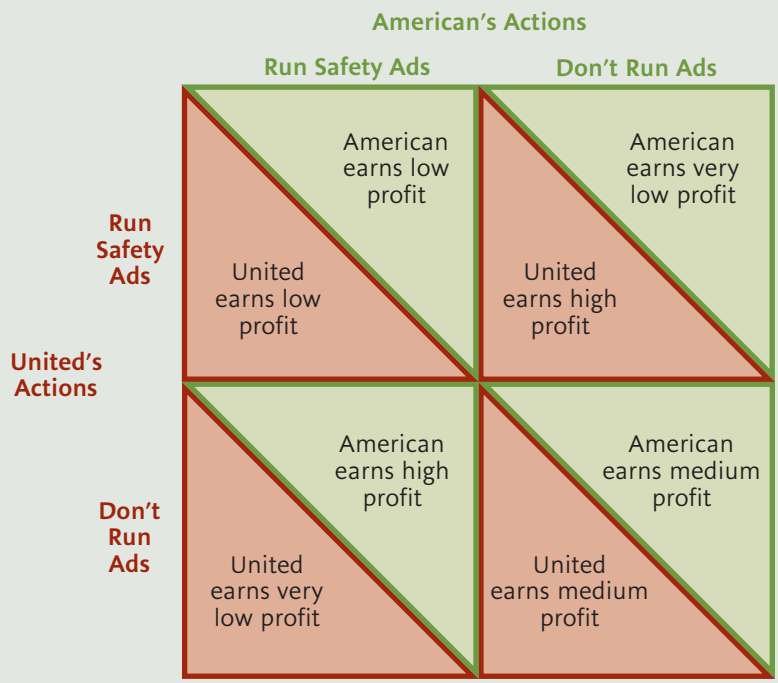
Let's take the airline industry as an example. Polls have shown that passengers are always very concerned about airline safety. Any airline that could convince the public of its superior safety record would profit considerably.

And there are so many different ways to interpret safety data that almost any airline could come up with a measure by which it would appear the "safest." Moreover, any airline that actually took steps to improve safety could tout those policies in its ads, taking business from other airlines. Yet airlines never run advertisements with information about their safety or security policies or attacked those of a competitor. Let's see why.

Figure 9 shows some hypothetical payoffs from this sort of advertising as seen by two firms, United Airlines and American Airlines, competing on a particular route. Focus first on the top, green-shaded entries, which show the payoffs for American. If neither firm ran safety ads, American would earn a level of profit we will call *medium,* as a benchmark. If American ran ads touting its own safety, but United did not, American's profit would certainly increase—to "high" in the payoff matrix. If both firms ran safety ads—especially negative ads that attacked their rival—the public's demand for airline tickets would certainly decline. Reminded of the dangers of flying, more consumers would choose to travel by train, bus, or car. American's profit in this case would be lower than if *neither* firm ran ads, so we have labeled it "low" in the payoff matrix. Finally, the worst possible result for American—"very low" in the figure—occurs when United touts its own safety record, but American does not.

Now consider American's possible strategies. If United decides to run the ads (the top row), American's best action is to run them as well. If United does not

FIGURE 9 **An Advertising Game**



run the ads (bottom row), American's best action is still to run the ads. Thus, American has a dominant strategy: Regardless of what United does, it should run the safety ads.

As you can verify, United, whose payoffs are in the lower, red-shaded entries, faces an entirely symmetrical situation. It, too, has the same dominant strategy: Run the ads. Thus, when each airline acts independently, the outcome of this game is shown in the upper left-hand corner, where each airline runs ads and earns a low profit.

So why don't we observe that outcome?

The answer is that the airlines are playing against each other repeatedly and reach the kind of cooperative equilibrium we discussed earlier. Each airline can punish its rival next time if it fails to cooperate this time. In the cooperative outcome, each airline plays the strategy that it will *not* run the ads as long as its rival does not. The game's outcome moves to the lower right-hand corner. Here, neither firm runs ads, and each earns medium rather than low profit. This is the result we see in the airline industry.

Until the 1980s, a similar collusive understanding seemed to characterize the automobile industry. As long as the "Big Three" dominated auto sales in the United States, the word *safety* was never heard in their advertising. There seemed to be an understanding that all three would earn greater profits if consumers were *not* reminded of the dangers of driving.

Things changed in the 1980s, however, as foreign firms' share of the U.S. market rose dramatically. One of the new players, Volvo, decided that its safety features were so far superior to its competitors that it no longer paid to play by the rules. Volvo began running television advertisements that not only stressed its own safety features but implied that competing products were dangerous. (On a rainy night, a worried father stops his son at the door, hands him some keys, and says, "Here, son, take the Volvo.") Once Volvo began running ads like these, the other automakers had no choice but to reciprocate. Now, automobile ads routinely mention safety features like antilock brakes and air bags.

The Four Market Structures: A Postscript

You have now been introduced to the four different market structures: perfect competition, monopoly, monopolistic competition, and oligopoly. Each has different characteristics, and each leads to different predictions about pricing, profit, nonprice competition, and firms' responses to changes in their environments.

Table 2 summarizes some of the assumptions and predictions associated with each of the four market structures. While the table is a useful review of the *models* we have studied, it is not a how-to guide for analyzing real-world markets: We cannot simply look at the array of markets we see around us and say, "This one is perfectly competitive," "That one is an oligopoly," and so on. Why not? Because markets in the real world will typically have characteristics of more than one kind of market structure. A particular barbecue restaurant, for example, may be viewed as a monopolistic competitor in the market for *restaurants* in Memphis, or an oligopolist in the market for *barbecue* restaurants in Memphis, or a monopolist in the market for barbecue restaurants *within walking distance of Graceland*.

But, as we've seen several times in this text, our choice of model is not really arbitrary. Rather, it depends on the *questions we are trying to answer.* Suppose we're interested in explaining why a *particular* barbecue restaurant with no nearby competitors earns economic profit year after year, or why it spends so much of its

TABLE 2

A Summary of Market Structures

	Perfect Competition	Monopolistic Competition	Oligopoly	Monopoly
ASSUMPTIONS:				
Number of firms	Very many	Many	Few	One
Output of different firms	Standardized	Differentiated	Standardized or differentiated	—
View of pricing	Price taker	Price setter	Price setter	Price setter
Barriers to entry or exit?	No	No	Yes	Yes
Strategic interdependence?	No	No	Yes	No
PREDICTIONS:				
Price and output decisions	$MC = MR$	$MC = MR$	Through strategic interdependence	$MC = MR$
Short-run profit	Positive, zero, or negative	Positive, zero, or negative	Positive, zero, or negative	Positive, zero, or negative
Long-run profit	Zero	Zero	Positive or zero	Positive or zero
Advertising?	Never	Almost always	Maybe, if differentiated product	Sometimes

profit on rent-seeking activity (lobbying the local zoning board). Then, we would most likely use the monopoly model. If we want to explain why *most* barbecue restaurants do *not* earn much economic profit, or why they pay for advertisements in the yellow pages and the local newspapers, or why there is so much excess capacity (empty tables) in the industry, we would use the model of monopolistic competition. To explain a price war among the few restaurants in a neighborhood, or to explore the possibility of explicit or tacit collusion in pricing or advertising, we would use the oligopoly model. And if we're interested in barbecue restaurants as an example of *restaurants in general,* and we want general explanations about restaurant prices, or the expansion or contraction of the restaurant industry in a city or country, we would use the perfectly competitive model (supply and demand curves for restaurant meals).

SUMMARY

Monopolistic competition is a market structure in which there are many small buyers and sellers, easy entry and exit, and firms sell differentiated products. As in monopoly, each firm faces a downward-sloping demand curve, chooses the profit-maximizing quantity where $MR = MC$, and charges the maximum price it can for that quantity. As in perfect competition, short-run profit attracts new entrants. As firms enter the industry, the demand curves facing existing firms shift leftward. Eventually, each firm earns zero economic profit and produces at greater than minimum average cost.

An *oligopoly* is a market structure dominated by a small number of strategically interacting firms. New entry is deterred by economies of scale, reputational barriers, strategic barriers, and legal barriers to entry. Because each firm, when making decisions, must anticipate its rivals' reactions, oligopoly behavior is hard to predict. However, one approach, *game theory,* has offered rich insights.

In game theory, a *payoff matrix* indicates the payoff to each firm for each combination of strategies adopted by that firm and its rivals. A *dominant strategy* is a strategy that is best for a particular firm regardless of what its rival does. If there is no cooperation among firms, any firm that has a dominant strategy will play it, and that helps predict the outcome of the game.

Sometimes oligopolists can cooperate to increase profits. *Explicit collusion*, in which managers meet to set prices, is illegal in the United States and many other countries. As a result, other forms of *tacit collusion* have evolved. Still, collusion is often countered by cheating, by technological change, and by the government's antitrust policies.

PROBLEM SET

Answers to even-numbered Questions and Problems can be found on the text Web site at www.cengage.com/economics/hall.

1. Draw the relevant curves to show a monopolistic competitor suffering a loss in the short run. What will this firm do in the long run if the situation does not improve? (Assume its *ATC* and *MC* curves don't change in the long run) How would this action affect *other* firms in this market?

2. Draw the relevant curves to show a monopolistic competitor earning an economic profit in the short run. Graphically show what this firm can expect to happen to this economic profit in the long run. (Assume its *ATC* and *MC* curves don't change in the long run)

3. The owner of an optometry practice, in a city with more than a hundred other such practices, has the following demand and cost schedules for eye exams:

Price per Eye Exam	Eye Exams per Week	Total Cost per Week	Total Revenue per Week	Marginal Revenue	Marginal Cost
$100	100	$10,500			
$ 80	140	$10,800			
$ 60	200	$11,300			
$ 40	310	$12,290			
$ 20	550	$14,762			

   a. Fill in the columns for total revenue, marginal revenue, and marginal cost. (Remember to put *MR* and *MC between* output levels.)

   b. Briefly explain why an optometry practice (like this one) might face a downward-sloping demand curve, even if it is one out of more than a hundred. (Hint: What might make this market monopolistically competitive rather than perfectly competitive?)

   c. Use the data you filled in for the marginal revenue and marginal cost columns to find the profit-maximizing price and the profit-maximizing number of eye exams per week for this practice.

4. Tino owns a taco stand in Houston, Texas, where there are dozens of other taco stands. He faces the following demand and cost schedules for his taco plates (two tacos and a side of refried beans):

Price per Taco Plate	Taco Plates per Week	Total Cost per Week	Total Revenue per Week	Marginal Revenue	Marginal Cost
$5	50	$ 30			
$4	80	$ 50			
$3	150	$ 176			
$2	800	$1,476			
$1	1,100	$2,136			

   a. Fill in the columns for total revenue, marginal revenue, and marginal cost and use the table to find the profit-maximizing price and the profit-maximizing number of taco plates per week for

Tino's Taco Stand. (Remember to put *MR* and *MC between* output levels.)

b. Redo the table to show what will happen in the short run if Tino spends $100 on an advertising campaign that increases the quantity demanded at each output level by 20 percent. What will happen to his profit-maximizing price and profit-maximizing number of taco plates per week? Do you expect this outcome to persist? Explain.

5. Suppose that the cost data in problem 3 are for the short run, and that the owner of the practice suddenly realizes that she forgot to include her only fixed cost: her license fee of $2,600 per year (which is $50 per week). Should the practice shut down in the short run? Why or why not?

6. Assume that the plastics business is monopolistically competitive.
   a. Draw a graph showing the long-run equilibrium situation for a typical firm in the industry. Clearly label the demand, *MR, MC,* and *ATC* curves.
   b. One of the major inputs into plastics is oil. Draw a new graph illustrating the short-run position of a plastics company after an increase in oil prices. Again, show all relevant curves.
   c. If oil prices remain at the new, higher level, what will happen to get firms in the plastics industry back to a long-run equilibrium? (Assume the firm's *ATC* and *MC* curves don't change in the long run)

7. Draw a diagram, including demand, marginal revenue, marginal cost, and any other curves necessary, to illustrate each of the following two situations for a monopolistic competitor:
   a. The firm is suffering a loss, and should shut down in the short run.
   b. The firm is suffering a loss, but should stay open in the short run.

8. In a small Nevada town, Ptomaine Flats, there are only two restaurants, the Road Kill Cafe and, for Italian fare, Sal Monella's. Each restaurant has to decide whether to clean up its act or to continue to ignore health code violations.

   Each restaurant currently makes $7,000 a year in profit. If they both tidy up a bit, they will attract more patrons but must bear the (substantial) cost of the cleanup; so they will both be left with a profit of $5,000. However, if one cleans up and the other doesn't, the influx of diners to the cleaner joint will more than cover the costs of the scrubbing; the more hygienic place ends up with $12,000, and the grubbier establishment incurs a loss of $3,000.
   a. Write out the payoff matrix for this game, clearly labeling strategies and payoffs to each player.
   b. What is each player's dominant strategy?
   c. What will be the outcome of the game? Explain your answer.
   d. Suppose the two restaurants believe they will face the same decision repeatedly. How might the outcome differ? Why?
   e. Assume that if one cleans up and one stays dirty, the cleaner restaurant makes only $6,000 in profit. All other payoffs are the same as before. What will the outcome of the game be now without cooperation? With cooperation?

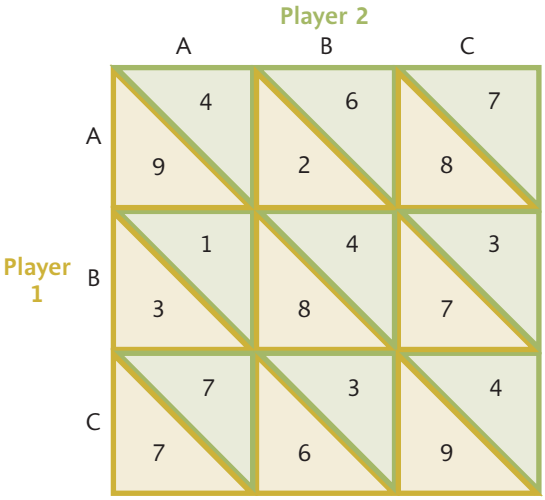
9. Professor Clemens has two students enrolled in his riverboat pilot course, Huck and Tom. The final exam counts as 100 percent of the course grade. If one student passes the exam and one student fails, Professor Clemens announces that he will assign the passer an A and the failer an F. If both students pass, he will give them both Bs. If both students fail, he will give them both Cs. Assume that if each student studies, he passes the exam; if he doesn't study, he fails. Finally, assume that although studying is hard, either student would prefer to study and get an A or a B than not study and get a C or lower.
   a. Write out the payoff matrix for Tom and Huck, clearly labeling strategies and identifying payoffs for each player for each combination of strategies (the payoffs will be letter grades).
   b. What is each player's dominant strategy?
   c. What will be the outcome of the game? Explain your answer.
   d. Could Huck and Tom benefit by cooperating (i.e., coordinating their strategies in this game)? Why or why not?
   e. Now suppose that Professor Clemens decides to penalize Huck for talking in class. He tells Huck that if he passes the exam and Tom does not, Huck will get a C instead of an A. All other payoffs will remain the same. What will be the likely outcome if the test is only offered once? What will be the likely outcome if the test is offered 50 times?

10. Assume that Nike and Adidas are the only sellers of athletic footwear in the United States. They are deciding how much to charge for similar shoes. The two choices are "High" (H) and "Outrageously High" (OH). Nike's payoffs are in the lower left of each cell in the payoff matrix below:
    a. Do both companies have dominant strategies? If so, what are they?
    b. What will be the outcome of the game?
    c. If Nike becomes the acknowledged price leader in the industry, what will be its dominant strategy? What will be the outcome of the game? Why?

More Challenging

11. Suppose that the government has decided to tax all the firms in a monopolistically competitive industry. Specifically, suppose it levies a fixed tax on each firm; that is, the amount of the tax is the same regardless of how much output the firm produces. In the short run, how would that tax affect the price, output level, and profit of the typical firm in that industry? What would be the effect in the long run?

12. To the right, you will find the payoff matrix for a two-player game, where each player has three possible strategies: *A*, *B*, and *C*. The payoff for player 1 is listed in the lower left portion of each cell. Assume there is no cooperation among players.
    a. Does either player have a dominant strategy? If so, which player or players, and what is the dominant strategy?
    b. Can we predict the outcome of this game from the payoff matrix using the methods you've learned in the chapter? Why or why not?
    c. Suppose that strategy C is no longer available to either player. Does either player have a dominant strategy now? Can you now predict the outcome of the game? Explain.



CHAPTER 12

Labor Markets

When we went to do the deals for Shrek 2, *they were made in one day. It was that fast and that easy. It was also probably the biggest payday in movie history. They were each paid $10 million for what is in effect 18 hours of work.*

Jeffrey Katzenberg, cofounder of DreamWorks SKG, referring to payments made to Eddie Murphy, Mike Myers, and Cameron Diaz for voiceovers.[1]

Imagine, for a pleasant moment, that you are someone like Eddie Murphy, Mike Myers, or Cameron Diaz, doing voiceovers for a major studio production. Your work day might begin in a limousine, escorting you to the site of the day's recording. From the moment you arrive, you are doted on by assistants whose sole job is to keep you happy. You spend a few hours reading from a script. If you make a mistake, everyone laughs good-naturedly, and you get another chance to get it right—as many chances as you need. And after doing this each day for a few weeks, you pick up a check for $10 million.

Now, let's switch gears. Imagine that you are the typical short-order cook at a coffeehouse. You spend the day sweating over a hot grill, cooking several hundred meals, dealing with the tempers of waiters and waitresses who want you to do it faster, and who glare at you if you forget that a customer wanted French fries instead of home fries. For this work, you would earn about $9.25 per hour, giving you an income of about $19,000 for the year.

Granted, this is an extreme comparison. But we observe sizeable differences in job earnings throughout the spectrum of the labor market. For example, in 2009, while the median hourly job earnings of high-school graduates was about $15, the median for those with bachelor's degrees was about $25.

To understand where pay differences come from—at the extremes and in the middle—you need to understand *labor markets,* the subject of this chapter. These are the markets where people supply their labor to employers: business firms, government agencies, and others.

Labor Markets in Perspective

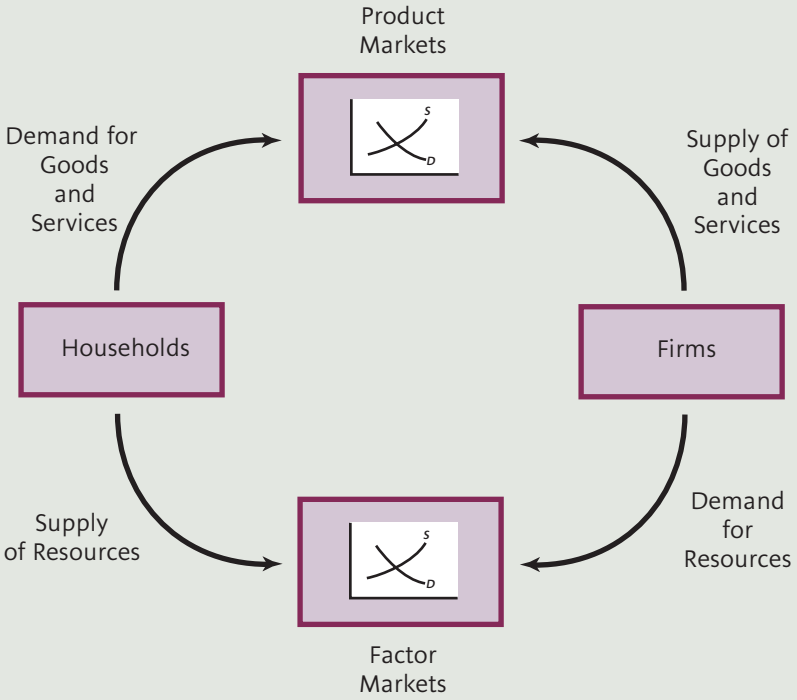
Before we focus on labor markets specifically, let's begin with some perspective. Most of the markets we've analyzed so far (maple syrup, crude oil, solar panels, pharmaceuticals, and more) were **product markets,** in which firms sell goods and

Product markets Markets in which firms sell goods and services to households.

[1] "Question and Answer: Movie Mogul Jeffrey Katzenberg," *Reel West* (Vol. 17, No. 4) July–August 2002.

FIGURE 1 Product Markets and Factor Markets

In product markets, households demand goods and services, and firms supply them. In factor markets, the roles are reversed: Firms demand resources—such as labor, capital, land, or households supply them.



services to households or other firms. Of course, products aren't made out of thin air, but rather from the economy's resources, such as labor, capital, and land. The use of these resources must be purchased from those who own them. Because resources are sometimes called factors of production, the markets in which they are traded are called **factor markets.** Labor markets, for example, are a type of factor market.

Factor markets Markets in which resources—labor, capital, land and natural resources, and entrepreneurship—are sold to firms.

In this and the next chapter, we switch our focus from product markets to factor markets. We'll be using some familiar tools: profit maximization, marginal decision making, equilibrium, and more. But factor markets differ from product markets in important ways.

Figure 1 illustrates one important difference. It shows another version of the circular flow model from Chapter 3. In this version, we've left out the money flows in order to highlight the roles of product and factor markets in the economy. As you can see, in product markets households typically demand the products and firms supply them. In factor markets, these roles are reversed: Firms demand resources such as labor, land, or capital, and they are supplied by the households who own these resources.

The figure also illustrates how behavior in product and factor markets is connected. When households demand more of a good, firms respond by producing more, which makes them demand greater quantities of resources. For example, if households want to buy more cars and Ford Motor Company produces more of them, the company will need to hire more labor, use more machinery, and so on. It will also use more inputs from other firms—glass, steel, tires—and these firms, in turn, will demand more resources. Thus, the demand for resources in the economy arises *from* the demand for goods and services.

The demand for a resource—such as labor—is a ***derived demand.*** *That is, it arises from, and will vary with, the demand for the firm's output.*

Derived demand The demand for a resource that arises from, and varies with, the demand for the product it helps to produce.

Defining a Labor Market

When you begin working or searching for a job after college, you will become a seller in a labor market. But which labor market? As we've suggested several times in this book, how broadly or narrowly we define a market depends on the specific questions we wish to answer.

For example, suppose we want to understand why college graduates, on average, earn more than those with only high school diplomas. For this purpose, we would look at two broadly defined labor markets: one for all college graduates in the country, another for all those who have only high school diplomas. If we want to know how salaries in some professions (say, physicians) are determined, we'd look at the market for all physicians in the country. Or if we want to know why physicians in Boston earn unusually high wages, we'd narrow our definition even further, to the market for physicians in Boston.

The Wage Rate

Much of our focus in this chapter will be on *price* in a labor market: the *wage rate*. It is common to think of a wage rate as an hourly rate, in part because many jobs pay by the hour. But we can calculate an hourly wage rate for *anyone* who works, including a manager paid a salary, a sales representative paid by commission, or a movie star with a contract. The hourly wage rate is simply their total earnings over a week, month, or year divided by the number of hours worked during that period.

For example, if someone earns $80,000 in commissions during the year and works 2,000 hours that year, the hourly wage rate would be $80,000/2,000 = $40.

Competitive Labor Markets

In the first part of this chapter, we'll be looking at *perfectly competitive* labor markets. Competition in labor markets is analogous to competition in product markets. More specifically, a **perfectly competitive labor market** has the following four characteristics:

Perfectly competitive labor market A market with many well-informed buyers and sellers of standardized labor, with no barriers to entry and exit.

1. *Many buyers and sellers:* So many firms demand labor, and so many households supply it, that no decisions by a single firm or worker has a noticeable effect on the labor demanded or supplied in the market.
2. *Standardized labor quality:* To employers, any worker who meets the basic skill requirements for the job is considered just as productive as any other worker.
3. *Easy Entry and Exit:* No artificial barriers prevent workers from entering or leaving a labor market or from acquiring the basic skills needed to work there.
4. *Well-Informed Buyers and Sellers:* Firms and households have all the information they need to make decisions about demanding or supplying labor.

Because labor quality is standardized, and each firm is such a small employer in the market, its employment decisions do not noticeably affect the market wage.

Just as we call a firm in a competitive product market a "price taker," we can call a firm in a competitive *labor* market a "wage taker." The firm takes the wage rate as given.

How is the wage rate determined in a competitive labor market? The same way that a price is determined in *any* competitive market: by the forces of supply and demand.

Labor Demand

When we refer to *labor demand* in a market, we mean the demand by *all* firms for the type of labor being traded in that market. For example, in the market for registered nurses in the United States, all employers of registered nurses—hospitals, medical practices, home health care providers, government agencies and more—would demand this labor. In a more narrowly defined market—say, the market for *hospital nurses* in *Chicago*—we would view only hospitals located in Chicago as demanding labor.

The Labor Demand Curve

Labor demand curve Curve indicating the total number of workers all firms in a labor market want to employ at each wage rate.

Let's look at the market for all nurses in the United States. Figure 2 shows a hypothetical **labor demand curve** (L^D) for this broadly defined labor market. It tells us the number of nurses that *all* U.S. employers would want to hire at each wage rate. Notice that the labor demand curve slopes *downward:* A rise in the wage rate (from $25 to $35)—holding all other influences on demand constant—causes the quantity of labor demanded to fall (from 1.0 million to 0.5 million). This suggests that

FIGURE 2 The Labor Demand Curve

The labor demand curve slopes downward because, all else equal, firms want to employ fewer workers from a labor market when the wage rate is higher. For the hypothetical labor demand curve shown here, firms will want to hire 1 million nurses when the wage rate is $25 per hour, but only 0.5 million nurses at a wage rate of $35.

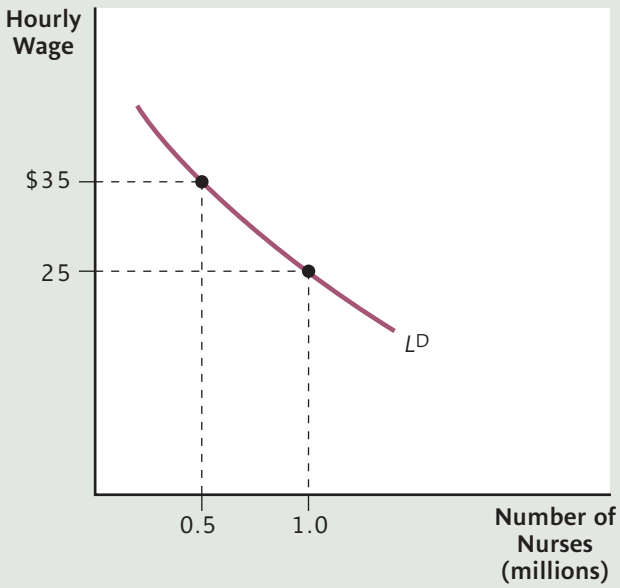
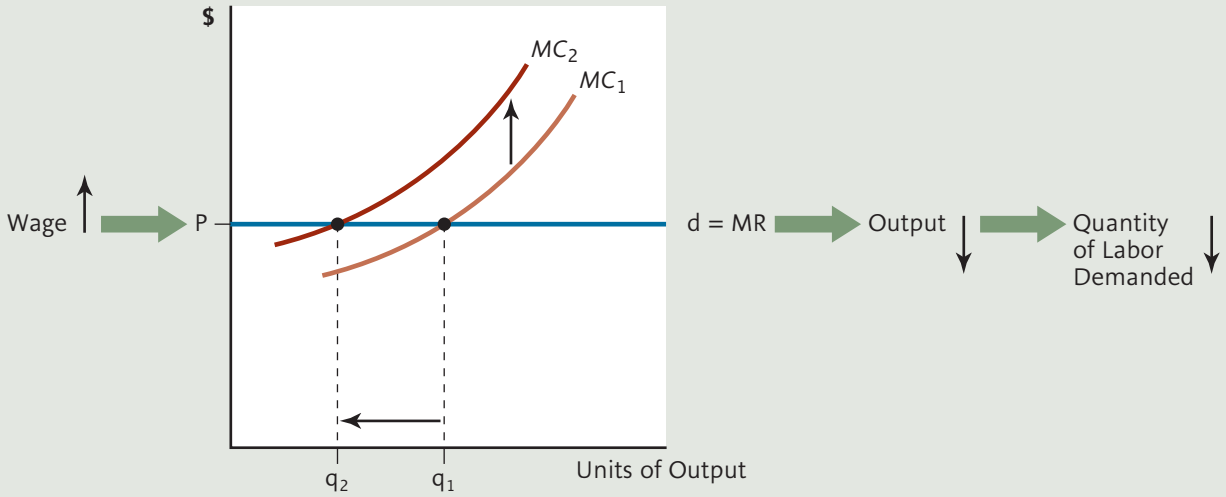


FIGURE 3 The Output Effect of a Rise in the Wage Rate



All else equal, a rise in the wage rate shifts up a firm's marginal cost curve, and therefore lowers its profit-maximizing output level (from q_1 to q_2). Because it is producing less output, the firm will want to employ fewer workers.

individual *firms* in this labor market want to employ fewer workers at higher wage rates. Why?

In the appendix, we show how to find the firm's quantity of labor demanded at each wage rate. Here, we focus on how the quantity of labor demanded *changes* as the wage rate changes.

The Output Effect

Over most time horizons, labor is considered a variable input, used to calculate the firm's marginal cost. So when the wage rate rises, a firm's *marginal cost curve* shifts upward. (It costs more to produce each *additional* unit of output when the wage is high than when the wage is low.) The upward shift in the marginal cost curve decreases the firm's profit-maximizing output level, as illustrated in Figure 3. Finally, remember that the demand for resources is a derived demand: As the firm's rate of *production* falls, so will the quantity of labor it wants to employ. This is the **output effect** of a wage change.

Output effect A change in the wage rate alters the profit-maximizing output level, and therefore changes the quantity of labor demanded.

For example, suppose the going wage rate for registered nurses rises from $25 to $35 per hour, and nothing else changes. Then a home health-care firm—facing an increase in marginal cost at each output level—would want to serve fewer clients, and would employ fewer nurses. The same would happen at other firms that employ nurses: higher wages would lead to higher marginal costs, lower output, and a decrease in the quantity of nurses demanded.

The Input-Substitution Effect

A higher wage rate in a labor market has a second effect: It raises the price of that labor *relative to the price of other inputs*. The other inputs—now relatively cheaper—may include capital or even other types of labor. As firms switch to using more of the

Input-substitution effect A change in the wage rate alters the price of labor relative to the costs of other inputs, and therefore changes the quantity of labor demanded.

other inputs whose prices have *not* risen, they demand less of the labor whose price *has* risen. This is the **input-substitution effect** of a wage change.

For example, a health-care firm may use several types of labor to provide patient care, including nurses, nurses aides, physician assistants, and more. All else equal, a rise in the wage rate of registered nurses would raise their price *relative* to the price of these other workers. To produce any *given* quantity of health care services, the firm's least-cost method may change: fewer nurses, more of the other types of labor.

In sum:

> *A market labor demand curve slopes downward because a rise in the wage rate has two effects: (1) It increases firms' marginal costs, causing them to decrease production and employ fewer workers (the output effect); and (2) it increases the relative cost of labor from that market, causing firms to substitute other inputs, such as capital or other types of labor (the input-substitution effect).*

Shifts in the Labor Demand Curve

You've seen that a change in the market wage rate moves us along a labor demand curve. But when the demand for labor is affected by something *other* than a change in the wage rate, the curve will shift. Let's discuss two important causes of such shifts.

Changes in Demand for the Product

Demand in a *product* market can increase for any number of reasons—increasing population, rising incomes, changes in tastes, and more (See Chapter 3). When demand increases in the product market, the labor demand curve will shift.

Let's take an example. Suppose that home health-care firms (which employ a significant fraction of the nation's nurses) sell in a competitive *product* market. If demand for home health care increases, then its price will rise. This, in turn, will raise the profit-maximizing output level (number of patients served) at each firm, increasing the number of nurses they want to employ. (In the long run, it might lead to entry of new firms, increasing the demand for labor further.)

Notice that we have not referred to any change in the wage rate in this story. When the demand for the product—health care—rises, each firm wants to employ more nurses—*at any given wage rate for nurses*. So the entire labor demand curve for nurses shifts rightward, as in Figure 4.[2]

Changes Involving Other Inputs

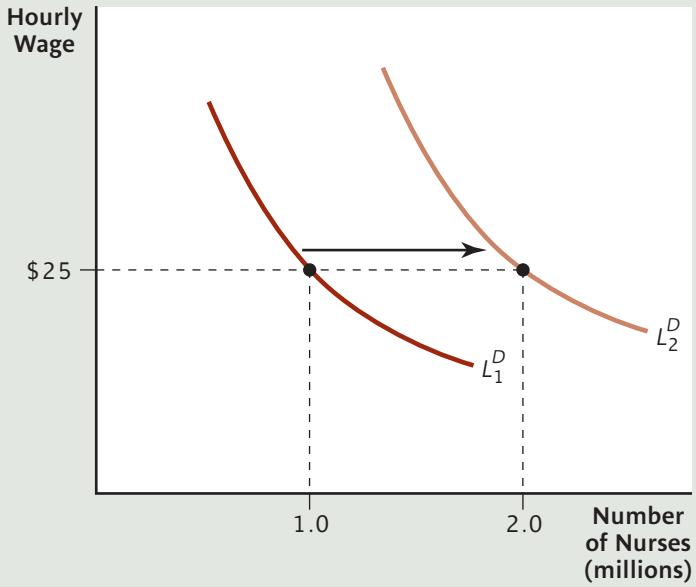
As we've discussed, firms use *other* inputs besides the labor we are considering. Changes in the price or availability of these other inputs can affect the labor demand curve. The effect depends on the relationship between the labor and the other input.

Complementary input An input that is used *by* a particular type of labor, making it more productive.

A **complimentary input** is one that is used *by* the labor we're analyzing, making it more productive and therefore more profitable for the firm to employ.

[2] Even if home health care is *not* sold in a perfectly competitive market, an increase in demand will still shift the labor demand curve rightward. As you've learned (see Chapters 10 and 11), when demand increases for the product of a monopoly or monopolistic competitor, its marginal revenue curve shifts upward (and rightward), and its profit-maximizing output level rises. Once again, because firms are producing more, they will want to use more labor at any given wage rate.

FIGURE 4 A Rightward Shift in the Labor Demand Curve



When firms in a labor market want to employ more labor at any given wage rate, the demand curve shifts rightward. Here, the number of nurses firms want to employ at a wage of $25 per hour rises from 1 million to 2 million. The rightward shift could be caused by an increased demand for a product that nurses help produce (such as home health care services), a new or cheaper complimentary input, or a rise in the price of a substitutable input.

When a technological advance creates an entirely new complementary input, or when an existing one becomes cheaper, the demand curve for labor that uses the input will shift rightward.

For example, sophisticated diagnostic equipment used by nurses is complimentary with their labor, because it enables them to care for more patients each day. When a drop in price causes firms to acquire more equipment, or when a new technology is put to use at the firm, nurses will be more productive and firms will want to hire more of them at each wage. The demand curve for nurses will shift rightward.

A **substitutable input** is one that can be used *instead* of a particular type of labor. We discussed an example of this type of input (other health care workers) earlier, as part of the input-substitution effect of a wage change for nurses. That effect moved us *along* a labor demand curve. But substitutable inputs can also cause the demand curve to *shift*.

Substitutable input An input that can be used *instead of* a particular type of labor.

When a technological change creates an entirely new substitutable input, or when an existing one becomes cheaper, the demand curve for the type of labor it replaces will shift leftward.

If, for example, the wage rate for physician assistants decreases, the least-cost combination of inputs for any level of production would change: more physician assistants, fewer nurses. This would shift the demand curve for nurses leftward.

Labor Supply

So far, we've considered the demand side of the labor market: firms that hire workers. Now we turn our attention to the *supply* side of the labor market: households that supply their labor to firms.

Variable Hours versus Fixed Hours

Some people have considerable freedom to vary how many hours they work. For example, many self-employed professionals—doctors, lawyers, freelance writers, and others—can adjust their work hours as they please, simply by increasing or decreasing the number of clients they serve. Even employees can sometimes vary their weekly hours by changing from full-time to part-time work or vice versa, or by accepting or refusing overtime. In cases like these—in which work hours can be varied—economists analyze the labor supply decision using a model of individual choice very similar to the one you learned for consumer theory in Chapter 6. However, instead of a limited budget that must be allocated to two different *goods,* the individual has a limited number of *hours* to allocate to two *activities*: working (earning income) and leisure.

In most labor markets, however, there is relatively little freedom to vary your weekly hours. Instead, your employer will expect you to work a pre-determined shift: typically eight hours a day, five days a week. Your choice is not how many *hours* you want to work, but rather *whether to work at all* in that labor market. In this chapter, we'll focus on *fixed-hours* labor markets like these, because they are so common.

The Labor Supply Curve

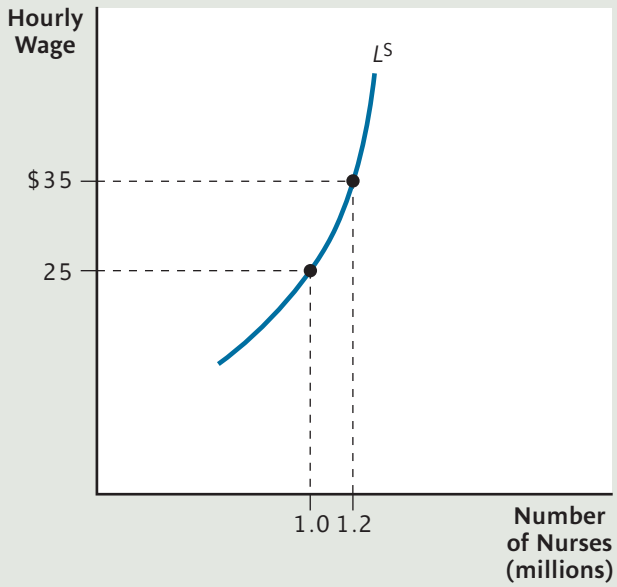
Labor supply curve A curve indicating the number of people who want jobs in a labor market at each wage rate.

Figure 5 shows a hypothetical **labor supply curve** (L^S) for registered nurses in the United States. It tells us the *number of people* who will want to work as nurses (supply their labor) at each hypothetical wage rate.

As you can see, the labor supply curve slopes upward: All else equal, a rise in the wage rate (from $25 to $35 in the figure) causes the quantity of labor supplied to rise (from 1.0 million to 1.2 million). Where would those 200,000 additional nurses come from? We'll suppose that the labor supply curve in the figure is drawn for a

FIGURE 5 The Labor Supply Curve

The labor supply curve in a market slopes upward because, all else equal (including the number of people qualified for that job), more people will want to work in that market when the wage rate rises. For the hypothetical labor supply curve shown here, 1 million qualified nurses want to work in the nursing market if the wage rate is $25 per hour, while 1.2 million want to work as nurses if the wage rate is $35.



short-run time horizon: It traces out changes in quantity supplied that would occur a few months after a change in the wage rate. This is not enough time for new workers—attracted by the higher wage—to enroll in nursing school and qualify to supply labor in this market. So the additional nurses would have to be those who already *are* qualified, but—before the wage increased—chose not to work as nurses. They might be retired, or temporarily not working, or working at some entirely different type of job. At a higher wage rate, some of them will decide it is worthwhile to re-enter this labor market and work as nurses.

The (short-run) supply of labor curve slopes upward. A rise in the wage rate causes some additional people—already qualified but previously not working in that labor market—to want to work there.

Shifts in the Labor Supply Curve

A change in the wage rate moves us *along* a labor supply curve. But *other* changes that affect labor supply decisions will shift the entire curve. Here we'll discuss two of them.

Changes in the Number of Qualified People

In Chapter 9, we derived the (short-run) market supply curve in a product market, along which the number of firms is held constant. Over time, if new firms enter the product market, the supply curve shifts rightward. Similarly, when we draw a short-run supply curve in a *labor* market, we assume that something is being held constant: The number of people *qualified* to work in that market. But if the number of qualified people rises, the labor supply curve will shift rightward.

In the market for nurses, for example, the qualifications include having a degree from an accredited nursing program. As we move *along* the supply curve, we assume the total number of people with nursing degrees remains constant. As the wage rate changes, they can either choose to work as nurses or not. But over time, if more people acquire nursing degrees, more will want to work as nurses at any *given* wage. This is represented by a rightward shift in the labor supply curve for nurses, as illustrated in Figure 6.

Changes in *Other* Labor Markets

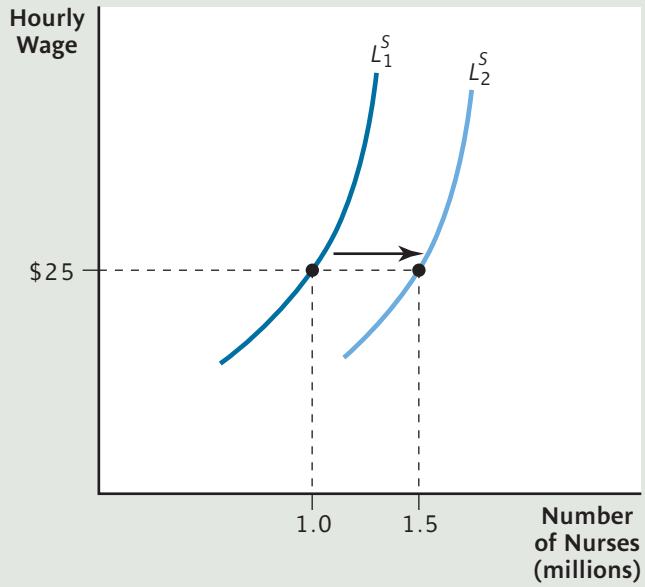
As long as some individuals can choose to supply their labor in more than one market, a change in the attractiveness of *other* jobs can shift the labor supply curve in the market we're analyzing. For example, suppose some of those with nursing degrees can instead teach science in a private high school, tutor other nursing students, or even open up a restaurant. If the wage rate falls or work in these other jobs becomes less attractive for other reasons, then nursing becomes relatively *more* attractive at any given wage rate. The labor supply curve for nurses shifts rightward.

Changes in Tastes

Some people like working with numbers and hate working with people; others prefer the reverse. Some like danger and excitement, whereas others like safety and routine. A change in tastes in favor of particular kinds of work or working conditions will shift labor supply curves rightward in those labor markets. When a type of work falls into disfavor, the labor supply curve shifts leftward.

FIGURE 6 A Rightward Shift in the Labor Supply Curve

When more people want to work in a labor market at any given wage rate, the supply curve shifts rightward. Here, the number of people who want to work as nurses at a wage of $25 per hour rises from 1 million to 1.5 million. The rightward shift could be caused by an increase in the number of qualified nurses over time, a decrease in the wage rate or attractiveness of jobs in alternative labor markets, or a change in tastes in favor of nursing work.



Tastes can also change for working *in general*. In the United States, women's labor force participation rate rose dramatically during the 1960s through the 1990s. The growth was especially rapid among married women: In 1960, less than one-third of married women were in the labor force; by 2000, almost two-thirds were. This change in the desire to work shifted labor supply curves rightward in many labor markets simultaneously.

Labor Market Equilibrium

Figure 7 shows demand and supply curves in the market for nurses. By now you can certainly guess that the market equilibrium occurs at point *A*, with a wage rate of $25 per hour, and employment of 1 million nurses. But on your own, it's worthwhile to review the logic of equilibrium in this new context. Make sure you can see that at any wage rate above $25, an excess *supply* of nurses would ultimately drive the market wage downward, as nurses competed to get available jobs. Similarly, at any wage rate lower than $25, an excess *demand* for labor would drive the wage rate upward, as firms competed to hire scarce nurses.

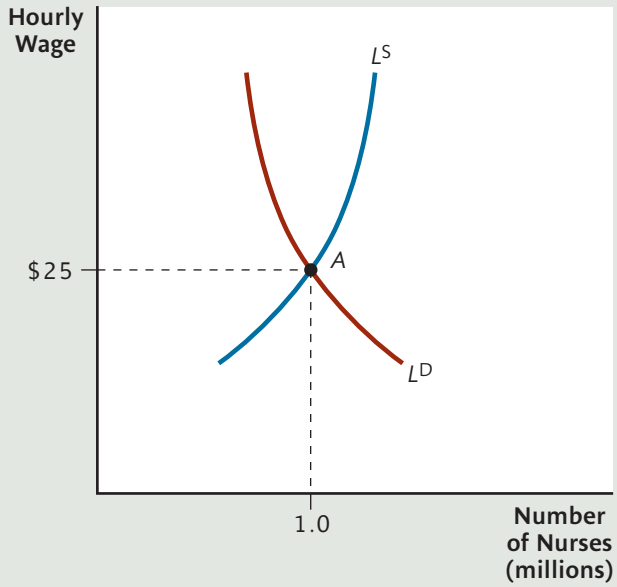
What Happens when Things Change

Events that shift a labor demand or labor supply curve will change both the equilibrium wage rate and employment level. Here we'll explore two examples.

An Increase in Labor Demand

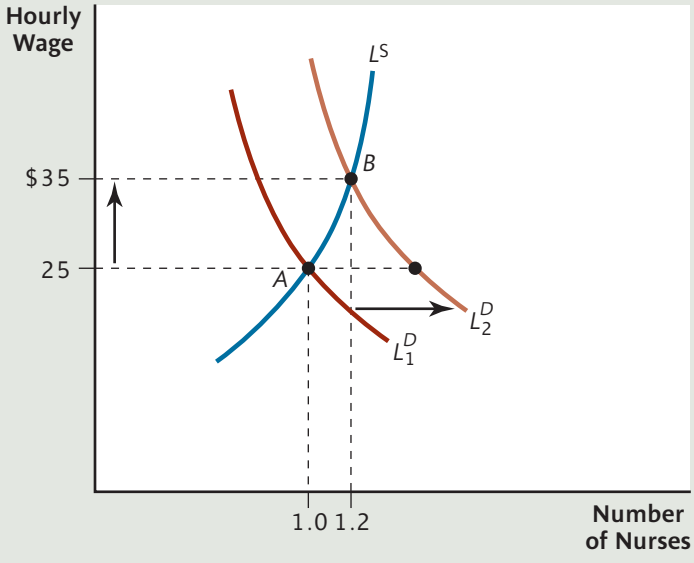
Figure 8 shows what happens when the labor demand curve for nurses shifts rightward. This might be caused, for example, by an increase in demand in the *product market* for health care services. Indeed, in recent years, the demand for health care

FIGURE 7 Equilibrium in a Labor Market



The equilibrium wage rate and employment level in a labor market firms are found where the labor supply and labor demand curves intersect. Here, the equilibrium wage rate for nurses is $25 per hour, and equilibrium employment is 1 million. If the wage rate were greater than $25, the quantity of nurses supplied would exceed the quantity demanded, and the wage rate would fall. If the wage rate were less than $25, the quantity demanded would exceed the quantity supplied, and the wage rate would rise.

FIGURE 8 An Increase in Labor Demand



Initially, the market for nurses is in equilibrium at point A, *with an hourly wage of $25 and 1 million people working as nurses. When labor demand increases (the labor demand curve shifts rightward), the equilibrium moves to point* B. *The wage rate rises to $35, and employment rises to 1.2 million.*

services has increased dramatically, due to an aging population, the availability of new treatments, and rising incomes.

In the figure, the demand curve shifts from L_1^D to L_2^D. At the old wage rate of $25, there is now an excess demand for nurses, causing the wage rate to rise. As the wage

rate increases, the quantity of nurses demanded falls, moving us leftward along L_2^D. At the same time, the rising wage rate attracts qualified nurses back into the labor market, moving us *rightward* along the short-run labor supply curve L^S. In the figure, once the wage rate reaches \$35, the quantity of nurses demanded and supplied are equal at the new, higher employment level of 1.2 million.

An increase in labor demand raises both the equilibrium wage rate and the equilibrium level of employment.[3]

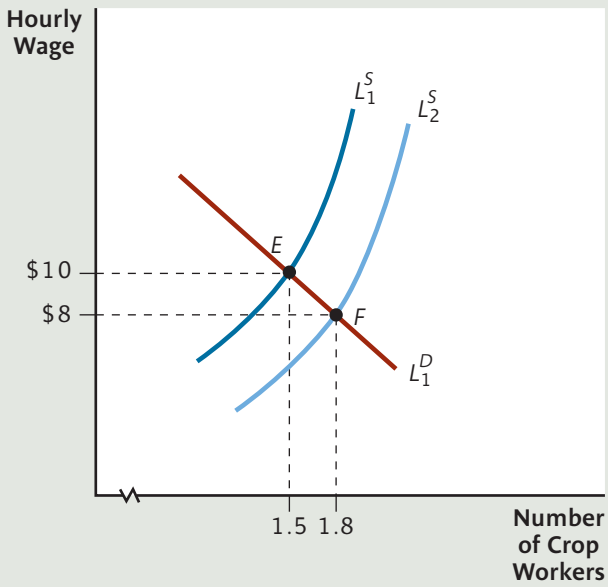
Note that now that the wage is higher, more people might decide to become nurses. As we move into the long run, this increase in the number of qualified people would start to shift the supply curve rightward as well. But unless something else changes, the supply curve should not shift by so much that the equilibrium wage rate falls all the way back to \$25 in the long run. After all, it is only the *increase* in the wage rate *above* \$25 that causes more people to want to qualify.

An Increase in Labor Supply

In Figure 9, we shift gears to another labor market: agricultural crop workers in the United States, most of whom are immigrants. In 2008 and 2009, the labor supply curve shifted rightward, due to events in *another* labor market. Recall (from Chapter 4) that during this period, the housing bubble had burst. Home prices were rapidly declining, and wages and job opportunities in housing construction declined. This made farm work *relatively* more attractive (compared to construction work).

FIGURE 9 An Increase in Labor Supply

Initially, the market for crop workers is in equilibrium at point E, *with an hourly wage of \$10 and 1.5 million people working as crop workers. When labor supply increases (the labor supply curve shifts rightward), the equilibrium moves to point* F. *The wage rate falls to \$8, and employment rises to 1.8 million*



[3] In Figure 8, after the wage rate rises, more people might decide to become nurses. As we move into the long run, this increase in the number of qualified people would shift the supply curve rightward and cause the wage rate to drop from \$35, its short-run equilibrium value. Where the wage rate ends up in the long run depends on the pay at which a *permanently* greater number of people will work as nurses.

TABLE 1

Hourly Earnings of Full-time Workers in Selected Occupations, 2007

Occupation	10th Percentile	Median (50th Percentile)	90th Percentile
Airline pilots and navigators	$48.99	**$122.95**	$ 172.89
Physicians, General Practice	$29.33	**$ 79.33**	$ 225.11
College teachers, business	$23.77	**$ 60.26**	$ 92.16
Pharmacists	$38.14	**$ 49.27**	$ 54.24
Marketing Managers	$27.64	**$ 46.47**	$ 79.33
College teachers, political science	$30.45	**$ 36.38**	$ 56.86
Registered nurses	$21.42	**$ 28.99**	$ 41.57
Electric installers and repairers	$17.37	**$ 27.80**	$ 35.27
Truck drivers, heavy and tractor-trailer	$11.80	**$ 16.50**	$ 24.61
Preschool teachers	$ 8.50	**$ 13.25**	$ 28.36
Cabinetmakers	$ 9.99	**$ 13.00**	$ 17.50
Bank tellers	$ 9.00	**$ 11.02**	$ 14.85
Cooks, short order	$ 7.00	**$ 8.58**	$ 12.93
Cashiers	$ 6.70	**$ 8.37**	$ 12.50

Source: Selected data from *National Compensation Survey: Occupational Wages in the United States, August 2008*, Table 15 Bureau of Labor Statistics. The BLS data are based on a sample survey of about 22,000 business establishments and government agencies.

Because crop workers could do either job and were relatively mobile, they poured into the agricultural job market.[4] In the figure, the labor supply curve shifts rightward, from L_1^S to L_2^S. The equilibrium wage rate declines from $10 to $8, while equilibrium employment rises from 1.5 million to 1.8 million.

Why Do Wages Differ?

Table 1 shows hourly earnings in 2007 for full-time workers in selected occupations. Each row of the table lists not only the median wage rate (in bold), but also the wage rate at the 10th and 90th percentiles for the occupation. For example, the second row shows that 10 percent of full-time physicians in general practice earned $29.33 per hour or less, while 90 percent earned $225.11 or less (so the top 10 percent earned $225.11 or more). And the bolded middle column tells us that in 2007, half of physicians earned less than $79.33 per hour while the other half earned more.

Note the inequality in wage rates among *different* occupations. These sharp differences occur even for jobs in the same industry, in which the work is often similar. Compare, for example, the median hourly wage rate of a business professor ($60.26) and a political science professor ($36.38), or that of a physician ($79.33) and a registered nurse ($28.99).

[4] "Immigrants Turn to Farm Work amid Building Bust," *Wall Street Journal,* June 13, 2008, p. A4.

But you can also see sharp differences in earnings *within* many occupations. For many of these jobs, the wage rate at the 90th percentile is many times greater than at the 10th.

This wage inequality—among and within occupations in the U.S. labor market—is *persistent*. The highest-paid and lowest-paid occupations have been so for decades. And year after year, the highest-paid workers within an occupation earn substantially more than the lowest.

Moreover, Table 1 understates differences in pay. It does not include bonuses, fringe benefits, or other additional labor earnings that are substantially greater for the highest-paying occupations and the highest-paid workers within each job. It also leaves out those at the very top—such as chief executive officers of top corporations, sports celebrities, and movie stars. For example, Eddie Murphy's wage rate on most films is between $10,000 and $20,000 per hour. (He earned a higher rate for *Shrek 2,* as you can calculate on your own). And the top 25 hedge fund managers averaged more than $100,000 per *hour* (that is not a typo) in 2008.[5]

How can one hour of human labor have such different values in the market?

An Imaginary World

To understand why wages differ in the real world, let's start by imagining an *unreal* world with three features:

1. All labor markets are perfectly competitive.
2. Except for differences in wages, all jobs are equally attractive to all workers.
3. In the long run, all workers can costlessly acquire the qualifications for any job.

In such a world, we would expect every worker to earn an *identical* wage in the long run. Let's see why.

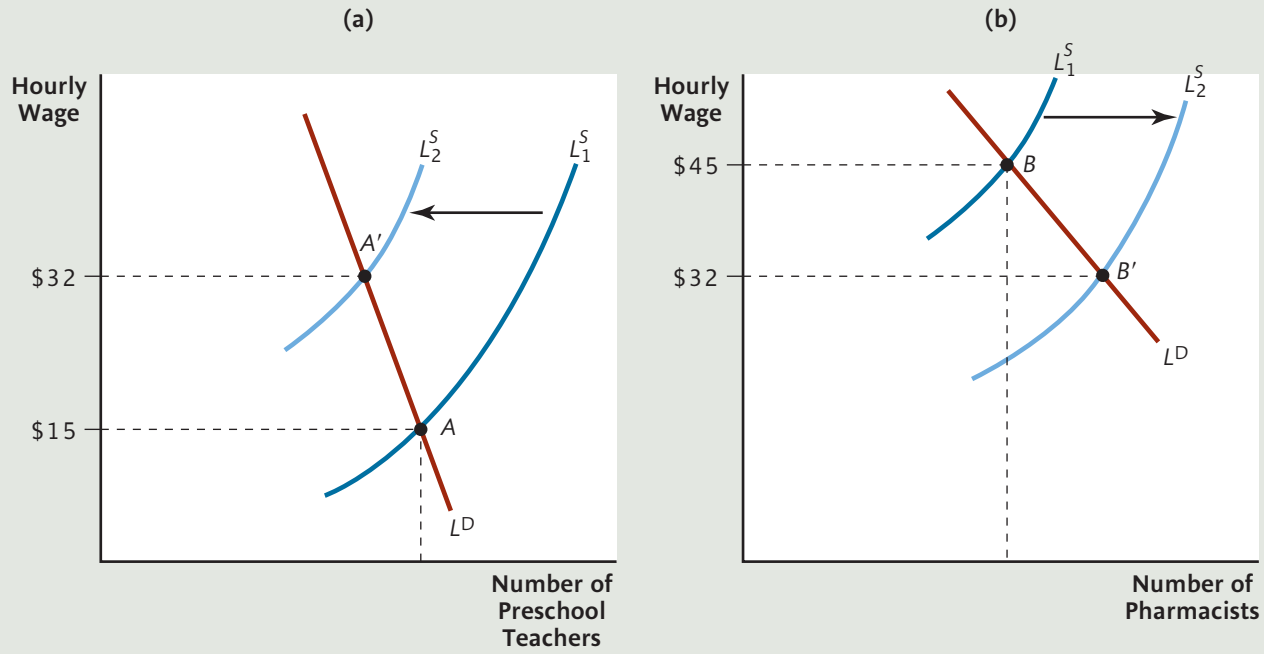
Figure 10 shows two different labor markets that initially have different wage rates. Panel (a) shows a local market for preschool teachers, with an initial equilibrium at point *A* and a wage of $15 per hour. Panel (b) shows the market for pharmacists, who, at point *B,* earn $45 per hour. In the figure, the labor supply curves are for a short-run period, during which the number qualified for each job is held constant. In our imaginary world, could this diagram describe the *long-run* equilibrium in these markets?

Absolutely not. Imagine that you are a preschool teacher. By our second assumption (all jobs are equally attractive), you would find being a pharmacist just as attractive as teaching school. But because pharmacists earn more, you would prefer to be one. By our third assumption, you can costlessly *qualify* to be a pharmacist. And by our first assumption (perfect competition), there are no barriers to prevent you from becoming one. Thus, you—and many of your fellow preschool teachers—will begin to obtain the qualifications to be pharmacists, and eventually seek work in that market. In panel (a) the labor supply curve will shift leftward (exit from the market for preschool teachers), and in panel (b) the labor supply curve will shift rightward (entry into the market for pharmacists). As these shifts occur, the market wage rate of preschool teachers will rise and that of pharmacists will fall.

When will the entry and exit stop? When there is no longer any reason for a preschool teacher to want to be a pharmacist—that is, when both labor markets are paying the same wage rate ($32 in our example). In the long run, the market for preschool school teachers reaches equilibrium at point A' and the market for pharmacists at point B'.

[5] David Walker, "$2.5 Billion in Pay Makes Simons Hedge Fund World's Top Earner," *Wall Street Journal,* March 25, 2009 and author's calculations, assuming a workweek of 80 hours. Assuming a more normal workweek of 40 hours, the wage rate would be $200,000 per hour.

FIGURE 10 **Disappearing Wage Differentials**



Initially, the supply and demand for preschool teachers in panel (a) determine an equilibrium wage of $15 per hour—at point A. In panel (b), the equilibrium wage for pharmacists is initially $45 per hour. If these markets are competitive, if the two jobs are equally attractive, and if all workers can costlessly acquire the qualifications to do either job, this wage differential cannot persist. Some preschool teachers will give up that occupation, reducing supply in panel (a), and become pharmacists, increasing supply in panel (b). This migration will continue until the wage in both markets is equal—at $32 in the figure.

Our conclusion about preschool teachers and pharmacists would apply to *any* pair of labor markets we might choose. In our imaginary world, bank tellers and physicians, kitchen workers and nurses—all would earn the same wage. In this world, labor is like water in a swimming pool: It flows freely from end to end ensuring that the level is the same everywhere. In our imaginary world, workers flow into labor markets with higher wages, evening out the wages in different jobs . . . *if* our three critical assumptions are satisfied.

But take any one of these assumptions away, and the equal-wage result disappears. This tells us where to look for the sources of wage inequality in the real world: a *violation* of one or more of our three assumptions.

Compensating Differentials

In our imaginary world, all jobs were equally attractive to all workers. But in the real world, jobs differ in hundreds of ways that matter to workers. When one job is more or less attractive than another for reasons other than a difference in pay, we can expect their wage rates to differ by a *compensating wage differential:*

*A **compensating wage differential** is the difference in wage rates that makes two jobs equally attractive to workers.*

Compensating wage differential A difference in wages that makes two jobs equally attractive to a worker.

To see how compensating wage differentials come about, let's consider some of the important ways in which jobs can differ.

Nonmonetary Job Characteristics

Suppose you are an office worker in a skyscraper, and you could earn $1 more per hour than you currently earn by washing the building's windows . . . from the *outside*. Would you "flow" to the windowwasher labor market, like water in a pool? Probably not. The harder work and greater risk of death just wouldn't be worth the one dollar-per-hour difference.

Nonwage job characteristic
Any aspect of a job—other than the wage—that matters to a potential or current employee.

Danger is an example of a **nonwage job characteristic.** It is an aspect of a job (good or bad) other than the wage rate that people care about. When you think about a career, whether you are aware of it or not, you are evaluating hundreds of nonwage job characteristics: the risk of death or injury, the amount of physical exertion required, the degree of intellectual stimulation, the potential for advancement, the cost of living and other conditions, where the job is located . . . the list goes on and on.

Differences in nonwage job characteristics alone can stop the flow of workers from one labor market to another before the wages are equalized. If this were the only one of our special features that were violated, we would expect the compensating differential to be just enough to make the two jobs equally attractive.

To take an extreme example, suppose people had identical tastes for work, and that *everyone* would prefer to teach preschoolers rather than work in a pharmacy. Further, suppose these preferences were so strong that it would take a wage differential of $30 per hour to make the two jobs equally attractive. Then the long-run equilibrium would remain at the initial points *A* and *B,* with pharmacists earning a compensating wage differential of $45 − $15 = $30 per hour to make up for the less desirable features of their job. Even with the higher wage rate for pharmacists, preschool teachers would not want to switch jobs.

The nonmonetary characteristics of different jobs give rise to compensating wage differentials. Jobs considered intrinsically less attractive will tend to pay higher wages, other things being equal.

Of course, different people have different tastes for working and living conditions. When labor markets are perfectly competitive, the entry and exit of workers automatically determines the compensating wage differential in each labor market.

What about *unusually attractive* jobs? Compared to the average job, these will generally pay *negative* compensating differentials. For example, many new college graduates are attracted to careers in the arts or the media. Since entry-level jobs in these industries are so desirable to so many people, they tend, on average, to pay lower wages than similar jobs in other industries.

Human Capital Requirements

In our imaginary world, everyone could costlessly acquire the qualifications to do any job. In the real world, many jobs require costly investments in human capital. All else equal, jobs that require more costly investments must pay more in the long run, or people wouldn't want to do them. (Although human capital may be a sunk

cost for those who already have it, it is *not* sunk for those thinking about entering a profession to replace those retiring from it.)

Differences in human capital requirements give rise to compensating differentials. Jobs that require more costly training will tend to pay higher wages.

For example, a pharmacist must have a PharmD degree, which requires at least six years of schooling after high school, including prerequisite science courses. In most states, substantially less schooling is required to qualify as a preschool teacher. If these two jobs paid the same wage, and everyone felt they were otherwise equally attractive, no one would want to become a pharmacist. The number qualified to be pharmacists would decrease (through retirement), and the labor supply curve would shift leftward. This would cause the market wage for pharmacists to rise—until the difference in wages compensated pharmacists for the additional costs of qualifying.

Compensating wage differentials explain some of the wage differences we observe between jobs. The relatively high earnings of doctors, attorneys, research scientists, and college professors reflect—at least in part—compensating differentials for the especially high costs of qualifying for their professions.

The idea of compensating wage differentials dates back to Adam Smith, who first observed that unpleasant jobs seem to pay more than other jobs that require similar skills and qualifications. It is a powerful concept, and it can explain many of the differences we observe in wages . . . but not all of them.

Differences in Ability

In our discussion of compensating wage differentials and human capital requirements, we assumed that anyone willing to pay for human capital can acquire it. We've also been assuming that labor is standardized (one of the assumptions of a competitive labor market). So once someone *becomes* qualified, they can do the job just as well as anyone else.

But in the real world, abilities differ. And these differences in ability can create wage differences in two ways.

Differences in Ability to *Become* Qualified

Not everyone has the intelligence needed to be an astrophysicist, the steady hand to be a neurosurgeon, the quick-thinking ability to be a commodities trader, the well-organized mind to be a business manager, or the talent to be an artist or a ballet dancer. If you don't have the intrinsic talent or ability required in one of these jobs, you will not qualify even if you were to invest in the training.

Going back to Figure 10, suppose a wage differential of $10 per hour would compensate for the higher human capital investment required of the job in the right panel, and that initially we are at points *A* and *B* (with a differential much greater than $10). But now suppose that those working for lower wages in the left panel cannot switch to the high wage job on the right because—regardless of training or effort—they cannot master the skills required for that job. Then the "flow of labor" from one market to the other, which would ordinarily reduce the differential to $10, will not occur. The result is a persistent wage differential that *exceeds* any compensating wage differential.

This leads to an important conclusion:

All else equal, jobs that require skills that relatively few people have the ability to acquire will pay persistently higher wage rates—in excess of compensating wage differentials.

Notice the word *relatively* in the highlighted statement. A wage rate is determined by both supply and demand. In order for your rare ability to acquire some skill to result in a higher wage, there must also be sufficient *demand* for the skill.

To take an extreme example, imagine that only twenty people on earth had the intrinsic ability, with proper training, to whistle out of their ears. But suppose that there is virtually no demand for employees or performers who develop that skill. Then, while the absolute number of people had it would be very few, the *relative* number who had it (relative to the demand) would be very many. So the ability, while rare, would not give anyone who exploits it a higher wage.

On the other hand, hundreds of thousands of people have the intrinsic ability to complete law school, medical school, or business school. But because the demand for these professionals is high *relative* to the number who are able to qualify, their wage remains *greater* than that needed to compensate for the cost of the training.

Differences in Ability *among* Those Qualified

Even among those qualified and working at a particular job, some are more able than others. This violates the "standardized labor" assumption of a perfectly competitive labor market. And it helps to explain much of the difference in pay we see *within* occupations in Table 1.

For example, the top 10 percent of marketing managers earn almost twice as much as the median, and almost three times as much as the bottom ten percent. Some of this difference may be due to differences in nonwage job characteristics (marketing managers work under a variety of different conditions). But much of it is based on differences in ability. If Alicia can design a better marketing campaign than Emily, and if every employer knows this, then hiring Alicia will earn a firm more revenue than hiring Emily. Thus, in an otherwise competitive market, firms will be willing to pay a greater wage to Alicia than to Emily.

All else equal, those with greater ability to perform a job better—based on talent, experience, motivation, or perseverance—will be more valuable to their employers, and will generally be able to command a higher wage rate.

Differences in ability also help explain why workers' pay tends to rise with age and experience on the job. Experience adds to a worker's human capital and can also reduce risk to employers. Hiring a new, untested worker—even one who seems to have great talent—is always a bit risky, because the worker's basic ability hasn't yet been proven. By contrast, hiring or continuing to employ someone with a history of advancement and accomplishments reduces this risk.

For these reasons, firms will typically pay more for a worker with a proven track record. Not surprisingly, when wage rates within an occupation are broken down by age (not shown in Table 1), workers who are older—and have been in their occupation longer—dominate the higher percentiles, while younger and newer entrants are more prevalent in the lower percentiles.

The Economics of Superstars

In 2000, at the age of 26, Alex Rodriguez signed a contract to play baseball for the Texas Rangers at an average salary of $25 million per year. (As of 2009, he is under contract with the New York Yankees, for an average annual salary of $27.5 million). No one would argue that this unusually high pay is a compensating wage differential for the "unpleasantness" of playing professional baseball, or that—at age 26—Rodriguez had spent more years honing his skills than the average attorney, doctor, or engineer. Rather, let's state the obvious: Rodriguez is an *outstanding* baseball player, better than 99.999 percent of the population could ever hope to be. This is largely due to Rodriguez's exceptional gifts: both physical (agility, coordination, strength) and emotional (the temperament and character needed to hone and exploit those gifts).

Alex Rodriguez is an example of a *superstar*—an individual widely viewed as among the top few in his or her profession. In recent years, superstars have included actors such as Will Smith, Julia Roberts, and Angelina Jolie; talk show hosts Jay Leno and David Letterman; news anchors Brian Williams and Katie Couric; and novelists Stephen King and J. K. Rowling.

Exceptional ability plays a role here. But when we try to explain the extremely high wage rates of these superstars based solely on ability, we confront a puzzle. Clearly Alex Rodriguez has more athletic ability and more skill in honing it, than almost anyone in the population, including other major-league baseball players. But can this explain a salary that is *25 times* that of the median major-league player? By any measure, is Rodriguez *25 times better* than these other players?

© DREAMWORKS/THE KOBAL COLLECTION/ PICTURE DESK

Many would agree that Eddie Murphy's voice work for *Shrek 2* was better than, say, the typical voiceovers on the Saturday morning cartoon shows. But Murphy's hourly pay—at about $550,000 per hour—was about 3,000 times that of the typical cartoon voice actor. Is he *that much better?*

The very top athletes, writers, rock stars, comedians, talk show hosts, and movie directors all earn wage premiums that seem vastly out of proportion to their additional abilities. Why? The explanation in all these cases *is* based on ability—and also the exaggerated rewards that certain markets bestow on those deemed the best or one of the best in their field.[6]

An example can help us understand how this works. Say you like to read one mystery novel a month for entertainment. If you can choose between the best novel published that month or one that is almost—but not quite—as good, you will naturally choose the one you think is best. Only people who read *two* mystery novels each month would choose the best *and* the second best; only those who read three will choose the top three. If most people rank recent mystery novels in the same order, then the best will sell millions of copies, the second best might sell hundreds of thousands, and the third best might sell only thousands.

Even though all three novels might be very close in quality, a publisher will earn *10 times* more revenue selling the best novel compared to the second best, and 10 times more revenue selling the second best compared to the third best, and so on. Accordingly, a publisher will be willing to pay the same multiples in advances and royalties when bidding for contracts with mystery novelists of different rankings. Even if the top

[6] The seminal article on this theory is Sherwin Rosen, "The Economics of Superstars," *American Economic Review,* American Economic Association, Vol. 71, No. 5, 1981, pp. 845–58. A less-technical version is Sherwin Rosen, "The Economics of Superstars," *The American Scholar,* Vol. 52, No. 4, 1983. See also Robert H. Frank and Philip J. Cook, *The Winner Take All Society* (New York: The Free Press, 1995).

author is viewed as only *slightly* better than the next one down, as long as the vast majority of readers agree on the ranking, she can end up earning 10 times as much.

The same thing happens in markets for athletes, rock stars, movie stars, and news broadcasters.

> *In labor markets for talented professionals, in which there is mass-distribution of their product and substantial agreement about rankings, small differences in ability can lead to disproportionate differences in pay.*

The owners of the Texas Rangers and the Yankees were willing to pay Alex Rodriguez $25 million or so each year because they believed that Rodriguez, as a superstar, would bring in *at least* that much additional revenue each year—from ticket sales, skybox rentals, TV and radio broadcasting fees, concession sales, parking fees, and more.

Beyond Superstars

The logic behind superstar pay applies beyond sports and entertainment markets. Suppose you had a net worth of $100 million, and you were being sued for all of it. And suppose the *best* attorney you could hire had a 90 percent chance of winning your case. The *second best,* who was very close in ability, had an 88 percent chance of success. How much more would you pay to hire the best one? And if the best isn't available, what premium would you pay for the second best, if the third-best has an 85 percent chance of success? With a little thought, you can apply the "economics of superstars" to many types of professions: physicians, political campaign strategists, asset managers, and corporate managers.[7] In all of these cases, the stakes are very high and small differences in perceived ability can lead to huge differences in outcome.


© MASTERFILE

Barriers to Entry

In our imaginary world, there are no artificial barriers to entering any trade or profession. But in the real world, barriers keep would-be entrants out of some labor markets, enabling those already there to continue earning higher wages than otherwise.

Occupational Licensing

In many labor markets, occupational licensing laws keep out potential entrants. Highly paid professionals such as doctors, lawyers, and dentists, as well as those who practice a trade, like barbers, beauticians, and plumbers, cannot legally sell their services without first obtaining a license. In many states, you cannot even sell the service of braiding hair without a license. In order to get the license, you must complete a course in cosmetology and pass an exam. Typically, a licensing board is comprised of people already in the occupation, who are sometimes exempt from new educational requirements.

These requirements raise the cost of acquiring human capital, and stop the flow of new workers into a profession before the wage can fall

[7] A long-standing social controversy over the pay of executives, asset managers and traders—especially at financial firms—intensified during the financial crisis of 2008–2009. A variety of forces have determined these wages, including what economists call "market failures." We'll revisit this issue when we discuss market failures in Chapter 15.

TABLE 2

Examples of Occupational License Requirements

Occupation	State(s)
Athletic Trainer	Most
Auctioneer	Several
Beekeeper	Maine
Elevator Operator	Massachusetts
Florist	Louisiana
Hair Braider	Several
Interior Designer	Several
Lobster Seller	Rhode Island
Prospector	Maine

Source: E. Frank Stephenson and Erin Wendt, "Occupational Licensing: Scant Treatment in Labor Texts," *Econ Journal Watch,* Volume 6, No. 2, May 2009, Table 3. (For a complete list of jobs requiring licenses by state, see *http: www.acinet.org/acinet/licensedoccupations*.)

to that in other, similar labor markets. Table 2 provides a few examples of occupational license requirements in various states. In almost all cases, the requirements are presented as measures to protect the public health or safety. In some cases, one can make a reasonable argument for this. In others . . . well, you be the judge.

Union Wage Setting

A labor union represents the collective interests of its members. Unions have many functions, including pressing for better and safer working conditions, operating apprenticeship programs, and administering pension programs. But a major objective of a union is to raise its members' pay.

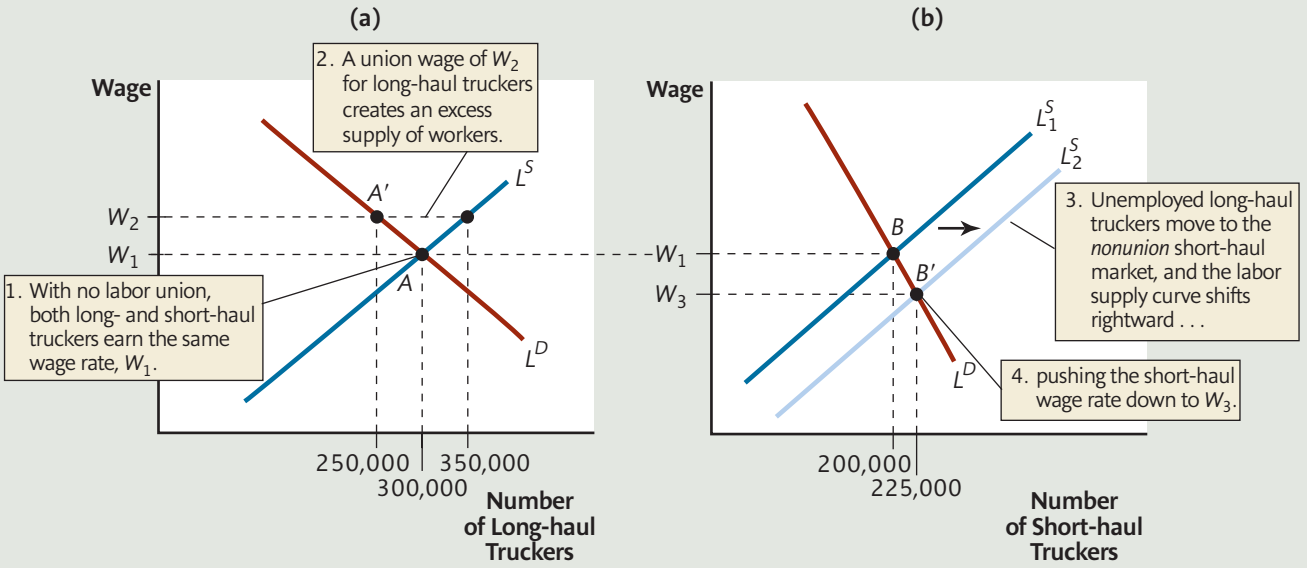
Federal law prohibits a union from creating an overt barrier to entry. Instead, the union creates a barrier *indirectly,* by using its power to negotiate a higher-than-competitive wage that the firm must pay. As we know from the last chapter, at a higher wage, the firm will have a lower profit-maximizing employment level. Thus, many potential workers are kept out of union jobs because the firm will not hire them at the union wage.

Figure 11 illustrates how unions can create wage differences. We assume that jobs in two industries—long-haul trucking and short-haul trucking—are equally attractive in all respects other than the wage rate. With no labor union, these two markets would reach equilibrium at points *A* and *B,* respectively, where both pay the same wage, W_1.

Now suppose that long-haul truckers are organized into a union that has negotiated a higher wage, W_2, with employers. In this market, employment drops from 300,000 to 250,000, (point A'), while the number who would like to work there rises to 350,000. There is an excess supply of long-haul truckers equal to 350,000 − 250,000 = 100,000. Ordinarily, an excess supply of labor drives the wage down, but the union wage agreement prevents this.

With fewer jobs available in the unionized sector, some former long-haul truckers will look for work as *nonunion,* short-haul truckers. Thus, in panel (b) the labor supply curve shifts rightward and the wage of short-haul truckers drops to W_3. The end result is a union–nonunion wage differential of $W_2 - W_3$. Notice that only *part* of the differential ($W_2 - W_1$) represents an increase in union wages; the other part ($W_1 - W_3$) comes from a decrease in *nonunion* wages.

FIGURE 11 Union Wage Differentials



In a competitive labor market, a union—by raising the wage firms must pay—decreases total employment in the union sector. This, in turn, causes wages in the nonunion sector to drop. The combined result is a wage differential between union and nonunion wages.

Several decades ago, when unions were a more powerful force in U.S. labor markets, the union wage differential was an important subject of study. But union membership in the United States has dwindled. In the mid-1950s, about 25 percent of the total U.S. labor force was unionized. Today, the comparable figure is less than 11 percent. Nevertheless, unions still maintain a significant presence in any industries, such as automobiles, steel, coal, construction, mining, trucking, and the government sector, including public education. They have certainly been responsible for at least *some* of the higher wages earned in those industries.

Of course, we've been viewing unions from one perspective only: to explain how wage differences can arise. The full effect of unions on labor markets is much more complex. For example, many of the features of modern work that we take for granted today—such as paid vacations and overtime pay—originated in union struggles with management.

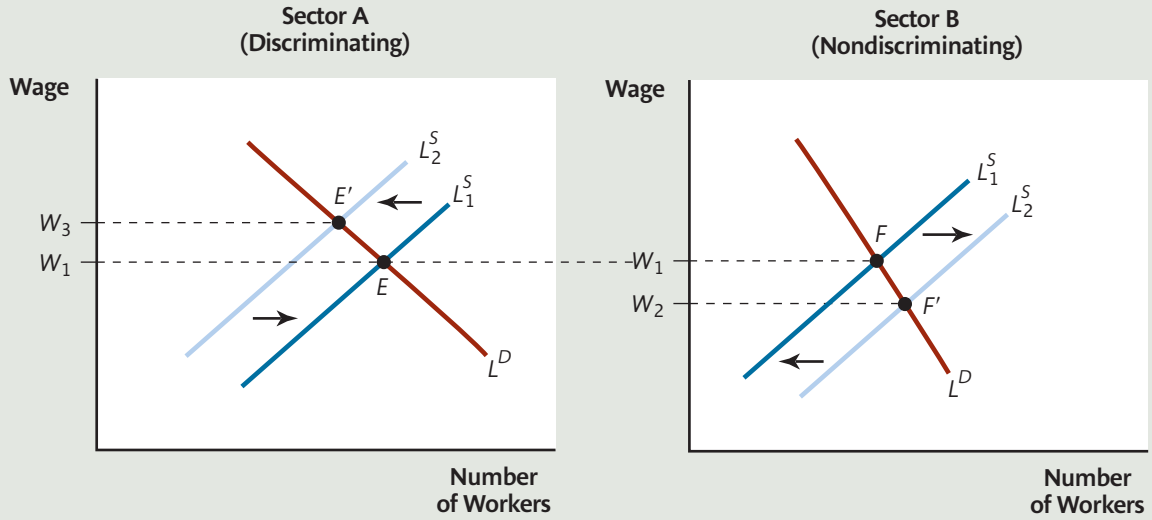
Moreover, through grievance procedures and other forms of communications with management, unions can raise worker morale, reduce labor turnover, and possibly increase worker productivity. This could actually *increase* the demand for labor by unionized firms, reducing (and possibly reversing) the drop in unionized employment caused by the higher wage.

Discrimination

Discrimination When a group of people have different opportunities because of personal characteristics that have nothing to do with their abilities.

Discrimination occurs when *the members of a group of people have different opportunities because of characteristics that have nothing to do with their abilities.* In recent U.S. history, discrimination against women and minorities has been

FIGURE 12 Employer Discrimination and Wage Rates



In the absence of discrimination, the wage rate would be W_1 in both sector A and sector B. If firms in sector A discriminate against some group—such as women—the group would seek work in the nondiscriminating sector, B. The increased labor supply in sector B causes the wage there to fall to W_2, while the decreased supply in sector A causes the wage there to rise to W_3. But only temporarily if the discrimination results from employer prejudice. As men migrate from sector B to the now higher wage sector A, the labor supply changes in both sectors are reversed. The wage returns to W_1 in both sectors.

widespread in housing, business loans, consumer services, and jobs. The last arena—jobs—is our focus here. While tough laws and government incentive programs have lessened overt job discrimination—such as the help wanted ads that asked for white males as late as the 1950s—less obvious forms of discrimination remain.

Can labor market discrimination create differences in wages? Our first step in thinking about this question is to distinguish two words that are often confused. Prejudice is an emotional dislike for members of a certain group; discrimination refers to the restricted opportunities offered to such a group leading, for example, to lower wages. As you will see, prejudice is neither necessary nor sufficient for discrimination to occur.

Employer Prejudice

When you think of job discrimination, your first image might be a manager who refuses to hire members of some group, such as African-Americans or women, because of pure prejudice. As a result, the victims of prejudice, prevented from working at high-paying jobs, must accept lower wages elsewhere. No doubt, many employers hire according to their personal prejudices. But it may surprise you to learn that economists generally consider employer prejudice one of the *least* important sources of labor market discrimination.

To see why, look at Figure 12, which shows the labor market divided into two broad sectors, A and B. To keep things simple, we'll assume that all workers have the same qualifications and that they find jobs in either sector equally attractive. Under these conditions, if there was *no* discrimination, both sectors would pay the same wage, W_1.

Now suppose the firms in sector A decide they no longer wish to employ members of some group—say, women. What would happen? Women would begin

looking for jobs in the *nondiscriminating* sector B, and the labor supply curve there would shift rightward. The equilibrium would move from F to F', decreasing the wage to W_2. At the same time, with women no longer welcome in sector A, the labor supply curve there would shift leftward, moving the market from E to E' and driving the wage up to W_3. It appears that employer discrimination would create a gender wage differential equal to $W_3 - W_2$.

But the differential would shrink over time. Why? With the wage rate in sector B now lower, *men* would exit that market and seek jobs in the higher-paying sector A. These movements would reverse the changes in labor supply, shrinking the wage differential.

If the migration of men does not *eliminate* the wage differential (because there aren't enough men), and their wage remains higher, there is still another force that works to reduce it further: competition in the product market. Because biased employers must pay higher wages to employ men, they will have higher average costs than unbiased employers. If biased firms sell their product in a competitive market, they will suffer losses and ultimately be forced to exit their industries. Over the long run, prejudiced employers should be replaced with unprejudiced ones.

And even if the product market is imperfectly competitive, firms still have their stockholders or owners to contend with. Unless *their* prejudice is so strong that they are willing to forego profit, management will be under pressure to hire qualified women rather than pay a premium to hire men. In either case,

> *When prejudice originates with employers, competitive labor markets work to discourage discrimination and reduce or eliminate any wage gap between the favored and the disfavored group.*

Employee and Customer Prejudice

What if *workers,* rather than employers, are prejudiced? Then our conclusions are very different. If, for example, a significant number of male assembly-line workers dislike supervision by women, then the nonprejudiced employer might suffer lower productivity, and higher costs, if he refused to discriminate against women. This could cause the firm to fail. The market does not help us solve this problem. In fact, it reinforces it.

The same argument applies if the prejudice originates with the firm's *customers*. For example, if many automobile owners distrust female mechanics, then an auto repair shop that hires them would lose customers and sacrifice profit. True, excluding qualified female mechanics is costly; it means paying higher wages to men and charging higher prices. But customers will be willing to *pay* higher prices, because they prefer male mechanics. Even in the long run, then, women might suffer lower wages or be excluded entirely from the auto mechanics trade.

More generally,

> *when prejudice originates with the firm's employees or its customers, market forces may encourage, rather than discourage, discrimination and can lead to a permanent wage gap between the favored and disfavored groups.*

Statistical Discrimination

Suppose you are in charge of hiring 10 new employees at your firm. Suppose that young, married women in your industry are twice as likely as men to quit their jobs within two years. Quits are very costly to your firm. Let's say that 20 people apply for the 10 positions—half are men and half women. All are equally qualified, and you have no

way of knowing which *individuals* among them are more likely to quit within two years. Who will you hire? If all you care about is maximizing profit, you will hire the men.

Statistical discrimination—so called because individuals are excluded based on the statistical probability of behavior in their group, rather than their own personal traits—is a case of discrimination without prejudice. It can lead a nonprejudiced profit-maximizing employer to discriminate against an individual member of a group, even though that particular individual might never engage in the costly behavior.

But, as some observers have pointed out, statistical discrimination can also be a cover for prejudice. For example, consider statistical discrimination against women. It may be true in some cases that young, married women are more likely to quit than men. But men may be more likely to develop alcohol and drug problems, which can lead to poor judgment and costly accidents on the job. Without prejudice, the risks associated with hiring men should be thrown into the equation. According to critics of the statistical discrimination theory, the negative behavior of the favored group (such as men in this example) is rarely considered by employers.

Discrimination and Wage Differentials

Table 3 shows median weekly earnings for different groups of full time workers in the United States. As you can see, men in each group earn more than women, and whites of both sexes earn substantially more than African Americans or Hispanics. Labor market discrimination may account for some of this difference. But figuring out how much is not easy.

Consider the black-white differential for men. In 2009, black men earned 30 percent less than white men, on average. But *some* of this difference is due to differences in education, job experience, job choice, and geographic location among whites and blacks. For example, while 19 percent of black adults have college degrees, about 30 percent of white adults do. Several studies suggest that if we limit comparisons to whites and blacks with the same educational background, geographic location, and, in some cases, the same ability (measured by a variety of different tests), 50 percent or more of this earnings difference disappears. Similar studies suggest that *most* of the male-female differential disappears when we control for other variables.

Can we conclude that discrimination is not responsible for the part of the wage differential explained by these other variables? Not really. Because some of the observed differences in education, geographic location, and ability may be the *result* of job market discrimination.

Figure 13 illustrates this vicious cycle. Let's assume that job discrimination initially causes a wage differential between equally qualified whites and blacks. With

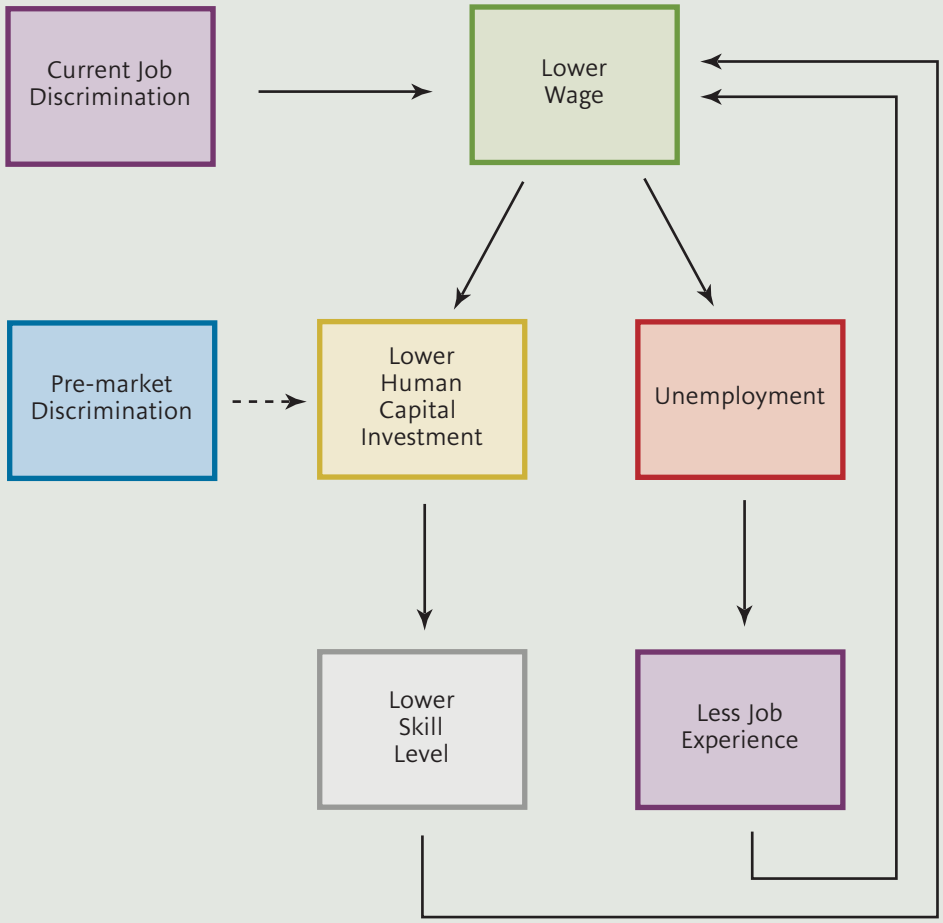
TABLE 3

Median Weekly Earnings, 2009 (Full-time Wage and Salary Workers Over Age 25)

	Median Income	Percentage of White Male Income
White Males	$855	100%
Black Males	$595	70%
Hispanic Males	$577	67%
White Females	$666	78%
Black Females	$559	65%
Hispanic Females	$510	60%

Source: Bureau of Labor Statistics News Release, *"Usual Weekly Earnings,"* April 10, 2009. Data are for first quarter of 2009. (*Note:* Persons of Hispanic origin may be of any race.)

FIGURE 13 The Vicious Cycle of Discrimination



a lower wage, blacks have less incentive to remain in the labor force or to invest in human capital, because they reap smaller rewards for these activities. The result is that blacks, on average, will have less education and less job experience than whites. At that point, even color-blind employers will hire disproportionately fewer blacks in high-paying jobs, perpetuating their lower wages.

In addition to job market discrimination, there may be *premarket* discrimination—unequal treatment in education and housing—that occurs *before* an individual enters the labor market. For example, regardless of black families' incomes, prior housing discrimination may have excluded them from neighborhoods with better public schools, resulting in fewer blacks being admitted to college. And a similar vicious cycle applies to the earnings gap between women and men.

The simple wage gap between two groups tends to overstate the impact of job-market discrimination on earnings, because it fails to account for differences in worker skill, experience, and job choice. However, controlling for these characteristics may understate the impact of discrimination, since these characteristics may in part result from discrimination.

The Minimum Wage Controversy

A policy frequently advocated to reduce wage inequality, at least at the lower end of the distribution, is an increase in the minimum wage. A minimum wage law makes it illegal to hire a worker for less than a specified wage, in any labor market covered by the law.

Many countries have minimum wage laws. In the United States, about 90 percent of the labor force is covered by the *federal* minimum wage law. Forty-four states have their own minimum wage laws, often higher than the federal minimum.

To most people, the benefits of a higher minimum wage are obvious: It increases the pay of those who earn the least, and thus helps to reduce economic inequality. But economists—even those who favor raising the minimum wage—regularly point out that the story is *not* that simple.

Who Pays for a Higher Minimum Wage?

A higher minimum wage rate raises average and marginal costs in industries that employ minimum wage labor. The result is a rise in product prices, much as occurs with an excise tax (see Chapter 4). In the short run, the higher prices enable firms to shift some, but not all, of the higher labor costs onto their customers. Thus, in the short run, the cost of the higher minimum wage is paid by both firms and consumers.

In the long run, if product markets are competitive, firms return to zero economic profit—the same situation they were in before the minimum wage was raised. But prices must remain permanently higher, to match the higher long-run average costs. Thus, the long-run burden of the minimum wage in competitive industries falls entirely on consumers.

This already suggests something peculiar about using the minimum wage for income redistribution. Instead of redistributing income from the rich to the poor, a higher minimum wage redistributes it from *customers* of minimum-wage paying industries (who may be rich or poor) to minimum wage workers . . . who may or may not be poor, as we discuss in the next section.

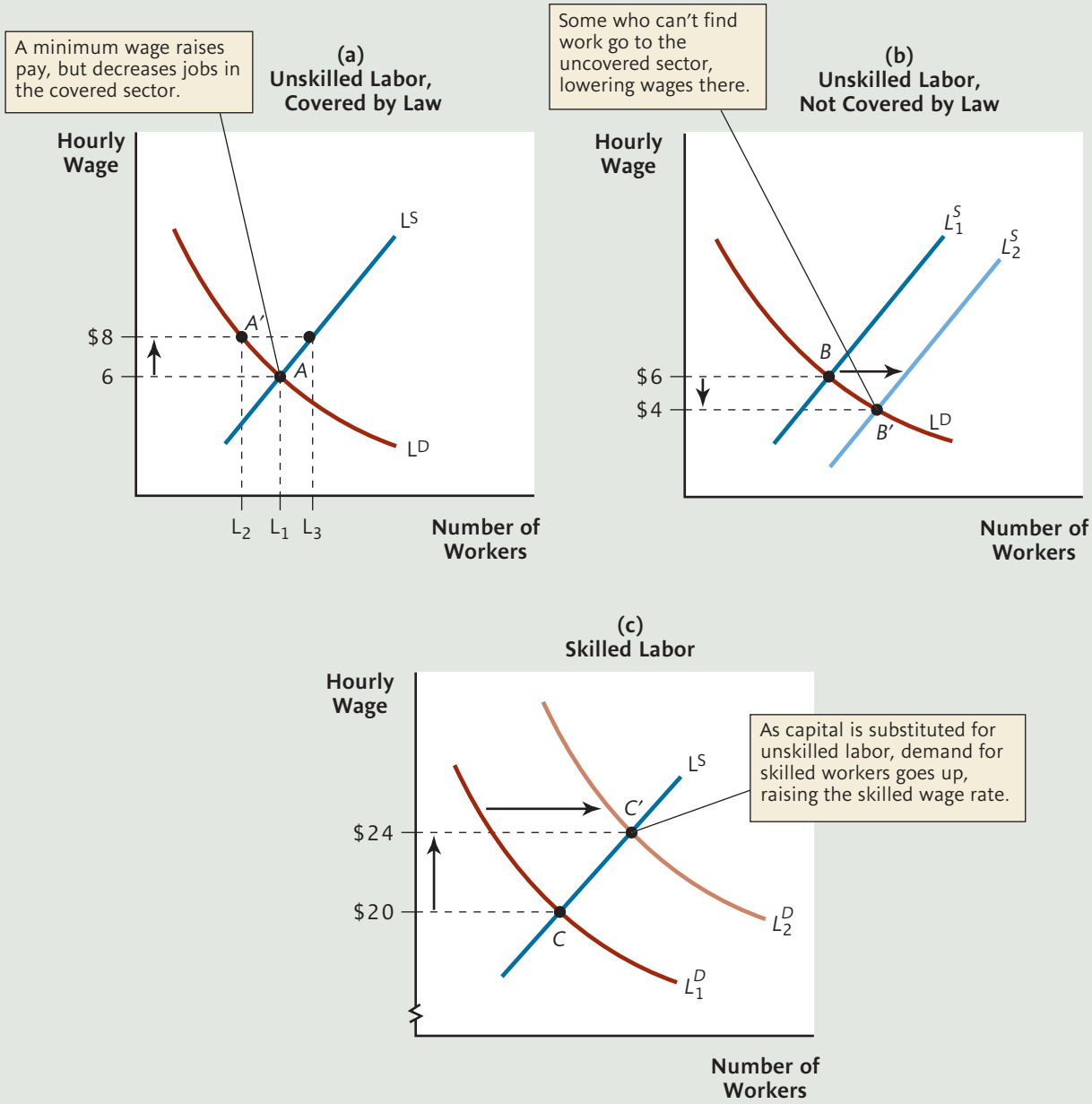
Who Benefits from a Higher Minimum Wage?

Anyone trying to support a family with a minimum wage job is certainly poor. At $7.25 per hour, a full-time minimum wage job would pay just $15,000 per year. But economists point out that a higher minimum wage is a rather blunt instrument for helping the poor because it applies to *any* worker earning the minimum, regardless of their economic situation.

According to the Bureau of Labor Statistics,[8] out of 75 million hourly workers, about 2.2 million earned at or below the minimum wage in 2008. Of these, more than a third were between the ages of 16 and 19, and more than half were under 24. Many of these young people (especially teenagers) live at home with families that run the gamut from the poor to middle class to rich. Thus, many who benefit from a higher minimum wage are economically better off than those who pay for it. This makes a higher minimum wage a rather blunt instrument for reducing inequality.

[8] "Characteristics of Minimum Wage Workers: 2008," Bureau of Labor Statistics, *www.bls.gov*.

FIGURE 14 **Minimum Wage Effects in Three Labor Markets**



Labor Market Effects of the Minimum Wage

Another problem with raising the minimum wage is that it can harm some of the very workers it is designed to help. To see why, look at Figure 14 which shows three different labor markets:

- the market for unskilled labor that is effectively *covered* by the minimum wage law (the law applies and is followed)

- the market for unskilled labor that is *not* covered (the law doesn't apply or is not effectively enforced, such as in some very small businesses)
- the market for skilled labor

Let's assume that, initially, there is *no* minimum wage and each labor market has reached its long-run equilibrium. The two unskilled labor markets in the first two panels operate at points *A* and *B*, respectively, with an hourly wage of $6. The wage is initially the same in both unskilled markets because, in the absence of a minimum wage law, workers would migrate to whichever market had the higher wage, eliminating any wage difference. In the skilled labor market in panel (c), the wage rate is considerably higher, at $20.

Now let's introduce a minimum wage of $8 for the *covered* unskilled market in panel (a). The minimum wage, just like any price floor, creates an excess supply in that market. Part of the excess supply comes from the decrease in the quantity of labor demanded (from L_1 to L_2). The other part arises from an increase in the quantity of labor supplied (from L_1 to L_3) because more people seek work in this market at the higher wage. Ordinarily, the excess supply of $L_3 - L_1$ would bring the wage rate down, but the minimum wage law prevents this. You can already see that the minimum wage benefits some low-skilled workers (those who keep their jobs and are paid more) but harms others (those who lose their jobs). The job losses in panel (a) can be especially harmful to young workers who are not college bound. A good performance at a first job—even a minimum wage job—can enable an unskilled worker to seek higher-waged employment later.

This is somewhat analogous to the way college students use unpaid internships—which are not covered by minimum wage law—to beef up their résumés and improve their future employment prospects. By reducing employment for unskilled workers, some young people will be deprived of this chance to build a job history.

But this is not the end of the story.

Some who lose their jobs in the covered sector move to the only labor market where they can find work—the uncovered market in panel (b). There the labor supply curve shifts rightward, from L_1^S to L_2^S. As the new entrants compete for jobs with those already working there, the wage rate falls—to $4 per hour, in our example.

Are skilled workers in panel (c) affected by the minimum wage? You might think not, because they already earn above the minimum. But remember: As the wage rises in panel (a), and employment falls, firms will substitute other inputs, such as capital, for unskilled labor. For example, an unskilled product packager might be replaced by a high-tech packaging machine that is produced, operated, and maintained by skilled workers. Substitution like this shifts the labor demand curve in panel (c) rightward, from L_1^D to L_2^D. The skilled wage rate rises—to $24 in our example.

To summarize:

A higher minimum wage benefits those unskilled workers who maintain their jobs and are paid more. It also benefits skilled workers by raising their equilibrium wages. But it harms those unskilled workers who cannot find work and those who work in the uncovered sector, where wages decrease.

Because this policy is such a blunt instrument for helping low-income people at best, and because it harms some of those it is supposed to help, many economists

oppose raising the minimum wage as a strategy to help low-income workers. Instead, they advocate a superior alternative.

The EITC Alternative

An alternative to the minimum wage, which has been very well regarded by economists, is the Earned Income Tax Credit (EITC). Begun in 1975 and expanded several times, the EITC supplements the incomes of low-income working people. In 2009, the EITC for a low-income worker supporting two children could reach more than $4,800 from the federal government, and often another $1,000 or more from a state EITC. This is in addition to income earned on the job, effectively adding as much as $2.40 an hour to the wage rate.

Many economists favor substantially increasing the EITC to help those earning low incomes. Increasing the EITC has several advantages over increasing the minimum wage:

- The minimum wage applies to *any* unskilled worker in the covered sector, regardless of family income or economic need. The EITC, by contrast, is only available to low-income households and provides greater benefits to those supporting children.
- The costs of the minimum wage are spread among households rather haphazardly, with no regard to income. The funds for the EITC come from a progressive federal tax system, making it genuinely redistributive from higher income to lower income households.
- The minimum wage is likely to reduce employment, but the EITC tends to increase employment. (By providing a subsidy to labor *suppliers*, it shifts the labor supply curve downward and rightward.)

Opposing Views

Given all the problems with the minimum wage and the existence of a superior alternative, you might think that economists overwhelmingly oppose any increase in the minimum wage. But that is generally not the case. Surveys of economists—including those specializing in labor markets—show majorities favor at least some increase in the minimum wage.[9] There are several reasons for this.

First, the employment effects of a small increase in the minimum wage are viewed by some as modest. Most of the employment effects would fall on teenagers, and the consensus view is that a 10-percent increase in the minimum wage (say, from $6.60 to $7.25) would cause total teenage employment to fall by about 2 percent.

[9] See, for example, Victor R. Fuchs, Alan B. Krueger, and James M. Porterba, "Economists' Views about Parameters, Values and Policies: Survey Results in Labor and Public Economics," *The Journal of Economic* Literature, Vol. 36, No. 3, 1998, pp. 1387–1425. For more recent arguments, see Daniel B. Klein and Stewart Dompe, "Reasons for Supporting the Minimum Wage: Asking Signatories of the 'Raise the Minimum Wage' Statement," *Econ Journal Watch*, Vol. 4, No. 1, January 2007 (pp. 125–167).

Some go even further, pointing to research by two prominent labor economists that suggests a rise in the minimum wage causes *no* decrease in employment.[10] The methods of this research have been hotly debated, and its conclusions have not been widely accepted by the profession.

Another reason for supporting a higher minimum wage is political. Funds for expanding the EITC, which would come out of general government revenues, are constrained by federal budget discipline. The costs of a higher minimum wage (higher prices for consumers and/or lower earnings for business owners) are non-budgetary and more easily hidden from view. Thus, even those who would prefer an expanded EITC may believe it is easier to get public support for a higher minimum wage.

Using the Theory

THE COLLEGE WAGE PREMIUM

Students have many motives for attending college, but one of the most important motives is to invest in their own human capital. College graduates on average earn higher incomes than those who don't attend college. Economists measure the earnings gains from college with the *college wage premium:* the percentage by which the average college graduate's wage rate exceeds that of the average high school graduate. This premium has been large—over 50% since 1980—and growing consistently over the last several decades.[11]

To explain the persistence—and rise—of this premium, we have to explain not only why the average wage rate for college graduates is rising, but why it is rising faster than that for high school graduates.

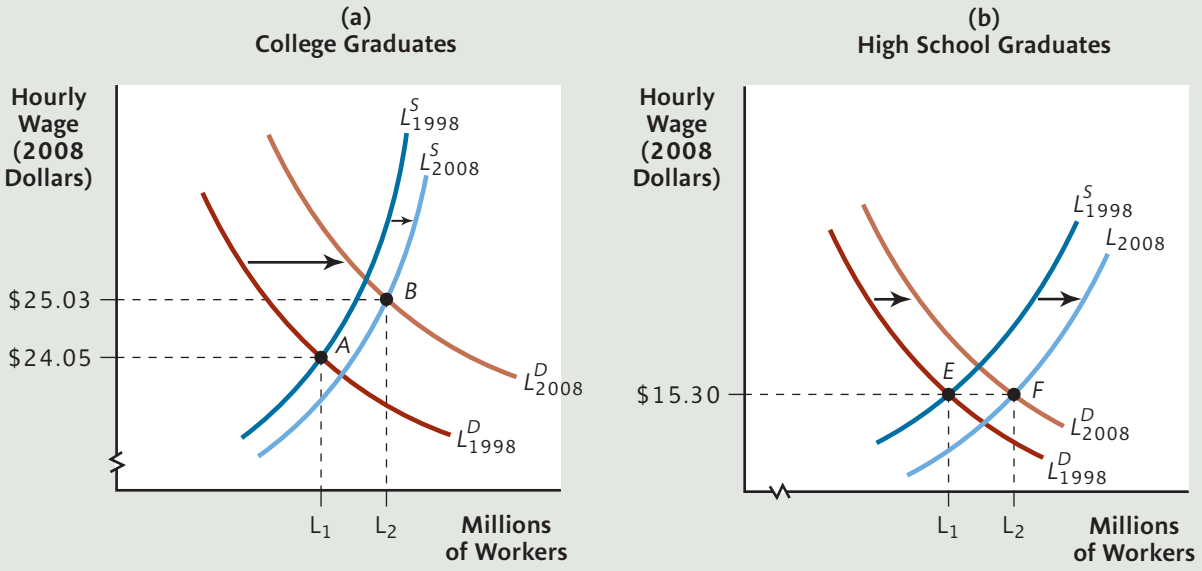
Supply and Demand for College Graduates versus High School Graduates

Figure 15 shows what has been happening in these two labor markets from early 1998 to early 2008. Look first at the left panel, for college graduates. In 1998, the median college graduate was earning $24.05 per hour. Over the next 10 years, several million additional college graduates entered the labor force, shifting the labor supply curve rightward. Nevertheless, the median wage rose to $25.03, because labor demand increased even more than labor supply.[12]

[10] David Card and Alan B. Krueger, *Myth and Measurement: The New Economics of the Minimum Wage* (Princeton, NJ: Princeton University Press, 1995).

[11] The college wage premium can be measured in various ways, controlling for variables such as age, gender, ethnic group, location, and more. Most of these measures behave similarly over time. In this case, we do not control for demographics at all. Our wage premium is the percentage by which the median wage for all those with a college (but no higher) degree exceeds the median wage for all those with a high school degree (and no college).

FIGURE 15 The College and High School Labor Markets over Time



In 1998, the average college graduate earned $24.05 per hour, at point A *in panel (a). That same year, the average high school graduate earned $15.30, at point* E *in panel (b). Over the next ten years, supply and demand curves shifted right in both markets. But in the college graduate market in panel (a), demand increased more rapidly than supply, so the equilibrium moved to point* B*, with a higher wage of $25.03. For high school graduates in panel (b), the increases in supply and demand were about equal. The equilibrium moved to point* F*, and the wage rate hardly changed at all.*

Why did the labor demand curve shift rightward? It was partly due to normal growth in the economy. As the economy grows, demand increases in product markets. Existing firms produce more and new firms are started. These forces create an ongoing increase in demand for most types of labor.

But the demand for college graduates rose even faster, due to *technological change*. New technologies were developed and used more widely—such as personal computers and the Internet. These technologies are strongly *complementary* with more skilled workers, who are disproportionately college graduates (as well as those with higher degrees). For example, anyone whose work requires manipulating, interpreting, or using information has been made more productive, and more valuable to firms, by the Internet. Doctors can find alternative treatments faster, attorneys can find legal precedents faster, and business managers can monitor costs and production more completely and more rapidly. Portable electronic devices and wireless technology have enhanced this effect, turning what used to be less-productive travel time into productive work hours.

[12] Source for median wages: Bureau of Labor Statistics, "Usual Weekly Earnings of Wage and Salary Workers," First Quarter 1998, and First Quarter 2008, Table 4, for full-time workers, and author calculations. We have converted 1998 earnings to 2008 dollars using the BLS's CPI–RS and converted weekly earnings to hourly wages assuming a 40-hour work week.

As you've learned, new complementary technologies shift the demand for labor rightward. And for the last several decades, including the one illustrated in the figure, the pace of technological change has been exceptionally rapid.

But technological change that is complimentary for one type of labor can be substitutable for another. Routine jobs such as adding up numbers, handling simple requests, sorting and delivering documents, or putting things together on an assembly line can increasingly be performed by computers and computer-controlled machines. These new technologies are thus *substitutable* for less-skilled labor, which is disproportionately those who do *not* graduate from college. As you've learned, a substitutable technology shifts the demand for labor leftward.

When a technological advance simultaneously increases the demand for high-skilled labor and decreases the demand for less-skilled labor, economists call it *skill-biased technological change*. For high school graduates, only normal growth in the economy shifts the demand curve rightward. Skill-biased technological change has the opposite effect.

The net result: the demand shift *relative* to the supply shift has been much smaller for high school than for college graduates. In fact, for high school graduates, the supply and demand curves have shifted rightward about equally over the past decade, as illustrated in the right panel of the figure. This explains why the average wage of high school graduates rose hardly at all—from \$15.30 to \$15.37.

Because the college wage rate grew faster, it follows that the college wage premium increased over the decade as well. In 1998, the premium was (\$24.05 − \$15.30)/\$15.30 = 0.57 or 57%. By 2008, it had risen to 63%.

The Remaining Puzzle

Illustrating these two labor markets with supply and demand curves helps us see why the college wage premium exists, and why it has grown. But a puzzle remains. Since 1980, the college wage premium has been high: more than 50%. The typical college graduate earns this premium for 40 years of work or more, adding more than \$1 million to lifetime income. This amount far exceeds any compensating wage differential for the costs of attending college, even with full tuition at a private school, and certainly at a less-expensive state college.[13]

Now recall our "imaginary world" where all wages are equal. If differences in human capital costs were the *only* departure from that world, we would expect massive migration from the high-school graduate labor market to the college graduate labor market. That is, we'd expect more rapid growth in college attendance. This would cause the labor supply curve to shift rightward more rapidly in the panel (a), slowing wage growth and, for a time, perhaps even reducing the wages of college graduates. Meanwhile, in panel (b), the labor supply curve for high school graduates would shift *less* rapidly, or even reverse direction. This would create faster wage growth for high school graduates. We'd expect these movements to bring the wages in the two markets closer together, until the difference between them was just enough to compensate for the cost of college. And in our imaginary world, this would take just a few years—the time required for the new, larger group of college graduates to hit the labor market.

[13] In the next chapter, we'll discuss how to calculate the dollar value of a college degree, using a technique called *present discounted value*.

But this has not happened. In fact, the opposite seems to be occurring: the college wage premium is rising, not falling. And research by two economists[14] suggests that since 1980, the major force driving the college wage premium higher is a *slowdown* in the growth of college attendance in the United States.

Why would growth in college degrees be slowing, rather than rising? This is the subject of much debate. Some economists point to easily-identified *barriers to entry* that are keeping potentially able people out of college. Examples include financial barriers (insufficient availability of student loans) or psychological barriers (inertia or anxiety that prevents families from filling out complex financial aid forms, as behavioral economists have found). For problems like these, modest changes in public policy could reverse the trend.

Others point to more complex social forces that limit college attendance, such as teenage pregnancy, extreme poverty, or failing public schools. Policy solutions for these problems are broader and more politically contentious. Finally, some argue that we may be approaching the upper limit of the fraction of the population intrinsically *capable* of succeeding at college.

Of course, a growing college wage premium has benefits as well as costs. Because you are reading this book, you will probably soon be a college graduate yourself. So for you, a higher premium means a greater income than you would otherwise have.

But there are social implications to a rising premium, and these depend on its origins. When the premium rises because of a more rapidly-growing *demand* for college graduates, it acts as a useful price signal, providing incentives for more people to attend college. Human capital grows more rapidly, and this—just like growth in physical capital—contributes to more rapid economic growth.

Yet when (as seems to be the case now) the premium is rising largely because of a slowdown in the *supply* of college graduates, it means *less* investment in human capital and *less* economic growth than we would otherwise have. Further, if artificial barriers or unresolved social problems are preventing able people from attending college, a large and rising college wage premium exacerbates social divisions and tensions.

SUMMARY

Resources are traded in *factor markets* in which firms are demanders and households are suppliers. The *labor market* is a key factor market. A *perfectly competitive labor market* is one in which there are many buyers and sellers of standardized labor, with easy entry and exit of workers, and well-informed workers and firms. The demand for labor by a firm is a *derived demand*—derived from the demand for the product the firm produces.

In a competitive labor market, each firm faces a market-determined wage rate. The labor demand curve slopes downward because of the output effect (at higher wage rates, the firm produces less output) and the input-substitution effect (at higher wage rates, the firm substitutes other inputs for that type of labor). It shifts rightward when the demand for the product increases. A new or cheaper input will shift the demand curve rightward if the input is complementary with that labor and leftward if the input is substitutable for that labor. On the supply side, the *labor supply curve* slopes upward. Given the number qualified, more people will want to work in a labor market at higher wage rates. The labor supply curve shifts rightward when more people become qualified, when alternative

[14] Claudia Goldin and Lawrence Katz, *The Race Between Education and Technology* (Harvard University Press, 2008), especially Chapter 8. See also their paper, "Long-Run Changes in the U.S. Wage Structure: Narrowing, Widening, Polarizing," *Brookings Papers on Economic Activity*, 2:2007.

labor markets become less attractive, or when tastes change in favor of that labor market.

The market labor supply and demand curves intersect to determine the market wage rate and employment in each labor market. Labor market equilibrium will change if either the labor demand or labor supply curve shifts. Wage differences can be explained by compensating wage differentials (for the nonwage characteristics of different jobs or for human capital costs), and because of differences in ability. Because of the "superstar" nature of many professions, small differences in ability can create extremely large wage differentials.

Barriers to entry created by occupational licensing, unions, or discrimination, can create wage differences. The market works against discrimination that arises from employer prejudice, but not against that arising from customer or employee prejudice, nor against statistical discrimination.

One frequent proposal for reducing income inequality is an increase in the *minimum wage*. This is controversial because the costs and benefits of the minimum wage do not fall exclusively on high- and low-income households, respectively. A higher minimum wage is also believed to reduce employment among low-skilled workers, especially teenagers, and to reduce the wages of unskilled workers in industries not covered by the law. A more efficient redistributive policy is the *earned income tax credit* (EITC), which targets low-income working households.

PROBLEM SET

Answers to even-numbered Questions and Problems can be found on the text Web site at www.cengage.com/economics/hall.

1. In the nation of Barronia, the market for construction workers is perfectly competitive. Explain what would happen to the equilibrium wage rate and equilibrium employment of construction workers under each of the following circumstances:
   a. Young adults in Barronia begin to develop a taste for living in their own homes and apartments, instead of living with their parents until marriage.
   b. Construction firms begin to use newly developed robots that perform many tasks formerly done by construction workers.
   c. Because of a war in neighboring Erronia, Erronian construction workers flee across the border to Barronia.
   d. There is an increased demand for automobiles in Barronia, and Barronian construction workers have the skills necessary to produce automobiles.
   e. Tuition at trade schools that qualify Barronians for construction work become cheaper.
2. Suppose that dehydrated meat is an inferior good. Discuss the effect on the equilibrium wage rate and level of employment in the dehydrated meat industry of an increase in national income.
3. Suppose the supply of people who want to be "extras" in a film (people who pretend to be minding their own business in the background of a film) is completely inelastic, because they will supply their labor to film production companies no matter what wage they are paid.
   a. Draw a diagram illustrating how the equilibrium wage and equilibrium employment level for extras is determined.
   b. On the same diagram, illustrate what will happen to the equilibrium wage and employment of extras if the demand in the product market (for films) increases.
4. Fifteen years ago, college professors frequently hired undergraduates as research assistants to gather basic information in the library. Today, most professors can get the information themselves using the Internet in less time than it would take to explain what is needed to a research assistant.
   a. In the labor market for undergraduate research assistants, has the Internet been a substitutable or complementary technological change?
   b. All else equal, what impact has the development of the Internet likely had on the wage and employment level of undergraduate research assistants?
   c. Many college professors find that *graduate* student research assistants are more productive than before because they can use the Internet. All else equal, has the Internet been a substitutable or complementary technological change for graduate student research assistants? What effect would this have on their equilibrium wage and employment level?
5. Suppose that in the market for U.S. meat packers, two things happen simultaneously: (1) Due to growth in less-developed countries, the demand for U.S. meat exports rises; and (2) Due to other job opportunities, fewer people find meat-packing an attractive job.
   a. Which curve or curves in the labor market for meat packers would be affected by these changes?
   b. If there were no other changes, could you predict the impact of these events on the equilibrium wage of meatpackers? On their equilibrium employment level? Explain.

6. The EITC is a subsidy given to workers for working. Suppose everyone in a particular unskilled labor market is receiving EITC payments of a given amount for each hour that they work.
   a. Draw a diagram showing the impact of the EITC on the labor market equilibrium. [Hint: The subsidy is paid directly to workers, not to firms.] On your graph, identify the wage rate and employment level both before and after EITC is introduced.
   b. What happens to the wage rate that employers pay to their workers after the EITC? Explain briefly.
   c. What happens to the wage rate the workers get (including the subsidy)? Explain briefly.

7. How would each of the following, ceteris paribus, affect the college wage premium over a long period of time?
   a. Government scholarships are sharply curtailed.
   b. Weird solar radiation renders all computers in the world inoperable.
   c. Computers become even more complicated to operate.

8. Each of the following observations, *ceteris paribus*, implies something about a particular labor market. For each one, (1) identify the relevant labor market; (2) state whether you expect the wage rate in that market to be higher than otherwise, or lower than otherwise; and (3) identify which conceptual explanation for wage differences (among those discussed in the chapter) you are applying to reach your conclusion.
   a. Elevator repairers have a higher accident rate than most other jobs.
   b. People who work in New York City have to pay higher rent than other people.
   c. People like to work in New York City because there is a lot to do there.
   d. It takes rare diligence, motivation, and practice to become a concert pianist.
   e. Los Angeles has passed legislation requiring that fortune tellers become certified by the city.
   f. Office cleaning services don't like to risk hiring ex-convicts, because employers believe they are more likely to steal equipment than other employees.

9. [Appendix] The following gives employment and daily output information for Your Mama, a perfectly competitive manufacturer of computer motherboards.

Number of Workers	Total Output
10	80
11	88
12	94
13	97
14	99

   A motherboard worker at Your Mama earns $80 a day, and motherboards sell for $27.50.
   a. How many workers will be employed? How do you know?
   b. Suppose the market wage for motherboard workers increases by $5 per day per worker, but the market price of motherboards remains unchanged. What will happen to employment at the firm? Why?

10. [Appendix] Add *MR* and *MC* columns to Table A.1 in the chapter and find the profit-maximizing output level using the *MR* and *MC* approach. When calculating *MR* and *MC*, don't forget to divide ΔTR and ΔTC by the change in output, which is *not* one unit in the table. Does the profit-maximizing output level differ from the one found in the appendix, using the $MRP = W$ approach? Explain briefly.

11. Many people think that immigration into the United States, because it causes competition for jobs, will lower the wage rates of U.S. workers. Yet, even though the United States admits hundreds of thousands of immigrants each year, the average U.S. wage has continued to grow. Can you explain why? Are there any groups of workers within the economy for whom the fear of lower wages is justified? Explain.

More Challenging

12. Some advocates of the minimum wage argue that any decrease in the employment of the unskilled will be slight. They assert that an increase in the minimum wage will actually increase the total amount paid to unskilled workers (i.e., wage × number of unskilled workers employed). Discuss what assumptions they are making about the wage elasticity of labor demand.

13. Suppose the demand for unskilled labor were completely *inelastic* with respect to the wage rate. Using graphs similar to those in Figure 14, but modified to reflect this new assumption, explain how a minimum wage above the equilibrium wage for covered unskilled workers would affect employment and the wage rate among:
   a. Covered, unskilled workers
   b. Uncovered, unskilled workers
   c. Skilled workers

The Profit-Maximizing Employment Level

This appendix presents the formal mechanics behind the firm's employment decision, and shows how to derive a firm's labor demand curve from information on production and costs. We start with a general rule you learned in Chapter 7: the *marginal approach to profit*:

> *The marginal approach to profit states that a firm should take any action that adds more to its revenue that it adds to its cost.*

In the previous chapters, you've seen this rule applied to one kind of action: increasing production. Now, we'll apply it to a different type of action: increasing employment. Our rule translates to: Increase *employment* whenever it adds more to revenue than it adds to cost. Our first step is to see how an increase in employment affects a firm's revenue.

Employment and Revenue

In Table A.1, we return to an example we met earlier: Spotless Car Wash. The first two columns are based on information in Table 1 of Chapter 7, where we assumed that Spotless has one automated car-washing line. Column 1 shows different numbers of workers that Spotless can employ, and Column 2 shows daily output (number of cars washed) for each level of employment.

We'll skip over Column 3 for now, in order to calculate total revenue for each level of employment. In the table, we assume that Spotless sells its product in a competitive market, i.e., it's a *price taker*. It can wash all the cars it wants at the market price, which is $8. To get total revenue, we can just multiply the quantity of output (column 2) by the constant price (column 4). This gives us the total revenue in column 5.

The Marginal Revenue Product (*MRP*) of Labor

Column 6 introduces a new concept: the *marginal revenue product* of labor.

> *The firm's* ***marginal revenue product*** **(MRP)** *of labor is the change in the firm's total revenue* (ΔTR) *divided by the change in the number of workers employed* (ΔL)*:*

$$MRP = \Delta TR/\Delta L$$

In words, the *MRP* of labor tells us how an additional worker will add to the firm's revenue. In the table, the *MRP* numbers are obtained by dividing the change in total revenue in Column 5 by the change in employment in column 1. For example, when employment increases from 2 to 3 workers (an increase of 1 worker), its daily revenue rises from $720 to $1,040 (an increase of $320). So, for this change in employment,

$$MRP \text{ of Labor} = \Delta TR/\Delta L = \$320/1 = \$320$$

Another Way to Calculate *MRP*

We can also calculate (and think of) the *MRP* in another way. Column 3 shows the marginal product of labor (*MPL*)—a concept you also learned about in Chapter 7. The MPL tells us the additional *output* produced when one more worker is hired. For example, when the firm hires the third worker, output rises from 90 to 130, so the *MPL* for this change in employment is 40.

Now, each time the firm hires another worker, its output (car washes) rises by the marginal product of labor (*MPL*). Each unit of this additional output sells for $8. Therefore, each time another worker is hired, *revenue* must increase by $MPL \times \$8$.

More generally,

$$MRP \text{ of Labor} = MPL \times P$$

TABLE A.1

Data for Spotless Car Wash

(1) Number of Workers	(2) Output (Cars Washed per Day)	(3) Marginal Product of Labor (*MPL*)	(4) Price per Car Wash	(5) Total Revenue	(6) Marginal Revenue Product (*MRP*)	(7) Daily Wage (*W*)
0	0		8	$ 0		$120
		30			$240	
1	30		8	$240		$120
		60			$480	
2	90		8	$720		$120
		40			$320	
3	130		8	$1,040		$120
		30			$240	
4	160		8	$1,280		$120
		24			$192	
5	**184**		**8**	**$1,472**		**$120**
		12			$96	
6	196		8	$1,568		$120
		4			$32	
7	200		8	$1,600		$120

You can verify that each value for *MRP* in column 6 is equal to the *MPL* in column 3 times the price in the column 4.[15]

Finally, notice how the *MRP* values change as we travel down column 6. As employment increases, *MRP* first rises (up to the hiring of the second worker), then falls (beginning with the hiring of the third worker). This mirrors the behavior of the *MPL* values in column 4. Recall, from Chapter 7, that at low levels of employment and output, *increasing marginal returns to labor* (rising *MPL*) are likely. But eventually, as employment continues to rise, a firm will eventually experience *diminishing marginal returns to labor* (falling *MPL*). So whenever hiring another worker causes *MPL* to rise or fall, it will cause $MRP = MPL \times P$ to rise or fall as well. The rise and fall of *MRP* is based on the behavior of *MPL*.

[15] $MRP = MPL \times P$ holds only when output is sold in a perfectly competitive market, in which the firm faces a horizontal demand curve for its product. If the firm faces a *downward-sloping* demand curve for its product, as in monopoly or monopolistic competition, there is a different relationship between *MRP* and *MPL*. Hiring another worker still increases output by the *MPL*, but now the firm must drop its price in order to sell the additional output. In this case, hiring another worker increases the firm's revenue by the additional output produced (*MPL*) times the increase in revenue per unit increase in output (*MR*). Thus, a more general equation for *MRP* is $MPL \times MR$. When the product market is competitive, MR is just the market price, P.

Employment and Cost

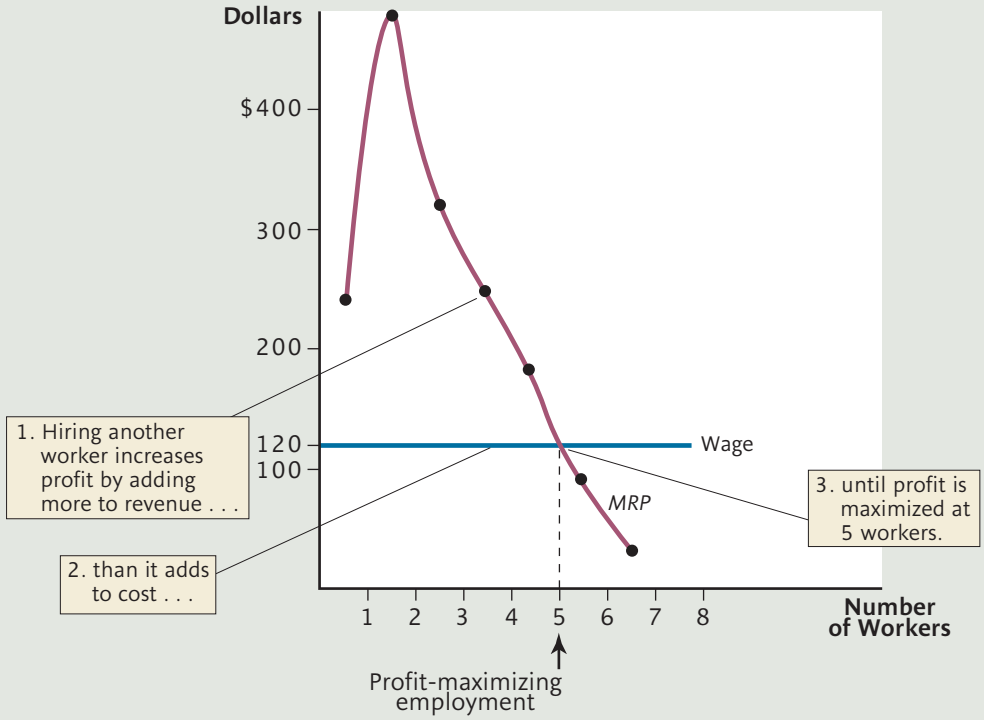
Let's leave revenue for now, and consider how hiring an additional worker affects the firm's cost. The *wage rate* is the number of dollars the firm pays for each employee *per unit of time*. It can be defined for any time period, such as an hour, a day, etc. But to enable direct comparisons, we should choose the same time period we use for measuring output and revenue. In our example, output and revenue are measured daily, so we'll use a *daily* wage.

As we did in the chapter, we'll assume that our firm hires its labor in a *competitive* labor market, so it acts as a wage taker. In Table A.1, we've assumed that the market wage that Spotless must pay is $120 per day. Therefore, all the entries in column 7 are $120.

The Profit-Maximizing Employment Level

We're finally ready to combine what we know about revenue, cost, and the marginal approach to profit to determine how many workers a firm will employ. The marginal approach to profit tells us that the firm should hire another worker whenever doing so adds more to revenue (*MRP*) than it adds to cost (the wage rate).

FIGURE A.1 The Profit-Maximizing Employment Level



Using W to represent the daily wage rate, we can state the firm's guiding principle this way:

> *Hire another worker when* MRP > W, *but not when* MRP < W.

Let's apply this guideline to Spotless Car Wash. When going from 0 to 1 worker, revenue rises by \$240 (MRP = \$240) and cost rises by \$120 (W = \$120). Because revenue rises more than cost ($MRP > W$), hiring this first worker will add to the firm's profit. The same is true when the second, third, fourth, and fifth workers are hired. (Verify this on your own.) But in moving from the fifth to the sixth worker, MRP = \$96, while W= \$120. Since $MRP < W$, the firm should *not* hire the sixth worker; it should stop at the fifth. We have found the firm's profit-maximizing level of employment: five workers.

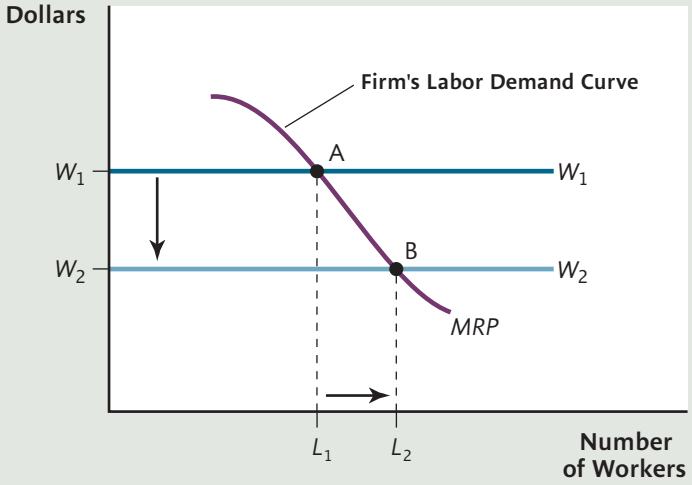
We can understand Spotless's employment decision even better by graphing the marginal data from Table A.1, as we've done in Figure A.1. As usual, marginal values are plotted *between* employment levels, since they tell us what happens as employment changes from one level to another. The value of MRP first rises and then falls as employment changes, so the *MRP curve* in the figure first slopes upward and then downward. The wage rate is always the same, as shown by the horizontal line at \$120.

As long as employment is less than five workers, the MRP curve lies above the wage line ($MRP > W$), so the firm should hire another worker. But suppose the firm has hired five workers and is considering hiring a sixth. For this move, the MRP curve lies *below* the wage line. Since $MRP < W$, increasing employment would *decrease* the firm's profit. The same is true for every increase in employment beyond five workers. Using Figure A.1, we see that the optimal employment level is five workers, just as we found earlier using Table A.1.

The profit-maximizing number of workers, five, is the employment level closest to where $MRP = W$—that is, closest where the MRP curve crosses the wage line. The reason is: For each change in employment that

FIGURE A.2 The Firm's Labor Demand Curve

As the wage rate varies, the firm moves along its MRP *curve in deciding how many workers to hire. As a result, the downward-sloping portion of the* MRP *curve is the firm's labor demand curve. It shows how many workers will be demanded at each wage rate.*



increases profit, the *MRP* curve will lie above the wage line. The first time that hiring a worker *decreases* profit, the *MRP* curve will cross the wage line and dip below it.

This observation allows us to state a simple rule for the firm's employment decision:

> *To maximize profit, the firm should hire the number of workers such that* MRP = W*—that is, closest to the point where the* MRP *curve intersects the wage line.*

A Proviso

There is one proviso for the previous highlighted statement: Profits are maximized only if the *MRP* curve crosses the wage line *from above*. That is, we must be on the *downward-sloping* portion of the *MRP* curve. To prove this to yourself, draw an example in which the *MRP* curve crosses the wage line twice—one as it slopes upward and once as it slopes downward. In your drawing, you'll notice that the *MRP* will always be greater than the wage to the right of the crossing point, so it will always pay for the firm to *increase* employment beyond that point. Thus, that crossing point cannot be the profit maximizing employment level. It is only the *downward-sloping* part of the *MRP* curve matters for the firm's employment decision.

Two Approaches to Profit Maximization

Let's compare two different approaches that a firm can use to maximize its profit. In previous chapters, we used the *MR* = *MC* approach to find profit-maximizing *output,* which requires a certain amount of labor. In this appendix, we've used the *MRP* = W approach to find profit-maximizing *employment,* which will produce a certain level of output. Can these two approaches lead to different employment and output decisions at the firm? For example, for Spotless, we've discovered that *MRP* = W tells Spotless to employ five workers. But five workers (as you can see in Table 1) implies a total output of 184 car washes per day. Could the *MC* and *MR* approach guide Spotless to some *other* level of output, say, 196 car washes?

The answer is: No, because these two "different" approaches are actually the *same* approach viewed differently. To see why, remember that hiring another *worker* increases the firm's output and therefore changes both its revenue and its cost. For example, in Table A.1, increasing employment from four to five workers raises output by 24 units (from 160 to 184 units). This, in turn, increases revenue by $192 and cost by $120. Since hiring the fifth worker increases revenue more than it raises cost (i.e., *MRP* > *W*), then it must also be true that increasing output by those *24 units* raises revenue

by more than it raises cost (i.e., $MR > MC$ for that increase in output).

This applies generally: Whenever $MRP > W$ for a change in employment, $MR > MC$ for the associated rise in output. Whenever $MRP < W$ for a change in employment, $MR < MC$ for the associated rise in output. And if $MRP = W$ for a change in employment, then it must be that $MR = MC$ for the associated change in output.[16]

The Firm's Labor Demand Curve

In Table 1, the firm took the wage rate of \$120 per day as given. But what if the wage rate had been different, say, \$90 per day? As you can verify by modifying the table, at this lower wage rate, the firm would have hired *six* workers instead of five. The optimal level of employment will always depend on the wage rate.

Figure A.2 shows what happens at the typical firm as the wage rate varies. For each wage rate, the optimal level of employment, where $MRP = W$, is found by traveling horizontally over to the MRP curve and then down to the horizontal axis. For example, with a wage rate of W_1, the firm will want to hire L_1 workers. If the wage drops to W_2, the optimal level of employment rises to L_2. As the wage rate drops, the firm moves along its MRP curve in deciding how many workers to employ. This is why we call the downward-sloping portion of the MRP curve the *firm's labor demand curve:*

The downward-sloping portion of the MRP curve is the firm's labor demand curve, telling us how much labor the firm will want to employ at each wage rate.

From Firm to Market

The labor market demand curve is found by simply adding the labor demand curves for every firm that employs labor in that market. For example, at a daily wage of \$120, Spotless's labor demand curve tells us it will employ five workers. Suppose a second firm would hire 12 workers from the same labor market at that wage, and a third firm would hire eight, and so on. On the *market* labor demand curve, the quantity of labor demanded at a wage of \$120 per day would be $5 + 12 + 8 + \cdots$ and so on (continuing to add the employment for all the firms in the market). As we've seen, at each of these firms, increasing the wage rate reduces employment. Therefore, in the labor market as a whole, increasing the wage rate reduces the quantity of labor demanded.

The market labor demand curve, like each firm's labor demand curve, slopes downward.

[16] To help you see the connection between these two approaches even more clearly, add two new columns for MR and MC in Table 1 and use them to find the profit-maximizing output level. But when calculating MR and MC, don't forget to divide ΔTR and ΔTC by the change in output as you move from row to row.

CHAPTER 13

Capital and Financial Markets

If you are like most people reading this book, you have already decided to make a tradeoff: to give up some income now, so you can have more income in the future. You made the decision when you chose to attend college rather than work full time. In return, after graduation, you can expect to earn more than you would without a college degree.

There are, of course, other reasons to attend college than just the boost in future income it will give you. But let's be narrow-minded for a moment, and think only about the money. Is the tradeoff worth it?

We'll discuss the answer at the end of this chapter. You'll see that the method we'll use helps us answer a variety of economic questions. The method is used by Starbucks when it must decide whether to open or close a store. And by Pfizer when it decides how much of the firm's yearly income to put into research and development on new drugs. It is also used by financial market analysts, when they are asked whether a share of Google stock is really worth its price of several hundred dollars. All of these decisions involve putting a value on dollars to be received *in the future*.

In this chapter, we'll be discussing how such decisions are made. We'll first look at decisions about investing in physical capital, such as factory buildings or machinery. Then we'll turn our attention to the value of financial assets, specifically stocks and bonds. Finally, in the Using the Theory section, we'll turn to the decision to invest in human capital—specifically, your decision to attend college.

Physical Capital and the Firm's Investment Decision

The concept of *capital* was introduced in the first chapter of this book. There, you learned that capital is one of society's *resources*, along with land, labor, and entrepreneurship. You also learned that we can classify capital into two categories: *physical capital,* such as the plant and equipment owned by business firms, and *human capital,* the skills and training of the labor force. In this section, we'll focus on firms' decisions about physical capital.

How does a business firm decide how much physical capital to use? In the same way that it makes any other decision. The firm's goal is to maximize its profit—not just this year, but over many years into the future. What guidelines should a firm use?

That depends on how the firm *pays* for its capital. We'll look first at the decision to *rent* capital. Then, we'll move to the more complex decision to *purchase* capital.

A First, Simple Approach: Renting Capital

In this section we assume that a firm *rents* its capital at a constant price per unit of time (per hour, day, etc.), just as it "rents" its labor. We'll make our discussion more concrete with an example. Imagine you are the fleet manager at Quicksilver Delivery Service. Your firm delivers packages for small retailers in the Chicago metropolitan area. You are responsible for determining the number of trucks the firm should use.

Your first step is to remember the *marginal approach to profit:*

> *The marginal approach to profit states that a firm should take any action that adds more to its revenue than it adds to its cost.*

Here, the action is "rent another truck." Keep in mind that when we use this approach, revenue and cost are measured *per period,* such as revenue and cost per year. So, the marginal approach says that you should rent another truck if doing so will increase yearly revenue more than it increases yearly cost.

Table 1 lists the data that will help you make your decision. The first column tells us different numbers of trucks you could rent. The second tells us the additional revenue each truck would generate for your firm each year. Notice that the additional annual revenue decreases as you use more trucks. For example, renting the first truck increases your annual revenue by $52,000. The second truck adds a bit less—only $50,000. This pattern makes sense: With just one truck, you would use it on your busiest route (the one where you can deliver the most packages each day). If you rent a second truck, you'd use it on the next busiest route, and so on. So each additional truck you rent will generate less additional revenue than the truck before.

The third column tells us all of the annual *incidental* expenses of using a truck (costs other than renting the truck itself). These include the annual costs for a driver's wages, insurance, gasoline, and maintenance. We've assumed for simplicity that these truck-related expenses are $40,000 per year for each truck, no matter how many trucks are rented.

Ultimately, we want to compare the additional revenue from a truck with its additional cost. So what should we do with these incidental expenses? One option is to include them as a cost, along with the rent. Another option is to deduct these incidental expenses from revenue, and then compare what's left (net revenue) to the rent. Either option will lead us to the same conclusion about how many trucks to

TABLE 1

Quicksilver Delivery's Truck Rental Decision

(1) **Trucks**	(2) **Additional Annual Revenue**	(3) **Additional Annual *Incidental* Costs**	(4) **Additional Annual *Net* Revenue** Column 2– Column 3	(5) **Additional Rental Cost**
First Truck	$52,000	$40,000	$12,000	$8,000
Second Truck	$50,000	$40,000	$10,000	$8,000
Third Truck	$47,500	$40,000	$ 7,500	$8,000
Fourth Truck	$46,000	$40,000	$ 6,000	$8,000
Fifth Truck	$44,000	$40,000	$ 4,000	$8,000

rent. But the latter approach—deducting these incidental expenses from revenue first—helps simplify things later on. So that's what we've done in Table 1.

Column 4 shows the result: the *net* revenue from each additional truck. For example, the second truck brings in annual revenue of $50,000, but after deducting expenses for the driver, gas, insurance, and so forth, the *net* additional revenue is $10,000. Notice that the *net* additional revenue in column 4, like the additional revenue, declines as you add more trucks.

Finally, in column 5, we come to the annual rent for each truck. We've assumed the annual rent is a constant $8,000 per truck, no matter how many trucks you rent.

Now we're ready to apply the marginal approach to profit, using the data in the table. You should *rent another truck* whenever its (net) additional revenue (column 4) is greater than the cost of renting the truck itself (column 5), because doing so will add to your profit.

Running down the table, we see that it makes sense to rent the first and second truck, because each one adds more in net revenue than it costs to rent. For example, the second truck adds $10,000 to net revenue each year, but costs only $8,000 to rent. But the third truck adds less to net revenue ($7,500) than to cost ($8,000). The same is true for the forth and fifth truck. Therefore, your profits are maximized by renting exactly two trucks.

When capital is rented, rather than purchased, the marginal approach to profit tells us the firm should rent another unit of capital whenever the additional (net) revenue per period is greater than the additional rental cost per period.

The Limits of the Simple Approach

Our first simple approach—which assumed that firms rent their capital—has a serious limitation. Our assumption that capital can be *rented*—while it may work for *some* firms—does not work for the economy in general. That's because every unit of capital in use is owned by someone or some firm. Even if Quicksilver Delivery Service rents its trucks, it will be renting them from a truck rental firm that purchased them. For any unit of capital employed in the economy, some firm—somewhere along the line—must have made the decision to purchase it. So if we want to understand decisions about capital investment in the economy, we must ultimately account for the decisions of firms that *purchase* the capital before it is used. For that reason, from this point on, we'll focus on the firms that *purchase* their capital.

This changes the decision-making process, because capital equipment, once purchased, will last a long time. At Quicksilver, for example, suppose a truck will last for 10 years. Then you'd have to compare the cost of buying a truck—which you have to pay *now*—with the additional net revenue the truck would bring in over the next ten years.

"That's easy," you might think. "I'll just add up the revenue over those 10 years. The fourth truck, for example, brings in $6,000 per year, so in 10 years, it will bring in a total of $60,000. As long as the truck costs less than $60,000, it adds to my profit."

But if you reason this way, you are making a serious error: You're treating each year's revenue as equally valuable, regardless of *when* the revenue is earned. In reality, the value of a future payment depends on *when* that payment is received. To see why, we'll have to take a detour from Quicksilver Delivery and explore the issue of future payments more generally. We'll come back to Quicksilver and its trucks when we're done.

The Value of Future Dollars

To see why the value of a future payment depends on *when* that payment is received, just run through the following thought experiment. Imagine that you are given the choice between receiving $1,000 now and $1,000, with certainty, one year from now. Do you have to think hard before making up your mind? It's always better to have dollars earlier rather than later. But the *economic* reason is this: If you get the $1,000 now, you could put it in the bank and earn interest for a year, giving you *more* than $1,000 a year from now.

Because present dollars can earn interest, it is always preferable to receive a given sum of money earlier rather than later. Therefore, a dollar received later has less value than a dollar received now.

Knowing that dollars received in the future are worth less than dollars received today is an important insight. But *how much* less is a future dollar worth?

To answer that question, we use a concept called *present value*.

*The **present value*** (PV) *of a future payment is its equivalent value in today's dollars. Alternatively, it is the most anyone would pay today for the right to receive the future payment with certainty.*

Present value The value, in today's dollars, of a sum of money to be received or paid at a specific date in the future with certainty.

To understand this concept better, let's work out a simple example: What is the present value of $1,000 to be received one year in the future? That is, what is the most you would pay *today* in order to receive $1,000 one year from today? The answer is clearly *not* $1,000. If you paid $1,000 today for a guaranteed $1,000 in one year, you would lose something: the interest you *could* have earned during the year. So it never makes sense to pay $1,000 now for $1,000 to be received one year from now.

But would you pay $900 for the guaranteed future payment? Or $800? That depends on how much interest you *could* earn by lending funds to someone else for a year. In fact, the most you'd pay is the amount that, if you lent it out for interest, would get you *exactly* $1,000 in one year. More concretely, suppose the interest rate you can earn if you lend out funds is 10 percent per year. Then the present value of $1,000 to be received on year from today is $909.09. Why? Because if you had $909.09 today, and loaned it out at 10 percent interest, then—one year from now—you'd have $909.09 × 1.10 = $1,000. As long as you can be sure you'll be repaid (an issue we'll discuss in more detail later), you shouldn't care whether you receive $909.09 today or $1,000 in the future.

But how did we *find* this present value of $909.09?

©ANTON FOLTIN/ISTOCKPHOTO

A truck, like most types of physical capital, will increase a firm's revenue for many years. As a result, the firm must calculate the present-dollar equivalent of future receipts.

Calculating Present Value

To find the present value of a $1,000 to be received one year in the future, we ask: What amount of money, loaned out today at a 10 percent annual interest rate, would give the lender $1,000 in one year? Let's call that amount of money "X." Then X has to satisfy the following equation: $(X)\ (1.10) = \$1{,}000$. Solving this equation for X, we find that

$$X = \frac{\$1{,}000}{1.10} = \$909.09.$$

Now remember that "X" in this equation is the number of today's dollars that you could turn into $1,000 in one year. That is, "X" is just the present value we've been seeking. So let's use the symbol "PV" (for Present Value) in place of X. Then, when the annual interest rate is 10 percent, we can say:

$$PV \text{ of } \$1{,}000 \text{ in one year} = \frac{\$1{,}000}{1.10} = \$909.09$$

In words, if you lent out $909.09 at 10 percent interest, you would have $1,000 one year from today. Therefore, $909.09 is the most you would be willing to give up today for $1,000 in one year. Or, more formally, *$909.09 is the present value of $1,000 received one year from now.*

We can generalize this result by noting that, if the interest rate had been something other than 0.10—we'll call it r—or the amount of money had been something other than $1,000—say, Y dollars—then the present value would satisfy the equation

$$PV \times (1 + r) = Y$$

or

$$PV = \frac{Y}{(1 + r)}.$$

What if the payment of $\$Y$ were to be received *two* years from now instead of one? Then we can use the same logic to find the present value. In that case, each dollar lent out now would become $(1 + r)$ dollars after one year. Then, when the dollar plus the earned interest was lent out again for a second year, it would become $(1 + r)(1 + r) = (1 + r)^2$ dollars at the end of the second year. Thus, the PV will satisfy

$$PV \times (1 + r)^2 = Y$$

and solving for PV, we obtain

$$PV = \frac{Y}{(1 + r)^2}.$$

Finally, for payments to be received one, two, or any number of years n in the future, we can state that

the present value of $\$Y$ *to be received n years in the future is equal to*

$$PV = \frac{Y}{(1 + r)^n}.$$

For example, with an interest rate of 10 percent, the present value of $1,000 to be received three years in the future would be

$$PV = \frac{\$1{,}000}{(1.10)^3} = \$751.31$$

Discounting The act of converting a future value into its present-day equivalent.

Discount rate The interest rate used in computing present values.

The process of making dollars of different dates comparable is called **discounting.** The value of r used in this process is often called the **discount rate.**[1]

Determinants of Present Value

Table 2 shows the present value (rounded to the nearest penny) of a dollar to be received at different times in the future, at different interest rates. To see where the numbers come from, let's do an example. In the table, locate the entry for present

[1] In macroeconomics, the term *discount rate* has a completely different meaning: It's the interest rate that the Federal Reserve charges banks when it lends them reserves. There is no connection between the two different meanings of the term.

TABLE 2

Present Values of $1 Future Payments

No. of Years in Future	Value of $1 to Be Received at Various Numbers of Years in the Future, at Different Discount Rates		
	5 Percent	10 Percent	15 Percent
0	$1.00	$1.00	$1.00
1	$0.95	$0.91	$0.87
2	$0.91	$0.83	$0.76
3	$0.86	$0.75	$0.66
4	$0.82	$0.68	$0.57
5	$0.78	$0.62	$0.50
10	$0.61	$0.39	$0.25
20	$0.38	$0.15	$0.06

value of $1 to be received 10 years from today when the interest rate is 10 percent. (The entry is $0.39.) Where did this number come from?

With an interest rate of 10 percent, our formula tells us that $PV = \$1/(1.10)^{10} = \$1/2.59 = \$0.39$. This tells us that, when the interest rate (discount rate) is 10 percent, anyone expecting to receive $1 ten years from today might just as well accept $0.39 now. After all, when loaned at 10 percent interest per year, 39 cents will get you $1 in ten years.

Some patterns in the table are worth noting. One pattern is found by running down any of the columns. Notice that the present value entries get smaller: As the $1 payment is postponed further into the future, its present value declines. This makes sense: The longer the payment is postponed, the less you would pay today to receive that future payment, because the interest you'd forgo by waiting is greater.

All else equal, a payment received later has a lower present value than a payment received earlier.

A second pattern is found in each row. For a given future payment date, as the interest rate rises from 5 percent to 10 percent to 15 percent, the entries get smaller. This make sense too: The higher the interest rate, the more interest you sacrifice by waiting to get your payment, so the less you'd pay for it now.

All else equal, the present value of a future payment is lower when the interest rate is higher.

Finally, although this is not seen in the table, the present value of a future payment is greater when the future payment itself is greater. In fact, you can find the present value of any number of dollars by multiplying the entries in the table by that number of dollars. For example, the present value of $500 to be received 10 years in the future is 500 times the present value of $1 received at that time, or

dangerous curves

Percentages and Decimals Be careful when working with interest rates: They can be expressed in either percentage form or decimal form. An interest rate of 5 percent (5%) can also be expressed in decimal form as 0.05.

In the expression $1+r$, r is always in decimal form. For example, $1+r$ is equal to 1.05 when the interest rate is 5 percent. Similarly, an interest rate of 0.5% (*one-half* of 1 percent) would translate to 0.005 in decimal form, and $1+r$ would then equal 1.005.

$500 \times \$0.39 = \$195$. However, because we've rounded the entries in the table to the nearest penny, the present values for larger dollar amounts will be approximations.

The Present Value of a Future *Stream* of Payments

The formula for present value calculations can also help us determine the value of a *stream* of future payments, with each individual payment to be received at a *different* time in the future. Consider the value, in today's dollars, of the following stream of future payments: $1,000 to be received one year from now, $900 to be received two years from now, and $600 to be received three years from now. To get the present value of this stream of payments, we first calculate the present value of each payment, and then we add those present values together:

$$PV = \frac{\$1,000}{(1+r)} + \frac{\$900}{(1+r)^2} + \frac{\$600}{(1+r)^3}$$

With an interest rate of 10 percent, the *total* present value of the entire stream of payments is equal to:

$$\begin{aligned} PV &= \frac{\$1{,}000}{(1.10)} + \frac{\$900}{(1.10)^2} + \frac{\$600}{(1.10)^3} \\ &= \$909.09 + \$743.80 + \$450.79 \\ &= \$2{,}103.68 \end{aligned}$$

The logic of present value shows us why anyone who expects to receive a stream of future payments must discount each of those payments before adding them together. The next section provides an example of how firms use present value to make decisions about investing in new capital.

Purchasing Capital

Let's return to your problem at Quicksilver Delivery Service, where you've decided to purchase, rather than rent, your trucks. How many trucks should you buy? Table 3 shows the present value calculations you'd need to make, under the following conditions:

- The net additional revenue per year for each truck is as given in Table 1, a few pages earlier, and will occur with *certainty*.
- Each truck has an expected useful life of 10 years.
- You have the option to lend funds without risk at an annual interest rate of 10%. (So 10% is the appropriate discount rate for your present value calculations.)

For example, the first truck gives Quicksilver $12,000 per year in additional revenue for 10 years. Since we're assuming for simplicity that each year's revenue is received at the end of each year (see the Dangerous Curves feature above), the present value of the first year's revenue is $12,000/(1.1); the present value of the second year's revenue is $12,000/$(1.1)^2$; and so on. When these present values are added

dangerous curves

When Are Future Payments Received? Businesses typically earn revenue every day they are in operation. However, in calculating present discounted value, it would be cumbersome to discount each day's revenue by the appropriate discount factor. As a useful approximation, we can treat each year's revenue as if it is all received in one lump sum at the *end* of the year. This is the convention followed in this book for all future payments. Thus, when we say that a firm or individual receives a payment of $10,000 "in the first year," we mean "at the end of the first year." (Can you see how we've used this assumption in Table 3?)

TABLE 3

The Present Value of Trucks at Quicksilver Delivery Service (with a Discount Rate of 10%)

Trucks	Additional Annual Net Revenue	Total Present Value of Additional Net Revenue over 10 years
First Truck	$12,000	$\frac{\$12,000}{(1.1)} + \frac{\$12,000}{(1.1)^2} + \cdots + \frac{\$12,000}{(1.1)^{15}} = \$73,735$
Second Truck	$10,000	$\frac{\$10,000}{(1.1)} + \frac{\$10,000}{(1.1)^2} + \cdots + \frac{\$10,000}{(1.1)^{15}} = \$61,446$
Third Truck	$ 7,500	$\frac{\$7,500}{(1.1)} + \frac{\$7,500}{(1.1)^2} + \cdots + \frac{\$7,500}{(1.1)^{15}} = \$46,084$
Fourth Truck	$ 6,000	$\frac{\$6,000}{(1.1)} + \frac{\$6,000}{(1.1)^2} + \cdots + \frac{\$6,000}{(1.1)^{15}} = \$36,867$
Fifth Truck	$ 4,000	$\frac{\$4,000}{(1.1)} + \frac{\$4,000}{(1.1)^2} + \cdots + \frac{\$4,000}{(1.1)^{15}} = \$24,578$

together for all 10 years, we find that the first truck gives the firm $73,735 in net additional revenue in present value terms. The other entries are calculated similarly.

Now that we know the total present value that you gain from each truck, do we know how many trucks you should buy? Almost, but not quite. There is still the matter of how much each truck *costs*. But now that we've translated *all* the additional revenue from each truck into a single, present value number, we know the benefits of each truck measured in *today's dollars*. That measure can be compared to the truck's cost, which must *also* be *paid* in today's dollars. If trucks cost $45,000, the firm gains more benefits (in future revenue) than costs for the first three trucks. But the purchase of the fourth truck, whose benefit to the firm is only $36,867 in today's dollars, does not make sense, since the cost in today's dollars is $45,000. Quicksilver should buy only three trucks.

Our examples have focused on a special type of capital—delivery trucks. But the same logic works for any other type of physical capital—automated assembly lines, desktop computers, filing cabinets, locomotives, and construction cranes. In each of these cases, the first step in making a decision about a capital purchase is to put a value on the benefits of the capital. This value is the total present value of the future revenue generated by the capital.

This first step—putting a value on physical capital—is so important and so widely applicable that we can refer to it as a general principle:

*The **principle of asset valuation** says that, when there is no uncertainty, the value of any asset is the sum of the present values of all the future benefits it will generate.*

Principle of asset valuation The idea that the value of an asset is equal to the total present value of all the future benefits it generates.

The principle of asset valuation tells us how to determine the marginal benefit from buying another unit of capital, such as another truck. Then, as we've done with Quicksilver, we compare this marginal benefit with the cost of the capital itself. As you've seen, the firm should then buy any unit of capital for which the marginal benefit (total present value of future revenue) is greater than the cost.

An Important Proviso: Risk

In our discussion so far, we've been assuming that your future receipts at Quicksilver are known with certainty. But realistically, future receipts are risky. The demand for package deliveries might decrease, or a competitor might drain away some of your

revenue, or the price of gasoline could spike. You can never be sure how much net revenue each truck will bring in to your firm.

Suppose that the net revenue entries in Table 1 are your best guess about the future. But there's also a small chance that each truck will bring in either $1,000 less or $1,000 more than your best guess. If you calculate the present value numbers for the *worst*-case scenario (in which each truck generates $1,000 less per year than in Table 1), the first two trucks would still have a present value greater than their cost. The third truck, however, would not.

Should you buy the third truck now? The worst case scenario isn't likely . . . but it's possible. On the other hand, the best case scenario—with every truck generating $1,000 more than your best guess each year—could happen as well. You now have a risky investment decision to make.

All else equal, most investors would prefer to face as little risk as possible. So the value of any future payment should be smaller when that payment is uncertain. Because we calculate present value assuming our "best guess" scenario is *certain* to occur, our present value calculations will *overestimate* the true value of future payments involving risk.

> *When there is uncertainty about future benefits, the principle of asset valuation is just our starting point. It tells us the* most *an asset would be worth. The greater the uncertainty or the greater one's aversion to risk, the more the value of a risky asset will fall short of its present value.*[2]

In this chapter, when there is uncertainty, we'll continue to calculate present values as if the "best guess" about the future is certain to occur. Therefore, risk will not affect present value itself. But risk will cause the true value of an asset to be *less* than its present value.

Going back to Quicksilver, under the "best guess" scenario, the third truck has a present value of $46,084—greater than its cost. But because of the risk, the third truck will be worth less than $46,084 to you—perhaps (if you dislike risk enough) less than $45,000. In that case, because of the uncertainty, you would not buy the third truck.

What Happens when Things Change: The Investment Curve

Investment Firms' purchases of new capital over some period of time.

Investment is the term economists use to describe firms' purchases of new capital over some period of time. In the example above, if trucks cost $45,000 each, Quicksilver should buy three of them. If it bought all three trucks this year, its investment expenditures for the year would be $45,000 × 3 = $135,000.

But this conclusion about investment is based on the assumption that the interest rate, and Quicksilver's discount rate, is 10 percent. With a lower interest rate—say, 5 percent—each year's revenue would have a *higher* present value, so the total present value of each truck would be higher. Our conclusion about Quicksilver's investment spending might then change. Similarly, a rise in the interest rate—say, to

[2] There are other ways to handle decision-making under risk. For example, risk can be incorporated into the present value calculation itself, by using a higher discount rate (the interest rate for riskless lending plus a "risk premium"). In that case, uncertainty will reduce the present value itself. In this textbook, however, risk will not affect the way we calculate present value.

TABLE 4

Present Value Calculations for Various Interest Rates

Trucks	Additional Annual Revenue	Total Present Value with a Discount Rate of: 5%	10%	15%
First Truck	$12,000	$92,661	$73,735	$69,213
Second Truck	$10,000	$77,217	$61,446	$57,678
Third Truck	$ 7,500	$57,913	$46,084	$43,258
Fourth Truck	$ 6,000	$46,330	$36,867	$34,606
Fifth Truck	$ 4,000	$30,887	$24,578	$23,071

15 percent—would *lower* the present value of each year's revenue, and *decrease* the total present value for each truck.

Table 4 shows how our total present value calculations for each truck change as the interest rate changes. The numbers in the last three columns are each calculated just as were the numbers we calculated in Table 3. The only difference is that, instead of always assuming a discount rate of 10 percent, Table 4 shows the total present value for each truck under three different interest rates. Notice what happens as we move from left to right in the table for any particular truck: The interest rate rises, from 5 percent to 10 percent to 15 percent, and the value of the truck to the firm falls.

Now, if trucks cost $45,000 each, how much will Quicksilver invest (spend on new trucks) at any given interest rate? Let's see. If the interest rate is 5 percent, Quicksilver should buy four trucks, because each of the first four trucks has a total present value greater than $45,000 at that interest rate. The fifth truck, however, has a total present value of only $30,887, so the firm should not buy that one. Thus, if the interest rate is 5 percent, Quicksilver's investment spending will be $45,000 × 4 = $180,000.

If the interest rate rises to 10 percent, we are back to the conclusion we reached in Table 3, which assumed a 10 percent interest rate: Quicksilver should buy three trucks when the interest rate is 10 percent. (You can also verify this using the middle column of Table 3.) Quicksilver's total investment spending would *decrease* to $45,000 × 3 = $135,000. Finally, if the interest rate rises to 15 percent, Quicksilver should buy only two trucks, so its total investment spending is $45,000 × 2 = $90,000.

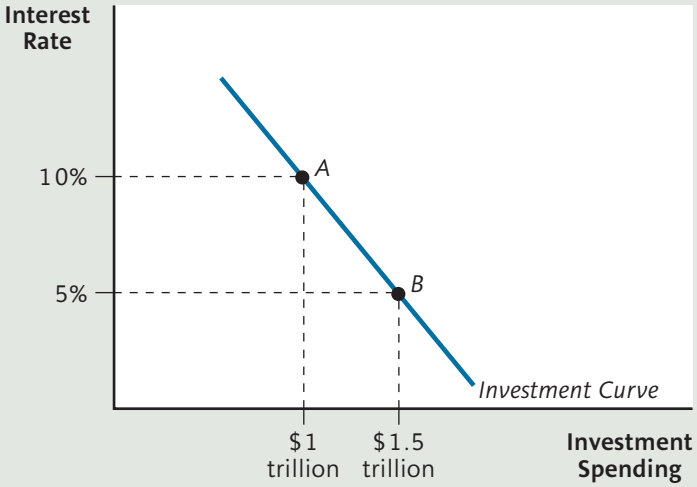
What is true for Quicksilver is true for every other truck-buying firm in the economy: The higher the interest rate, the fewer trucks delivery services and other truck-buying firms will want to purchase, and the smaller will be investment expenditures on trucks during the year.

Take a moment to think about why this happens. The trucks themselves are the same, and they are just as productive at any of the three interest rates. But each truck is less valuable to firms in *present-dollar* terms. That's because—with a higher interest rate—the future additional revenue from each truck is worth *less* in today's dollars. But the truck still costs the same in today's dollars, regardless of the interest rate. So with a high interest rate, each firm will want fewer trucks at any given price.

The same logic applies to other capital purchases. At high interest rates, U.S. firms end up buying less of all different kinds of capital—not just delivery trucks, but also other durable goods such as computers, machine tools, combines, and

FIGURE 1 The Investment Curve

As the interest rate falls from 10 percent to 5 percent, each firm that buys a particular type of capital will buy more of it. As a result, the economy's total investment in physical capital rises from $1 trillion to $1.5 trillion. This is shown as the movement from point A to point B along the investment curve in the figure.



printing presses. It should be no surprise, then, that we come to the following conclusion:

> *As the interest rate rises, each business firm in the economy—using the principle of asset valuation—will place a lower value on additional capital, and decide to purchase less of it. Therefore, in the economy as a whole, a rise in the interest rate causes a decrease in investment expenditures.*

The Economy's Investment Curve

The relationship between the interest rate and investment expenditure is illustrated by the economy's investment curve, shown in Figure 1. The curve slopes downward, indicating that a drop in the interest rate causes investment spending to rise. When you study *macroeconomics,* you'll learn that the investment curve is important for the performance of the overall economy, for several reasons. But here's a hint as to one of them: When the interest rate falls, the increased investment in new capital means that the nation's *capital stock*—the total quantity of installed capital—will grow more rapidly than it otherwise would. With more capital, labor will be more productive, and our standard of living will be higher.

To recap:

> *Lower interest rates increase firms' investment in physical capital, causing the capital stock to be larger, and our overall standard of living to be higher.*

Markets for Financial Assets

You may be wondering what markets for financial assets, like stocks and bonds, have to do with the other subject of this chapter: markets for capital. After all, capital—like a machine or a factory—is something *real;* it enables a firm to produce real goods and services.

But in financial markets, the things being traded are just *pieces of paper,* which don't directly help anyone to produce anything. So what do these pieces of paper have to do with capital?

Actually, quite a bit. The pieces of paper being traded in financial markets are **financial assets**—promises to pay future income to their owners. Because capital lasts for many years, most firms get the funds to purchase their capital by issuing and selling these financial assets. For example, the firm might issue and sell shares of *stock* in the company, obligating it to pay those who hold the shares part of the firm's future profits. Or it might sell *bonds*, which are promises to pay back a sum of money in the future, along with interest payments. This leaves the firm with long-lasting capital, but also a long-lasting obligation to make future payments. So there is a close *economic* connection between a firm's decision to demand capital equipment and its decision to supply financial assets.

Financial asset A promise to pay future income in some form, such as future profits or future interest payments.

Primary and Secondary Asset Markets

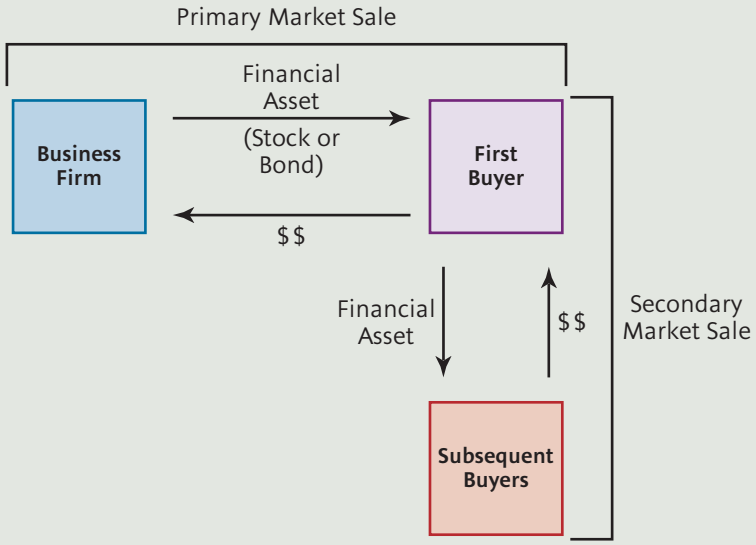
When a newly issued financial asset is sold for the first time, we label this a **primary market** trade. But the buyer can then sell the asset to someone else, in which case we call it a **secondary market** trade. The difference between these two types of trades is illustrated in Figure 2.

Primary market The market in which newly issued financial assets are sold for the first time.

Secondary market The market in which previously issued financial assets are sold.

Once a financial asset is issued and sold in the primary market, it can change hands many times in the secondary market. In fact, almost all of the trading that takes place in financial markets on any given day is secondary market trading. But it is only through primary market sales that firms actually obtain funds for investment projects. Secondary market trading is an exchange

FIGURE 2 Primary and Secondary Asset Markets



When a corporation creates a financial asset (such as a share of stock or a bond), it is sold to the public in the primary market. *This is how the corporation raises funds for investment. The financial asset can then be resold by its original purchaser to someone else in the* secondary market. *Secondary market transactions have no direct impact on the firm that originally issued the assets*

between private parties, and the original issuing firm or government agency is not involved.

Still, firms and government agencies follow secondary markets closely. Why? Because financial asset prices in the secondary and primary markets are closely related. In fact, at any given time, the prices of two otherwise identical assets in the primary market and secondary markets should be identical.

For example, suppose that Microsoft wants to obtain funds by issuing new shares of stock in the primary market. In order to attract buyers, it will have to sell these *new* shares at the same price that *old* Microsoft shares are selling for in the secondary market. After all, every regular share of Microsoft stock provides the same benefits to its owner. So a buyer has no reason to prefer a new share to an old one, or vice versa. The same is true of bonds: When a newly issued bond promises the same future payments on the same dates as a previously issued bond, buyers have no preference for one over the other.

While firms that issue financial assets are not direct participants in secondary market trading, they are affected by what happens in the secondary market. More specifically, if an asset's price rises in the secondary market, so will the price the firm can charge for similar, newly issued assets in the primary market.

Now you can begin to see why firms care so much about secondary market trading. Each newly issued financial asset obligates the firm to make payments to its holder in the future. The higher the price at which the firm can *sell* the new financial asset, the more dollars it will get in exchange for the future obligations it is taking on. Or, viewed another way, the higher the selling price of a financial asset, the fewer assets the firm will have to issue to raise any given amount of money now.

Financial Assets and Present Value

We've discussed one connection between capital and financial markets. But here's another: Financial assets, just like capital equipment, gives their owners a stream of future benefits—in this case, dollar payments. Therefore, the value of a financial asset is calculated in the same way as the value of any other asset, such as a truck or a computer: We find the *total present value* of the future payments the asset will generate. Thus, the method of valuation is another connection between markets for capital and markets for financial assets.

The principle of asset valuation applies to financial assets as well as physical assets. In each case, the value of the asset is the sum of the present values of the future benefits.

In the rest of this chapter, we'll explore two types of financial assets: bonds and stocks. We'll also analyze the very well-publicized markets in which these assets are traded.

The Bond Market

If a firm wants to buy a new fleet of trucks, build a new factory, or upgrade its computer system, it must decide how to finance that purchase. One way to do this is to sell **bonds**. A bond is simply a promise to pay a certain amount of money, called the **principal** or **face value**, at some future date. Although $10,000 is the most common principal amount, you can also find bonds with face values of $100,000, $5,000, and other amounts.

Bond A promise to pay back borrowed funds, issued by a corporation or government agency.

Principal (face value) The amount of money a bond promises to pay when it matures.

A bond's **maturity date** is the date on which the principal will be paid to the bond's owner. If a bond has a maturity date 30 years after the date on which it was first sold, we'd call it a 30-year bond. Other bonds have shorter maturities—15 years, 10 years, 1 year, 6 months, or even 3 months.

Maturity date The date at which a bond's principal amount will be paid to the bond's owner.

Some bonds, including many of those sold by the U.S. federal government, are **pure discount bonds.** A discount bond is one that does not make any payments except for the principal it pays at maturity.

Pure discount bond A bond that promises no payments except for the principal it pays at maturity.

For example, a pure discount bond might promise to pay its owner $10,000 when it matures in two years. If the bond originally sells for $8,900, then the total interest—the difference between what the bond originally sold for and what the owner will receive at maturity—is $10,000 − $8,900 = $1,100.

A bond's **yield** (more formally, "yield to maturity") is the constant annual rate of return that the buyer would earn on his investment in the bond if it were held to maturity. For example, if you buy a pure discount bond paying $10,000 in two years for $8,900, your yield is 6%. How do we know? Because (as you can verify) if you took that same $8,900 and earned interest at a constant rate of 6% per year for two years, you'd end up with $10,000—the same as you'd get from the bond at maturity.

Yield The annual rate of return a bond earns for its owner.

Most bonds are a bit more complicated than pure discount bonds. In addition to repayment of principal, they also make a series of interim payments called **coupon payments.** For example, a 30-year $10,000 bond might promise a coupon payment—say $600—each year for the next 30 years, and then pay $10,000 at maturity. But every bond—even those with multiple coupon payments—has a yield. In the case of a coupon bond, the calculations are more difficult. (Financial calculators come in handy!) But the yield has a similar meaning: It's the constant annual rate of return that—by timing the withdrawals just right—would enable you to receive the same schedule of future payments as you would get from the bond.

Coupon payments A series of periodic payments that a bond promises before maturity.

How Much Is a Bond Worth?

To determine the value of a bond, let's start with a simple example: a pure discount bond that promises to pay $10,000 when it matures in exactly one year. The $10,000 is a future payment, and our method of calculating its value should not surprise you: It involves *present value*. Let's suppose (for now) that the bond will pay just what it promises, with certainty. That is, we'll assume no risk of *default* (failure to pay as promised).

Let's also suppose the interest rate at which you can lend funds elsewhere with just as much certainty is 10 percent. Then we can calculate (to the nearest dollar) the present value of the bond with our discounting formula as follows:

$$PV = \frac{\$Y}{(1+r)} = \frac{\$10{,}000}{1.10} = \$9{,}091.$$

Because the present value of $10,000 to be received in one year is $9,091, that is the most you should pay for the bond. It is also the lowest price at which its current

owner will sell the bond to you. We conclude that this bond will sell for $9,091, no more and no less.

The same principle applies to more complicated types of bonds, such as discount bonds that don't pay off for many years, or coupon bonds. For example, suppose a bond maturing in five years has a principal of $10,000, and also promises a coupon payment of $600 each year until maturity, with the first payment made one year from today. The total present value of this bond would be:

$$PV = \frac{\$600}{(1.10)} + \frac{\$600}{(1.10)^2} + \frac{\$600}{(1.10)^3} + \frac{\$600}{(1.10)^4} + \frac{\$600}{(1.10)^5} + \frac{\$10{,}000}{(1.10)^5} = \$8{,}484.$$

Once again, this total present value—$8,484—is what the bond is worth, and this is the price at which it will trade, as long as buyers and sellers use the same discount rate of 10 percent in their calculations.

Bond Prices and Bond Yields

There is an important relationship between the price of a bond and the yield or rate of return the bond earns for its owner. This is easiest to see with a pure discount bond, such as the bond that pays $10,000 in one year in our earlier example. Suppose you bought this bond for $8,000. Then, at the end of the year, you would earn interest of $10,000 − $8,000 = $2,000 on an asset that cost you $8,000. Your yield would be $2,000/$8,000 = 0.25 or 25 percent.

But now suppose you paid $9,000 for that same bond. Then your interest earnings would be $10,000 − $9,000 = $1,000, and your yield would be $1,000/$9,000 = 0.111 or 11.1 percent.

As you can see, the yield a bond gives you depends on it's price. For each price, there is a different yield. And the greater the price of a bond, the lower the yield on that bond. This applies not only to simple discount bonds, but also to more complicated bonds with coupon payments. And the reasoning is the same in both cases: A bond promises to pay fixed amounts of dollars at fixed dates in the future. The more you end up paying for those promised future payments, the lower your rate of return.

More generally:

> *There is an inverse relationship between bond prices and bond yields. The higher the price of any given bond, the lower the yield on that bond.*

Bond Yields and Interest Rates

You may have noticed something interesting in some of our examples. When there is no risk of default, a bond's yield is equal to the interest rate used to determine its present value. For example, take the pure discount bond that pays $10,000 in two years. Suppose the interest rate for present value calculations (the rate you can earn on other riskless lending) is 6 percent. Then we can calculate the present value of the bond as $10,000/(1.06)² = $8,900.

dangerous curves

Bond Prices, Bond Yields, and Opportunity Cost You might think that the inverse relationship between bond prices and yields applies only to those *buying* bonds, and not those already holding them. After all, your yield is determined when you first buy a bond, and then it never changes, right?

Not right. If the price changes, the yield changes too, even for those who bought earlier at a different price. That's because the *opportunity cost* of continuing to own a bond is its *current* price, not the price you originally paid.

For example, suppose you initially pay $9,091 for a bond that gives you $10,000 in one year. The yield is initially $909/$9,091 or 10 percent. But now suppose, the day after you buy the bond, its price jumps to $9,500. If you continue to hold the bond, you give up $9,500 that you *could* have by selling it. For this $9,500 sacrifice, you will get $10,000 in one year. So your yield is now $500/$9,500 or 5.3 percent. This is the same yield that someone *buying* the bond for $9,500 would get. The point to remember is: When bond prices rise, the yield for *every* bondholder falls.

But, as we've seen previously, if you buy that bond for a price equal to this present value, your yield is 6 percent—the same as the annual interest rate on riskless lending.

If you think about it for a moment, this makes perfect sense. If you have other opportunities to lend safely and earn 6 percent per year, you won't buy a bond unless it offers at least 6 percent. And the bond won't yield any *more* than 6 percent, because other buyers would find it attractive enough to buy when the yield is 6 percent.

All else equal, a riskless bond will have a yield equal to the interest rate on other riskless loans—that is, the interest rate used to calculate its present value.

Why Do Bond Yields Differ?

Thousands of different kinds of bonds are traded in financial markets every day. There are corporate bonds of various maturities and bonds issued by local, state, and federal governments and government agencies. Bonds issued by foreign firms and governments are also traded in the United States. And each bond has its own unique yield. Why is this? Why don't all bonds give the same yield? That is, why doesn't each bond sell at a price that makes its yield identical to the yield on any other bond?

The answer is that bonds differ in important ways from one another.

Default Risk

One difference among bonds is in their *default risk*. A bond is a promise to pay in the future, and there is always a danger that the promise won't be kept. When a firm goes bankrupt, the holders of its bonds may receive none, or only some, of the payments they were promised. For example, when General Motors declared bankruptcy in June 2009, it defaulted on bonds with a face value of $27 billion, declaring that there would be no further coupon or principal payments. In their place, bondholders were given shares of stock in a newly-formed successor corporation. But these shares were worth substantially less than the bonds they replaced.

Governments, too, sometimes default on their bonds. In January 2009, the government of Ecuador defaulted on about $4 billion in bonds issued by the previous administration, declaring they were not legally binding on the new administration.

Now recall that when we calculated the present value of a bond—what the bond is "worth"—we used the interest rate at which you could lend out funds *safely*. The present value we obtained is what you should be willing to pay for a bond that has *no* risk of default. But if there is a risk of default, you will *not* be willing to buy the bond at that present-value price. Instead, you (and everyone else) will only buy it a *lower* price, to compensate you for the extra risk you are taking. Of course, with a lower price, the bond will have a higher yield (albeit a risky one).

Bonds with default risk sell for less than the present value of their promised payments. All else equal, the greater the risk of default, the lower the bond's price, and the greater its yield.

Table 5 shows that the market does value bonds in this way. In the table, bonds are listed in the order of increasing risk, according to Fitch Ratings, a private corporation that analyzes corporations and municipalities that issue bonds and estimates the likelihood that they will default. U.S. Treasury bonds top the list. They are backed by the promise of the U.S. government and have virtually zero probability of

TABLE 5

Interest Rates on 5-Year Bonds, June 15, 2009

Rating	Yield
U.S. Treasury bond	2.72 percent
AAA corporate bond	3.15 percent
AA corporate bond	3.95 percent
A corporate bond	4.73 percent
BBB corporate bond	5.98 percent
BB corporate bond	6.64 percent
B corporate bond	9.07 percent

Source: http://finance.yahoo.com. Yields for bonds below A rating are based on sample of individual bond quotes on the same date.

default. Next is AAA, the highest rating given to the most credit-worthy corporations and municipalities. The ratings continue down through AA, A, BBB, and so on.

Notice how the yields diverged on June 15, 2009. For example, the difference between the riskless yield of 2.72 percent on U.S. Treasury bonds and the more risky BB yield was about 4 percentage points. That difference compensates investors for the chance that a BB bond will go into default before it matures.

Other Reasons that Bond Yields Differ

Although we've stressed differences in default risk, bond yields can differ for other reasons as well. These include differences in maturity dates, frequency of coupon payments, differences in how the interest is taxed, or because one bond is more widely traded (and therefore easier to sell on short notice) than another. If you study economics further, you'll learn how each of these contributes to differences in yields in a course on "Money and Banking" or "Financial Economics."

Explaining Bond Prices

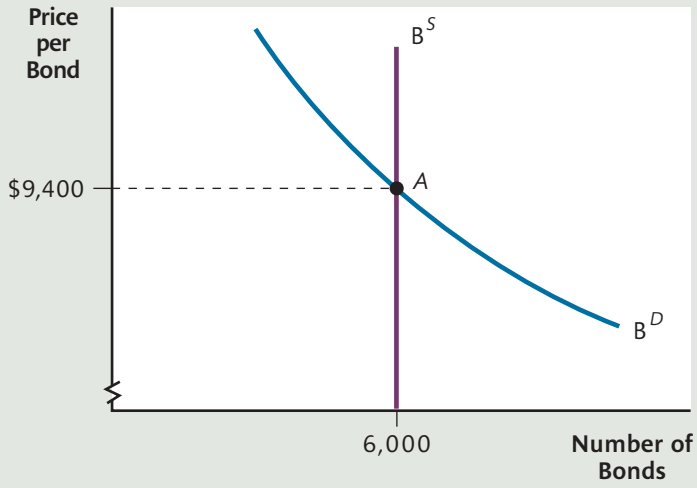
Bond prices at any given time are determined by supply and demand. Figure 3 provides an example. It shows supply and demand curves for a Verizon Communications bond that matures in exactly one year, with a face value of $10,000 and no coupon. One year from today, the owner of this bond will receive $10,000 from Verizon. The number of these bonds is on the horizontal axis, and the price of each bond is on the vertical axis.

Notice that the supply curve, B^S, is vertical. You've seen a vertical supply curve like this before—for housing, in Chapter 4. Recall the reason that the curve was vertical there: We were viewing housing as a "stock" variable: the number of homes in existence at a given time. The supply of homes was the number available for ownership, and the demand for homes was the number that people wanted to own.

We analyze the bond market in a similar way. We view supply and demand for Verizon bonds *not* as the quantity people want to sell and buy each period, but rather as the quantity available and the quantity that people want to own *at any given moment*. (That is, we treat bonds as a "stock" variable, rather than a "flow" variable).

In the figure, we assume that Verizon has issued a total of 6,000 bonds in the past. So on any given day, the total number of these bonds in existence—and available for people to own—is 6,000, regardless of the price on that day. The supply of

FIGURE 3 The Market for One-Year GM Bonds



In the market for Verizon bonds that pay $10,000 in one year, equilibrium occurs where the number of bonds people want to hold equals the number in existence (6,000). In the figure, the equilibrium occurs at point A, with a price of $9,400. At this price, each bond earns $600 in interest, for an annual interest rate (yield) of $600/$9,400 = 0.064 or 6.4%.

bonds can change only if Verizon issues more of them (shifting the curve rightward) or retires some by buying them back from their current owners (shifting the curve leftward). Unless that happens, the supply curve remains the vertical line at 6,000.

The demand curve, B^D, tells us the quantity of bonds that people want to own on a given day, at different hypothetical prices. It slopes downward, telling us that the lower the price, the more of these bonds people will want to own.

Why do people want to own more bonds at a lower price? In large part, because people are different. They have different attitudes toward risk and different beliefs about how risky a bond really is. Let's explore this further.

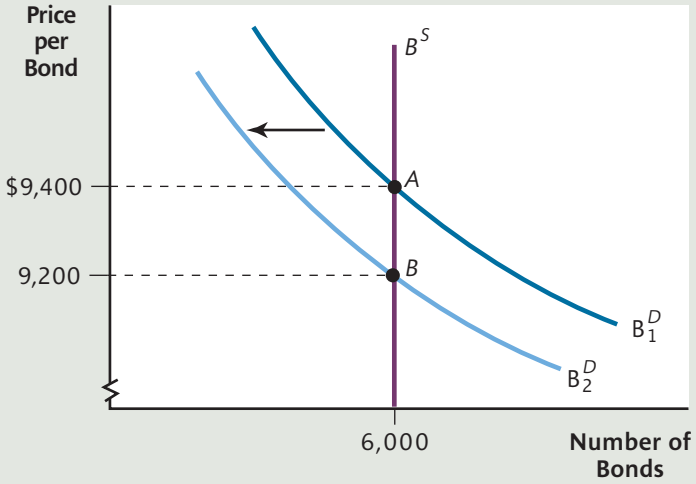
Suppose that the interest rate for riskless lending is 5 percent. Then the present value of this bond is $10,000 / 1.05 = $9,524. As you can verify, anyone who paid $9,524 for this bond would enjoy a yield of 5 percent—*if* it pays what it promises.

But because there is some risk that Verizon will default, most people would prefer the riskless alternative to owning this bond for a 5 percent yield. They will only want to hold this bond at a price *less* than $9,524 (providing a yield *greater* than 5 percent). Someone who is very averse to risks, or who thinks that Verizon bonds are particularly risky, might not want to buy it for any price greater than $9,000 (a yield of $1,000 / $9,000 = 11.1%). Someone else, who doesn't mind the risk as much, or who thinks Verizon bonds aren't that risky, would pay a price closer to the bond's present value of $9,524 (for a yield closer to 5%). Because people differ in their opinions and attitudes, each will have a different price at which they are willing to own this bond. Thus, the lower the price, the more bonds in total people will want to own.

You can see in the figure that at any price greater than $9,400, the quantity of bonds people *are* holding (the number in existence, given by the supply curve) would exceed the number that people *want* to hold (on the demand curve). There would be an excess supply of this bond, and people would try to sell it. This would drive the price down to $9,400 at which point people would be willing to hold all 6,000 bonds in existence. At any price lower than $9,400 people want to hold *more* of these bonds than they are holding, and they will drive the price up to acquire them.

FIGURE 4 A Decrease in Demand for Verizon Bonds

When the demand for one-year Verizon bonds decreases, the equilibrium moves from point A *to point* B*, and the price drops to $9,200. Because the bonds still pay $10,000 in one year, the annual yield is now $800/$9,200 = 0.087 or 8.7%*



When the market is in equilibrium at a price of $9,400 this one-year Verizon bond will pay interest of $10,000 − $9,400 = $600. Anyone who buys it will have a yield of $600/$9,400 = 0.064 or 6.4 percent.

Bond prices achieve their equilibrium value almost instantly. It takes just a few seconds for people to call in an offer to sell or buy bonds when there is an excess supply or demand. Only at the equilibrium price will every seller find a buyer and every buyer find a seller, so that every bond will be willingly held.

What Happens When Things Change

Most bonds' prices change every day, and even minute by minute. Because bonds virtually always trade at their equilibrium prices, and because the supply curve shifts only rarely (when new bonds are issued or old ones are retired), it must be changes in demand that cause these frequent price changes.

Figure 4 shows an example: a leftward shift in the demand curve for Verizon bonds, from B_1^D to B_2^D. As a result of the shift, the new equilibrium price drops to $9,200. At this price, the annual interest payment is now $10,000 − $9,200 = $800, so the yield is $800/$9,200 = 0.087 or 8.7 percent.

What might cause the demand curve to shift leftward, as in Figure 4? Some important reasons are:

- *An increase in the (riskless) interest rate.* This raises the annual interest rate used in calculating the *PV* of every bond and decreases their present values. With lower *PV,* but the same risk, people will not want to own fewer bonds at any price.
- *An increase in the attractiveness of other assets* (such as stocks, real estate, or *other* bonds). This makes people want to hold more of their wealth in these other assets and want to hold fewer of the bonds we're analyzing at any price.
- *An increase in the perceived riskiness of the bond.* Risk makes the value of a bond fall short of its *PV*. The greater the risk, the greater the shortfall. At any given price, people will want to hold fewer bonds.

- *Expectations of any of the above*. If people expect any of the above events to occur in the future, they will expect the demand curve B^D to shift leftward as well, and expect the price of the bond to drop. As a result, people will want to hold fewer bonds at any given price *now*.

An Example: Bond Prices in September 2008

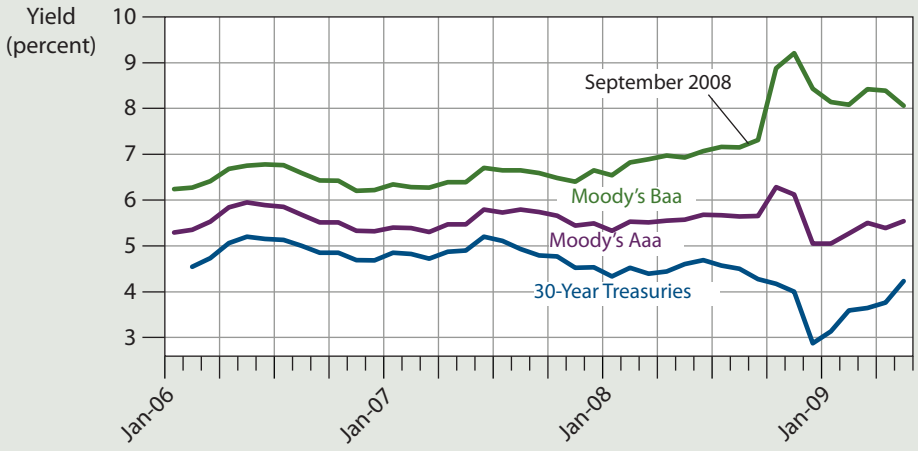
An example of a sudden and dramatic change in bond prices occurred in September 2008. Recall from Chapter 4 that by this time, housing prices had begun collapsing, and many people began defaulting on their mortgage loans. Banks and other financial institutions that had themselves borrowed money to lend it to homebuyers, and relied on mortgage payments to pay back *their* loans, were suddenly in danger of going bankrupt. At the same time, the economy was in the midst of a severe recession. With incomes and spending declining, losses mounted at businesses across the country. Many firms—retail chains, manufacturers, and more—were teetering on the edge of bankruptcy. The public, needless to say, was nervous about corporate bonds.

Then, in September of 2008, several events happened that shook confidence even further. One of the most important was the bankruptcy of a major financial institution: Lehman Brothers. The holders of more than $100 billion in outstanding Lehman bonds suddenly realized they would receive far less than they had been promised. Another financial institution—American International Group—was saved from bankruptcy only by a last-minute government takeover. Events were unfolding rapidly, and everyone wondered which corporations would be the next to fall.

People began to perceive an increase risk of default on corporate bonds of all types. Figure 5 shows what happened. The two top lines show yields on two different types of 30-years corporate bonds from 2006 through mid-2009. The top line is for bonds rated "Baa" (moderately risky) by Moody's bond rating service, and the middle line is for those rated "Aaa" (highest quality).

Whatever Moody's and the other ratings agencies were saying at the time, people seemed to believe that these bonds had suddenly become more risky. In September 2008, demand curves for both types of bonds shifted leftward (as illustrated for

FIGURE 5 Yield-Spreads for 30-Year Bonds, 2006–2009



Source: U.S. Federal Reserve

The yield on thirty year bonds depends on the risk of default. Those rated Baa are more risky and have higher yields than those rated Aaa, which in turn have higher yields than the least-risky 30-year Treasury bonds. Differences in yields are called "spreads." In September 2008, when fears of major corporate bankruptcies arose, the spread between more-risky and less-risky corporate bonds rose. So did the spread between corporate bonds of all types and safe Treasury bonds.

Verizon bonds in Figure 4), so their prices dropped and their yields rose. Notice the especially dramatic increase in the more risky Baa-rated bonds in Figure 5.

Economists sometimes focus on the difference in yields between two types of bonds—such as the Baa and Aaa bonds in the figure. This difference in yields is called the **spread**, which in the figure is the (changing) distance between the two upper lines. The spread rose in September 2008. Then, even as the panic subsided, and yields on both types of bonds began to come down, the spread remained high. A rising spread between more-risky and less-risky corporate bonds is sometimes called a "flight to quality."

Finally, look at the bottom line, which tracks yields on 30-years U.S. Treasury bonds. This yield actually *dropped* as the panic hit, and continued to drop for several months. As other assets—corporate bonds, stocks, real estate, and more—were viewed as more risky, these safe government bonds suddenly looked more attractive. Demand for safe government bonds increased, their prices rose and their yields fell.

Now look at the difference between yields on government bonds and yields on corporate bonds of either type. As you can see, these spreads remained elevated through much of 2009. A rising spread between the yields on corporate assets and government bonds is sometimes called a "flight to safety," because government bonds are considered the safest assets of all. As the figure suggests, for several months after the panic of September 2008, investors had fled to safety, and many were still refusing to take the flight back.

Flights to quality or safety—the growing spreads in the figure—can be costly to the economy. Remember that when a firm wants to raise funds for investment purposes, it issues new financial assets—such as bonds—in the primary market. But the yields on these new bonds must be the same as those in the secondary market. Thus, higher bond yields increase firms borrowing costs for new investment projects. And when spreads increase, it is the newest, most innovative (and therefore riskiest) investment projects that, faced with such high borrowing costs, are most likely to be cancelled.

The Stock Market

Share of stock A share of ownership in a corporation.

A **share of stock**, like a bond, is a financial asset that promises its owner future payments. But the nature of the promise is different. When a corporation issues a bond, it is *borrowing* funds and promising to pay them back. But when a corporation issues a share of stock, it brings in new ownership of the firm itself. In fact, a share of stock is, by definition, a *share of ownership* in the corporation. Those who buy the shares provide the firm with funds, and, in return, they are entitled to a share of the firm's future profits.

Why Do People Hold Stock?

Why do so many individuals and fund managers choose to put their money into stocks? You already know part of the answer: When you own a share of stock, you own part of the corporation. Indeed, the fraction of the corporation that you own is equal to the fraction of the company's total stock that you own. For example, in June 2009, Starbucks Corporation had 736 million shares outstanding. If you owned 7,360 shares of Starbucks stock, then you owned 7,360 / 736,000,000 = .00001, or about one-thousandth of 1 percent of that firm. This means you are, in essence, entitled to a thousandth of a percent of the firm's after-tax profit.

In practice, however, most firms do not pay out *all* of their profit to shareholders. Instead, some of the profit is kept as *retained earnings,* for later use by the firm. The part of profit that is distributed to shareholders is called **dividends.** Of course, as a part owner of a firm, you are part owner of any retained earnings as well, even if you will not benefit from them until later.

Dividends Part of a firm's current profit that is distributed to shareholders.

Aside from dividends, a second—and usually more important—reason that people hold stocks is that they hope to enjoy **capital gains:** the return someone gets when they sell an asset at a higher price than they paid for it. For example, if you buy shares of Hewlett Packard at $30 per share, and later sell them at $35 per share, your capital gain is $5 per share. This is in addition to any dividends the firm paid to you while you owned the stock.

Capital gain The return someone gets by selling a financial asset at a price higher than they paid for it.

Some stocks pay no dividends at all, because the management believes that stockholders are best served by reinvesting all profits within the firm so that *future* profits will be even higher. If the firm uses this money well, then future profits (and future dividends) can be even greater. And in the meantime, higher profits raise the price of the stock so that shareholders can get capital gains when they sell it.

Until 2003, Microsoft had never paid a dividend. But by plowing its profits back into the company, the firm's shares grew to a total value of almost $300 billion in mid-2003. The company's shareholders had great faith that they would eventually get cash from the firm, and in March 2003, it happened: Microsoft paid its first dividend.

Valuing a Share of Stock

Let's begin our attempt to value of share of stock with the simplest possible case. Imagine we know, with certainty, that a corporation's after-tax profit will be $10 million per year, every year. In that case, the value of a share should equal the present value of those future after-tax profits. (We use after-tax profits because this is what belongs to the firm's shareholders, whether they receive them as dividends or not. Any profits *not* received as dividends are plowed back into the firm on *behalf* of the shareholders.)

But the present value calculation, even in this simple case, is a little different than for a bond. A bond has a known maturity date, after which no further payments are made. A share of stock, by contrast, continues to entitle the owner to a share of the corporation's after-tax profit as long as the corporation exists. This is essentially forever, unless the firm is anticipated to go out of business.

Fortunately, there is a formula for calculating the present value in a case like this.

> *If a firm will earn a constant $Y in profit after taxes each year forever, then the total present value of these future profits is $Y/r, where r is the discount rate.*[3]

Let's apply this formula to our example of $10 million in after-tax profit forever. If the discount rate—the interest rate you could earn on riskless lending—is 10 percent, the formula tells us that the *PV* of those future profits is $10 million / 0.10 = $100 million.

What about the value of a single *share* of this firm's stock? If there are 1 million shares of stock outstanding for this firm, then each share should be worth one

[3] There are other present value formulas for more complicated earnings forecasts, such as earnings that are expected to *grow* at a constant rate, or grow for some period and then stabilize, and more. If you go further in your study of economics or finance, you will learn some of them.

one-millionth of the firm's total value. So each share's price should be $100 million / 1 million = $100.

Now, our example was very simple in many respects. For one thing, profits at most firms do not remain constant year after year. There are other present value formulas for profit that is expected to *grow* at a constant rate, or grow for some period and then stabilize, and more. If you go further in your study of economics or finance, you will learn some of them. What's important here is that, regardless of which *PV* formula is called for,

> *if there were no uncertainty, the value of a share of stock would equal the total present value of its after-tax profit, divided by the number of shares outstanding.*

What about the value of a single *share* of this firm's stock? If there are 1 million shares of stock outstanding for this firm, then each share should be worth one one-millionth of the firm's total value. So each share's price should be $100 million / 1 million = $100.

> *Ignoring risk, the value of a share of stock in a firm is equal to the total present value of the firm's after-tax profit divided by the number of shares outstanding.*

Explaining Stock Prices

The prices of most stocks change every minute, and sometimes make large jumps in seconds. Why? Like all asset prices, stock prices are determined by supply and demand. However, as with bonds, our supply and demand curves require careful interpretation.

Figure 6 presents a supply and demand diagram for the shares of FedEx. The supply curve tells us the quantity of FedEx shares *in existence* at any moment in time. This is the number of shares that people are *actually* holding.

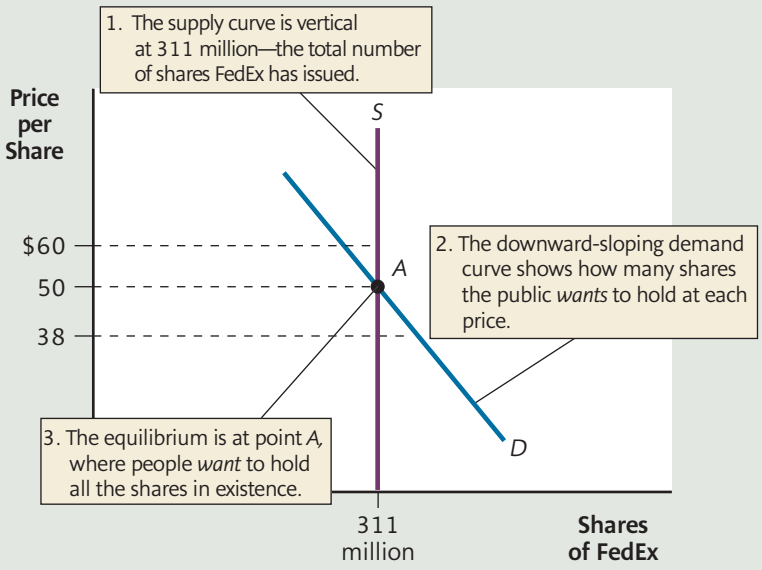
On any given day, the number of FedEx shares in existence is just the number that FedEx has issued and sold previously. Therefore, no matter what happens to the price today, the number of shares remains unchanged. Through June, 2009, FedEx had issued 311 million shares, so the supply curve is a vertical line at 311 million.

The quantity of FedEx shares that people want to *own* is given by the downward-sloping demand curve. As you can see, the lower the price of the stock, the more shares of FedEx people will want to hold. Why is this?

As discussed earlier, if FedEx's future profits per share were known by everyone with certainty, then everyone would place the same value on a share: the present value of those future profits. Everyone would want to own a share if they could buy it for any price equal to or less than that present value. This is a good starting point for valuing a share of stock.

The real world, of course, is more complicated. For one thing, people disagree about what FedEx's future profits are likely to be. If you think FedEx will be highly profitable in the future, you'll calculate a higher present value than someone who thinks FedEx's best days are behind it. Thus, even if everyone were certain about their own predictions, they would each calculate different present values for the shares, and each would have a different price at which they would be willing to own the stock.

FIGURE 6 The Market for FedEx Shares



Second, aside from different expectations of future profits, people may differ in their views about how *risky* those profits will be, and in their *attitudes* toward that risk. Because shares are risky, people will ordinarily hold them only if they can buy them for a price *less* than their basic present value. But for each person, the price that makes the stock attractive will be different.

Other factors, too, can influence the prices people are willing to pay, such as how many shares of FedEx they own already, what *other* assets they hold, how much wealth they have, and more. Clearly, the prices at which people want to own this stock will be different for different people. As the price of FedEx shares falls, and it becomes attractive to more potential investors, more of them will want to own it. This is what the downward sloping demand curve tells us.

In Figure 6, you can see that at any price other than $100 per share, the number of shares people *are* holding (on the supply curve) will differ from the number they *want* to hold (on the demand curve). For example, at a price of $38 per share, people would want to hold more shares than they are currently holding. Many would try to buy the stock, and the price would be bid up. At $62 per share, the opposite occurs: People find themselves holding more shares than they want to hold, and they will try to get rid of the excess by selling them. The sudden sales would cause the price to drop. Only at the equilibrium price of $50, where the supply and demand curves intersect, are people satisfied holding the number of shares they are *actually* holding.

As with bonds, stocks achieve their equilibrium prices almost instantly. Legions of stock traders—both individuals and professional fund managers—sit poised at their computers, ready to buy or sell a particular firm's shares the minute they feel they have an excess supply or a shortage of those shares. Thus, we can have confidence that the price of a share at any time is the equilibrium price.

What Happens when Things Change?

Why do stock prices *change* so often? Or, since stocks sell at their equilibrium prices at almost every instant, we can ask: Why do shares' *equilibrium* prices change so often?

Since a supply curve, like that in Figure 6, only shifts when the firm issues new shares (an infrequent occurrence), the day-to-day changes in equilibrium prices cannot be caused by shifts in the supply curve. So they must be caused by shifts in *demand* curve. Figure 7 shows how a rightward shift in the demand curve for shares of FedEx could cause the price to rise to $70 per share. Indeed, on rare occasions, the demand curve for a firm's shares has shifted so far rightward in a single day that the share price doubled or even tripled.

But what causes these sudden shifts in demand for a share of stock? Here are some important examples, each causing the demand curve to shift *rightward:*

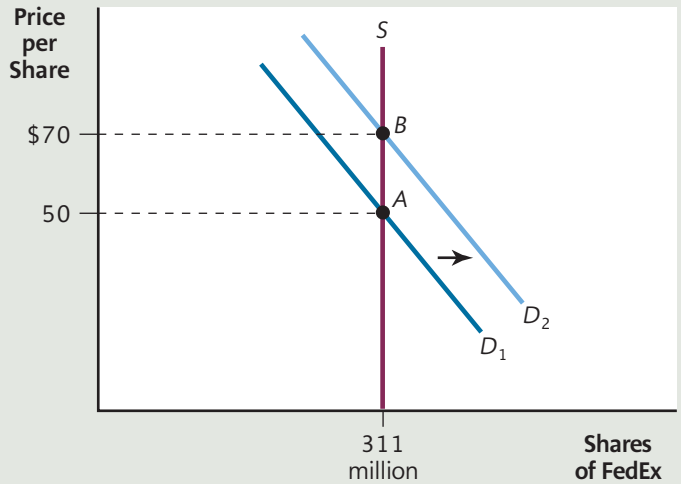
Release of New Information. New information that suggests future profits will be higher than previously anticipated will increase the present value of a share. Remember that present value is our *starting point* for valuing a share. But if other factors (e.g., risk) remain unchanged, a greater present value will increase the number of shares they want to own at any price.

Drop in Interest Rates. With lower interest rates—including the interest rate for riskless lending—the discount rate used in present value calculations will be lower as well. This increases the present value of any given share and—all else equal—makes people want to own more of them at any price.

Other Assets Less Attractive. All else equal, when other assets (bonds, real estate, commodities like gold and silver, or *other* firms' shares) become less attractive, people want to have more of their wealth in the shares we are analyzing, increasing demand.

FIGURE 7 An Increase in Demand for Shares of FedEx

With 311 million FedEx shares available, and demand curve D_1, the equilibrium price is $50 per share. When demand for FedEx shares increases (to demand curve D_2), an excess demand is created at the original price: People want to hold more than the 311 million shares they are actually holding. Their efforts to buy more shares drive up the price, until a new equilibrium at $70 is reached. At this new, higher equilibrium price, the quantity of shares demanded is once again equal to the 311 million available.



Decrease in Risk. With less uncertainty surrounding future profits, the price that makes a share attractive to potential owners will rise closer to its present value. (Remember: we calculate present value assuming no uncertainty.) This implies that, at any given price, less uncertainty makes people want to own more shares.

Expectations of Any of the Above. If people expect any of the above changes to occur in the future, they will also expect the demand curve for shares to shift rightward, and the price of the stock to rise. The anticipation of capital gains in the future (when the price rises) makes people want to hold more shares at any given price *now.*

An Example: Stock Prices in September 2008

The financial panic that came to a head in September 2008, discussed earlier for corporate bonds, also affected the stock market. It illustrates some of the factors that can shift the demand curve for shares—in this case, causing a leftward shift.

Specifically, during the crisis, new economic data was released daily, suggesting that the ongoing recession was worsening, and that future corporate profits would plummet. Even worse was the risk of widespread corporate bankruptcies. When a firm goes bankrupt, stockholders are last in line to receive any value that might be left. When General Motors declared bankruptcy in 2009, for example, its stock became worthless.

These events shifted demand curves for most corporate shares leftward, and their equilibrium prices dropped. Figure 8 gives an idea of how dramatic the drop was. It shows the behavior of the Standard and Poor's 500 (S&P 500), which tracks the average price of 500 of the largest U.S. corporations. As you can see, stock prices plunged downward beginning in mid-September. Over the next five weeks alone, the S&P 500 (as well as other, broader indexes) fell by 30 percent, and corporate shares lost more than $2 trillion of their value.

The Efficient Markets View

Every day, financial news programs on cable stations like CNBC or Fox Business Network offer advice to millions of television viewers. The analysts interviewed on these shows tell us that they have done some careful research, or that they have a secret formula, and that by following their advice, you'll earn more dividends, interest, and capital gains than you could hope to earn on your own. Of course, for the *really* good predictions, you'll have to pay a price and subscribe to their private newsletter or use them as your money manager.

Economists believe that analysts can often explain stock and bond price movements in the *past*. But they are extremely skeptical about anyone's ability to *predict* such price changes in advance. This is because economists tend to take the **efficient markets** view of markets in which assets are widely traded. According to this view, financial markets digest new information that might affect asset prices *efficiently,* that is, rapidly and thoroughly. To see what this view means, let's consider its implications for the stock market.

Efficient market A market that instantaneously incorporates all available information relevant to a stock's price.

The Meaning of an Efficient Stock Market

The efficient markets view implies that you cannot, on average, beat the market by doing research and finding and buying underpriced (or selling overpriced) stocks. You cannot do this because any research that *you* do will also be done by others and

FIGURE 8 The S&P 500 Stock Market Index, 2006–2009

The S&P 500 Index is a popular stock market index, which tracks share prices for 500 large corporations. In September 2008, as expected corporate profits fell and uncertainty rose, share prices plummeted. The S&P 500 dropped 30 percent in about a month.

is therefore already incorporated into the stock's price. That means—if the goal is to outperform a broad stock market average like the Standard & Poor's 500—you are largely wasting your time.

For example, suppose that you spend a lot of time investigating the XYZ corporation, which is experimenting with a new, patented technology. If successful, it will generate enormous profits. Every day, you look at press clippings, search Web pages, and learn all about this new technology. You even do some complicated statistical work to estimate the potential profits the technology could earn for XYZ, Inc.

One day, XYZ, Inc., reports that a very preliminary experiment was promising. Only very savvy people like you could read the details and come to the proper conclusion: The chances that the new technology will work have just doubled. Instantly, you go to the Web and look up the stock's price . . . only to find that it has already jumped higher. Why? Because there were hundreds of others just like you—some of them professionals managing billions of dollars of other people's wealth—who did the same research you did. They all have recognized the likelihood of greater future profit, and have attached a greater present value to this corporation's stock. By the time you go to buy shares, the demand curve has already shifted rightward.

But wait. *Someone* had to get there first. Surely *that* person benefited from their research, right? Actually, no. Assuming that the news was a surprise, only those lucky enough to be *already* holding the stock at the time of the announcement will

benefit. They will immediately adjust their asking price upward, so no one will be able to buy at the lower, preannouncement price.

The same is true of strategies to profit from *patterns* in stock trading, often called *technical analysis*. In the efficient markets view, this is a waste of time as well. Imagine a very simple pattern: Because of exuberance or fatigue or superstition, people are more likely to buy stocks than to sell them on Friday, so on average, stock prices rise every Friday. Since everyone would anticipate this pattern, they would buy stocks on Thursday, hoping to profit from the Friday runup. But this would cause stocks to rise on Thursday, not Friday, so people would buy on Wednesday, and so on. Soon, there would be no patterns at all; every day would be like any other day. Even though this is a very simple example, the logic applies to *any* pattern an analyst might uncover.

> *According to efficient markets theory, any information that helps one predict the future price of a specific stock or the market in general is instantly incorporated into stock prices. The only ones who benefit are those who are lucky enough to be holding the stock before the information became available.*[4]

The theory of efficient markets is one of the most exhaustively tested theories in all of economics. Thousands of studies have confirmed the efficiency of stock prices with respect to all sorts of information.

Common Objections to Efficient Markets Theory

When students first learn about efficient markets theory, some objections often come to mind. Here is how economists often respond to them.

Efficient markets theory can't be true. Otherwise, why would financial institutions spend so much money doing research on particular stocks, or on the market in general? The cynical response to this objection is that many financial institutions earn their income from commissions when their clients buy and sell stocks. Frequent research reports on stocks—"what's hot and what's not"—help these institutions convince their clients to trade more frequently.

A less cynical response draws on what you've learned about perfectly competitive markets in Chapter 8. Because there are no artificial barriers to doing stock market research and trading for profit, efficient markets theory assumes individuals and firms will continue to enter the market until the profit is reduced to zero.

But remember: It is *economic* profit that is reduced to zero through entry. In long-run equilibrium, stock researchers would earn just enough *accounting* profit to compensate for the implicit costs of their activities, such as foregone income. If it takes some special talent or ability to do this kind of research, then annual accounting profit would just cover the salary that people with these special talents or abilities could earn elsewhere. Thus, there is room to justify a positive salary for the thousands of professionals who analyze the market.

However, these professionals manage enormous amounts of other people's wealth. If they beat the market by only a tiny—almost imperceptible—fraction, they have covered their salary. For example, suppose a professional managing a portfolio

[4] One exception to this rule is *insiders*—those with connections to the firm and access to information *before* it becomes public. They can buy or sell stock early, before information is reflected in the price of the stock. Profiting from insider information is illegal. Those who do so, if they are caught, pay stiff fines and sometimes even go to jail. However, enforcement of insider trading laws is difficult, since it is often hard to detect.

of $3 billion in wealth can beat the market averages by 0.01 percent per year. That tiny edge would generate $300,000 in extra revenue for his or her firm—enough to justify up to that much in salary.

Now consider someone managing his or her own portfolio of, say, $100,000. Once entry by professionals has reduced the return to doing research down to 0.01 percent, this individual investor—doing his or her own, full-time research, or using the research of others—increases his or her earnings by only $10 per year.

Efficient markets theory can't be true. I just saw a money manager interviewed on CNBC, and he beat the market by more than 5 percentage points, three years in a row! Efficient markets theory does not rule out luck. With thousands of professionally managed stock funds, we would expect some of them to be unusually lucky even if they operated by sheer guesswork. And we would expect some small number to do so many years in a row. An unusually good performance by a few—in a process involving many—does not prove any actual skill.

To prove this to yourself, imagine that 10,000 investment managers each picked their stocks by throwing darts at the stock page, and timed their buying and selling by flipping a coin every day. If we ignore trading commissions, this group would, on average, perform about as well as the stock market as a whole. But through luck alone, *some* would be very successful in any given year—beating the market averages significantly 5 percentage points or more.

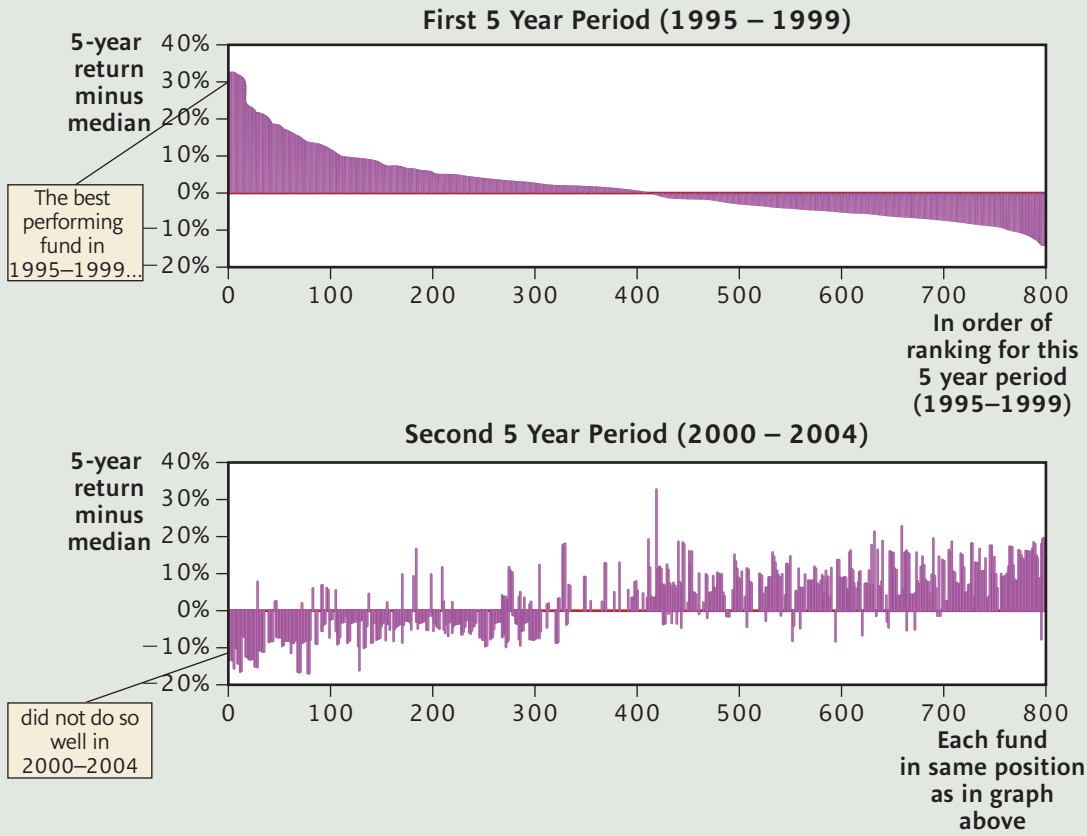
Let's say that the first year, by chance alone, 10 percent of these managers significantly beat the market. That's 1,000 managers. The second year, 10 percent of this group—or 100 managers—would outperform the market again. And the third year, 10 percent of that group—or 10 managers—would do it again. These 10 managers—especially if interviewed on CNBC—would appear to have some special skill or talent at picking stocks or timing buys and sells. But remember: In our thought experiment, they achieved their success by chance alone. The fourth year, each of them would be no more likely to beat the market than would any other randomly selected manager.[5]

This is not just academic theory—it is supported by the data. Figure 9 provides a startling example. It shows the performance of 800 professional stock funds during two five-year periods. In the upper panel, the funds are ordered by their performance for the period 1995 to 1999, with the best performing on the left. The height of the graph shows the annual rate of return for a fund minus the annual return for the median fund. For example, from 1995 to 1999, the most successful (leftmost) fund earned about 33 percentage points more per year than did the median fund. The worst (rightmost) fund did about 12 percentage points worse per year than the median.

Now look at the lower panel, where each of the funds occupies the same position along the horizontal axis as it did in the upper panel. But the lower panel shows performance over the *next* five years, from 2000 to 2004. As you can see, there was no tendency for a top performing fund in the first period to be among the top in the next period. This is just what we'd expect if the performances in the first period were based on chance alone. (In fact, this data shows a tendency for top performers to drop to the bottom, and bottom performers to rise to the top. But that odd result may be due to chance as well.)

[5] A brief personal note: In December 2008, after one of us finished a lecture about efficient markets theory, a perplexed student came up after class. While he accepted the ideas in theory, he and his family had clearly found an exception: An investment manager who had greatly outperformed the market every year for more than a decade. The student asked if he could prove this by bringing a prospectus and some quarterly statements to office hours. Two days later, he e-mailed that he would not need to keep the appointment. The investment manager was Bernard Madoff, who had been arrested the night before for fraud—most notably, for creating fictitious investment returns.

FIGURE 9 **Lack of Persistence in 5 Year Performance of 800 Mutual Fund Managers**



In the upper panel, the graph shows the percentage points by which the returns of each fund manager exceeded the returns of the median fund, during the 5-year period 1995-1999. They are ordered from best performer to worst performer during those five years. For example, the best performing fund earned a 5-year return that was 32 percentage points greater than the median fund's return. In the lower panel, the funds are located in the same horizontal positions as in the upper panel. But now the graph shows by how much each fund beat the median during the subsequent 5-year period (2000 - 2004). Comparing panels, the better performing funds in the first five-year period were not likely to be the better performing funds in the second 5-year period.

Source: Index Funds Advisors, Inc., http://www.ifa.com/12steps/step5/step5page2.asp, accessed on June 17, 2009. Return is the number of percentage points by which each fund beat the median for each 5-year period.

I've heard that economists no longer believe in efficient markets theory, because of asset price bubbles. There has been some confusion about whether the efficient markets view still makes sense in light of economic research on asset price bubbles—episodes of exaggerated price spikes not connected to any fundamental change in underlying values. The confusion is partly based on different interpretations of the phrase "efficient markets."

The strongest interpretation—one that few economists hold to—is that markets always determine the "correct price" for an asset, one that accurately reflects the benefits and risks of owning it at the time. The mere fact that asset prices sometimes change rapidly—and that people gain and lose fortunes—does not by itself prove that the earlier prices were "incorrect." Prices are based on forecasts, and even a

good forecast may not come true. If prices for particular stocks (or stocks in general) change rapidly, it could be because new information has changed the forecast.

Still, few economists today adhere to this "correct price" interpretation of efficient markets. Asset markets, including the stock market, *do* seem to experience occasional bubbles. There is a vigorous debate among economists about what causes them, and their implications for policy. Research in the field of *behavioral finance* suggests that various forms of herd behavior—some rational and some not—can cause bubbles to form, and exaggerate those that start for other reasons. In any case, it is widely agreed that, in the midst of a bubble, markets are not delivering the "correct" price.

But research continues to justify, and most economists still adhere, to the more *modest* interpretation of efficient markets theory we've been stressing: that markets for widely traded assets digest information rapidly, and that individual investors cannot systematically outperform the market averages.

That last statement might surprise you, in light of our discussion of bubbles. After all, isn't it easy to spot asset bubbles? And don't they always burst? So can't you "beat the market" just by betting that prices will come down? The answers to these questions, in order, are: No, yes, and not necessarily.

First, while an asset bubble may seem obvious after it has burst, it may *not* be so obvious while it is growing. There is always the possibility that asset prices are soaring because of a change in long-term fundamentals rather than temporary herd behavior. For example, in the 1980s, many market analysts believed that the rapidly growing price-differential between coastal and inland homes was a speculative "coastal home" bubble, which would soon burst. But the premium for coastal real estate remains as strong as ever. In retrospect, it appears to have been due to long-term market and institutional forces, rather than short-term herd behavior.

Second, even if you *could* identify a bubble in progress, you might not be able to profit from its inevitable bursting. You'd also have to correctly forecast *when* it will burst. Fortunes have been lost by smart investors whose timing was off.

For example, one common strategy to profit when asset prices come down is to "sell short." In the stock market, for example, you sell shares you don't actually own while their price is high, and commit to buy and deliver them later after their price has crashed. When you sell short, you must guarantee that you'll deliver by putting up some of your own money now. And there's the rub. If the bubble continues, and share prices keep rising, you have to put up more and more cash. If you run out of money, you'll have to close out your position and buy the shares at an even higher price than you paid. Thus, even if you correctly spot the bubble, you can lose a fortune. A quote often attributed to the famous economist John Maynard Keynes (1883–1946) summarizes this problem: "The market can stay irrational longer than you can stay solvent."

Efficient Markets Theory and the Average Investor

Although the efficient markets view rules out many investment strategies, that doesn't mean you shouldn't invest in the stock market at all. The average stock's price, over long periods of time, tends to rise. In fact, if dividends and capital gains are added together, stocks—over the very long run—have earned their holders a better yield than bonds. That's because stocks are more risky, and shareholders must be compensated for bearing that risk.

If you do invest in the stock market, efficient markets theory suggests some strategies that can help you do so more wisely.

First, knowing that stocks are more risky (one reason for their past higher returns), you should be cautious about putting all of your wealth in stocks. Be especially

cautious with funds you may need to access suddenly, such as to pay for college or graduate school, or to make a down payment on a home.

Second, make sure your stock market investments are diversified—consisting of different stocks that tend not to rise or fall together. A stock index fund that spreads your investments automatically among hundreds or thousands of shares in the index is a very low-cost way to do this.

Third, because you have to pay commissions when you trade stocks, you should trade as little as possible. Instead of worrying about what's-hot-and-what's-not every few weeks, use a "buy and hold" strategy that prevents commissions from eating into your returns.

Fourth, if someone asks you to pay for their stock-picking advice, don't. If you want the fun (and anguish) of investing in individual stocks, you can do just as well by picking them on your own, even if you pick them randomly. The stocks you pick will be as likely to rise or fall as stocks chosen by an expert.

All of the markets we have studied in this chapter enable us to save funds and earn a rate of return, and they enable firms to invest and grow. They help relax the economic constraints imposed by scarcity. And they certainly contribute to the high standard of living we enjoy. When savers and borrowers come together in financial markets, both sides benefit.

Using the Theory

© TOM STEWART/CORBIS

THE PRESENT VALUE OF A COLLEGE DEGREE

Previously in this book, we've discussed some of the economic aspects of attending college. In Chapter 1, we analyzed the costs of college. In Chapter 12, we discussed one of the benefits of college: the *wage premium* that college graduates earn over high school graduates. Now, we'll compare these costs and benefits to ask: Is college a wise financial investment?

Traditionally, the answer has always been a resounding yes. But in recent years, tuition—one component of costs—has been growing rapidly. Has the answer changed? Let's see.

The decision to attend college is remarkably similar to the other sorts of decisions we've discussed in this chapter. Getting a college education is an investment in *human capital*. It involves purchasing an asset—a college degree—that is expected to yield a stream of future benefits, in the form of a greater annual income. The benefits will accrue over your working life. But the costs have to be paid sooner—over just four years. As with any other problem involving benefits and costs over various future periods, the right answer is found by comparing the *present value* of these benefits and costs.

The Costs of College

To calculate the present value of the *costs* of college, let's use the most expensive choice: a private college, with no scholarship or grant aid. (A student loan will not ordinarily change the present value of the costs; it will just redistribute when they are actually paid.)

Totaling the additional explicit costs (tuition, books, and other supplies) and the implicit costs (foregone earnings from not working during the academic year), the annual cost for a private college averages $44,197.

Since this cost is paid each year over four years, our next step is to compute the present value. We'll use 5 percent as an estimate of the rate for riskless lending. And to make the problem simpler, we'll assume that each year's costs are paid at the end of each year (our rule for approximating present value), and that the costs of college remain constant over the period. (In the end-of-chapter problems, you'll be asked to re-do the calculations under different assumptions.)

Using our simplifying assumptions, the present value of the cost of four years of college is

$$PV \text{ of cost of private college}$$

$$= \frac{\$44{,}197}{1.05} + \frac{\$44{,}197}{(1.05)^2} + \frac{\$44{,}197}{(1.05)^3} + \frac{\$44{,}197}{(1.05)^4} = 156{,}720$$

In words, with a 5 percent discount rate, the costs you will pay over the four years to attend college are equivalent to paying $156,720 today. Therefore, if the present value of your future benefit exceeds $156,720, and if you can be relative sure about this, then you are getting a "good deal." That is, the return on your investment will be higher than for an alternative investment paying 5 percent per year (such as putting $156,720 into Treasury bills).

The Financial Benefits of College

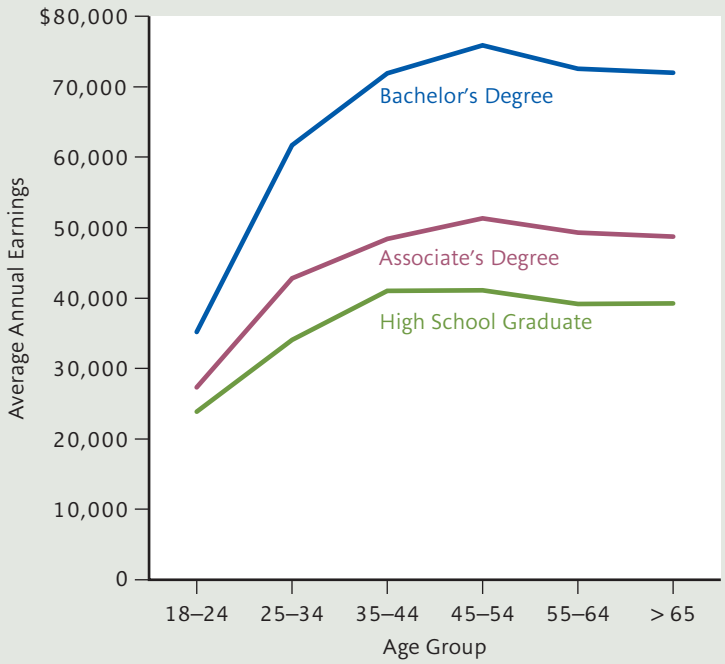
Studies of income by age and education show that (1) earnings generally increase with age (and therefore experience) for all education groups; (2) earnings increase with education for all age groups (once the education is completed); and (3) earnings rise with age more sharply for those with more education.

Figure 10 illustrates all three of these observations. It shows the average earnings of full-time workers in different age groups in 2007, with each line corresponding to a different level of education.

FIGURE 10 Age-Earnings Profiles in 2007

Income tends to rise with both education and age. As you can see in the graph, within each age group, more education is associated with greater annual earnings. And those with more education enjoy faster growth in earnings over their first few decades of work.

Source: U.S. Census Bureau, Table P-32 Educational Attainment—Full-Time, Year-Round Workers 18 Years Old and Over by Mean Earnings, Age, and Sex: Weighted average of men and women was calculated by authors.



As you can see in the figure, high school graduates (the lowest line) have the lowest earnings at *all* age groups. Within any age group, the greater the education, the greater the earnings. For example, in the 35–44 age group, the average high school graduate who worked full-time earned about $40,985, while the average college graduate (with no higher degree) earned about $71,900. Those who went on to earn professional degrees (not shown in the diagram) do even better at each age. In the 35–44 age group, their average earnings are $143,160.

These earnings in 2007 don't necessarily predict what will happen to today's graduates. But the general relationship between earnings and age or education has been present for decades and is expected to continue into the future.

One frequently cited statistic is that today's college graduate can expect to earn more than $1 million more over their working lifetime than a high school graduate. But remember: Most of the additional earnings from college will be received far into the future, when they are worth substantially less in present value terms. For example, with a 5 percent discount rate, one dollar in additional earnings received 30 years from now has a present value of only 23 cents. To properly evaluate the benefits, we must use the present value of these additional earnings.

As you can see in Figure 10, the *additional* earnings from attending college (the distance between the top and bottom lines) grow with time, starting at about $11,000 after graduation, and rising to about $34,000 during the peak earnings years. Using these numbers, and a 5 percent discount rate, we find (through tedious calculations) that the *increase* in earnings from college, over a typical working life, has a present value of about $423,000.

Comparing Costs and Benefits

We've seen that in present value terms, attending college gives expected benefits worth $423,000 today. For the private school option under the assumptions we've made, the *costs* of college have a present value of $143,860. Comparing the two, the benefits are substantially greater. College is a good investment.

How good? Interpreting these numbers in different ways can help us get some perspective. For example, we can ask: How much would college have to cost today to make it a poor investment? The answer is: about $423,000. If college costs, say, $430,000 in present value terms, you could expect a better lifetime financial result putting that amount into Treasury bills.

We can also ask: What is the annual rate of return provided by an investment in a college degree? Or, to put it another way, what rate of return would have to be available on an alternative investment to make it financially equivalent to obtaining a college degree? To answer, we must find the discount rate that would make the PDV of the benefits of college equal to the PDV of its costs. Once again, the calculations are tedious, but the result for our example (a private institution with no financial aid and the typical age–earnings differentials shown in Figure 7) is an annual average rate of return of 13 percent. Thus, on purely financial grounds, if you can find some other investment with a consistent annual rate of return greater than 13 percent over the next 40 or so years, with an equivalent degree of certainty, it would dominate an investment in a college degree. If you find such an investment, please let us know.

Some Provisos

Some of the assumptions in our example artificially inflate the investment value of a college degree. For example, we've assumed that none of the costs of college will rise over the four years—neither tuition nor the salary you could earn with a high

school degree. We've also assumed there is no risk in the returns to college. In fact, both the average wage premium, and the premium for any individual—are uncertain. Accounting for these factors would reduce the value of a college degree to some extent—but not by much.

On the other hand, some of our assumptions have led us to *under*state the value of a college degree for many students. For example, we've been assuming a high-tuition private college, with no grants or scholarships. Many students attend state or community colleges and/or receive scholarships.

But there are other, more theoretical questions about our analysis. It is certainly true that college graduates earn more at every age (after graduation) than high school graduates. But why? Do they *learn* things in college that make them more productive, and therefore more valuable, to a future employer? Or by graduating from college, do they "signal" to employers that they have certain characteristics (ability, intelligence, perseverance, social skills, organizational skills—all needed to get the diploma) perseverance—that make them worth a higher salary? In either case, college would still be a good investment, because the degree would be required to get the higher income.

But there could be another factor behind the data that could undermine this conclusion. Suppose that people of higher-than-average intelligence and greater-than-average initiative would earn higher incomes whether they attended college or not, just because of these traits. And what if these people disproportionately attend college because they have a *taste* for it—they like to learn and like the intellectual environment that college provides. In that case, we would still observe the patterns of Figure 10—people with college degrees would earn more—but not for the reasons we've suggested. Instead, people with higher earnings just *happen* to have more education because they like education. But they would have had those higher earnings with or without the degree. (Bill Gates, who dropped out of Harvard in his freshman year, is an extreme example.)

If this explained *all* of the higher earnings of college graduates, college would be a bad financial investment. You'd be paying a present value of perhaps $143,000 for future benefits of . . . zero (in strictly monetary terms). Labor economists have looked at this question in a variety of ways (including studies of identical twins).[6] The consensus is that only a small amount (10 percent in one study) of the higher earnings of college graduates can be attributed to causes other than the degree. So whether you are actually learning valuable skills in college, or just signaling to employers that you would be a productive employee, college remains a very good financial investment.

[6] For a good summary, see David Card, "The Causal Effect of Education on Earnings," in *Handbook of Labor Economics*, Vol. 3A (Orley C. Ashenfelter and David Card, editors.), pp. 1801–1863.

SUMMARY

A firm that rents its capital equipment can use the marginal approach to profit to decide how much to rent. The firm will increase its use of capital as long as the additional net revenue per period exceeds the rental cost of the capital. However, the decision to purchase, rather than rent, capital is more complex, requiring a comparison of future receipts with the current purchase price. Present value calculations help make these comparisons.

The principle of asset valuation tells us that—when there is completely certainty about future receipts—the value of any asset is the total present value of all the future income the asset will generate. With no risk, the firm should buy any unit of capital for which the total present value of all future years' net revenue is greater than the purchase price. This total present value will be smaller when interest rates are higher. Therefore, higher interest rates discourage investment in physical capital. When there is uncertainty, the value of physical capital will fall short of its present value by an amount that depends on the degree of risk and investor attitudes toward risk.

There are many types of financial markets, including those for bonds and corporate stock. Ignoring risk, the price of a bond will equal the total present value of its future payments. There is an inverse relationship between the price of a bond and its yield (rate of return). When there is risk of default, the price of a bond will be less than its present value, and its yield will be higher than under certainty. The price of any particular bond is determined by supply and demand. The supply curve is vertical, reflecting the fixed quantity of bonds at any time. The demand curve slopes downward, reflecting differences in beliefs and attitudes about risk among potential bondholders.

The value of a share of corporate stock, in the absence of uncertainty, would equal the total present value of the future after-tax profits of the firm, divided by the number of shares outstanding. With uncertainty about future profit, the value of a share to any individual will be less than the presented value. The price of a share, like the price of a bond, is determined by the intersection of a vertical supply curve and a downward-sloping demand curve.

Efficient markets theory suggests that the price of an asset, such as a share of stock, rapidly incorporates all available information that would enable someone to predict its future price. Therefore, research to help pick individual stocks or to detect trading patterns is not worth an investor's time or money. In an asset price bubble, stock prices can depart substantially from their underlying value, but this does not undermine the conclusions of efficient markets theory for individual investors.

PROBLEM SET

Answers to even-numbered Questions and Problems can be found on the text Web site at www.cengage.com/economics/hall.

1. You are considering buying a new laser printer to use in your part-time desktop publishing business. The printer will cost $380, and you are certain it will generate additional net revenue of $100 per year for each of the next five years. At the end of the fifth year, it will be worthless. Answer the following questions:
   a. What is the value of the printer if you could lend funds safely at an annual interest rate of 10 percent? Is the purchase of the printer justified?
   b. Would your answer to part (a) change if the interest rate were 8 percent? Is the purchase justified in that case? Explain.
   c. Would your answer to part (a) change if the printer cost $350? Is the purchase justified in that case?
   d. Would your answer to part (a) change if the printer could be sold for $500 at the end of the fifth year? Is the purchase justified in that case? Explain.

2. Your inventory manager has asked you to approve the purchase of a new inventory control software package. The software will cost $200,000 and will last for four years, after which it will become obsolete. If you do not approve this purchase, your company will have to hire two new inventory clerks, paying each $30,000 per year. Answer the following questions:
   a. Should you approve the purchase of the inventory control software if the relevant annual interest rate is 7 percent?
   b. Would your answer to part (a) change if the annual interest rate is 9 percent? Explain.
   c. Would your answer to part (a) change if the software cost $220,000? Explain.
   d. Would your answer to part (a) change if the software would not become obsolete until the last day of its sixth year?

3. Ice Age Ice is trying to decide how many $150,000 commercial ice makers to buy. Assume that each machine is expected to last for seven years. Complete the following table for an interest rate of 5 percent (and assuming no uncertainty). How many ice makers should Ice Age Ice purchase? How low would the price per machine have to fall before the firm would buy four ice makers?

Ice Machines	Additional Annual Revenue	Total Present Value of Additional Revenue over Seven Years
1	$26,000	
2	$25,000	
3	$16,000	
4	$12,000	
5	$ 6,000	

4. Your firm is considering purchasing some computers. Each computer costs $2,600, and each will add to your net revenue by known amounts. Because you plan to use the computers for different purposes, you have ranked those purposes in descending order of annual additional revenue as follows:

Computer	Net Additional Revenue per Year
1	$3,000
2	$2,000
3	$1,000
4	$ 500

   a. Assume that each computer has a useful life of three years, and no value thereafter. If the annual interest rate is 10 percent per year, how many computers should you purchase?
   b. If, before you purchased the computers, the interest rate dropped to 5 percent per year, how many computers would you purchase?

5. A drug manufacturer is considering how many of four new drugs to develop. Suppose it takes one year and $10 million to develop a new drug, with the entire cost being paid up front (immediately). The yearly profits from the new drugs will begin in the *second* year (with profits, as always, assumed to come at the end of the year.), and are given in the table below:

Drug	Annual Profit
A	$7 million
B	$5.5 million
C	$5 million
D	$4 million

   These profits, which are certain, accrue *only* while the drug is protected by a patent; once the patent runs out, profit is zero.
   a. If the annual interest rate is 10 percent and patents are granted for just two years, which drugs should be developed?
   b. If the annual interest rate is 10 percent and patents are granted for three years, which drugs should be developed?
   c. Answer (a) and (b) again, this time assuming the discount rate is 5 percent.
   d. Based on your answers above, what is the relationship between new drug development and (1) the discount rate; (2) the duration of patent protection?
   e. Would the relationships in d. still hold in the more realistic case where profits from new drugs are uncertain?
   f. Is there any downside to a change in patent duration designed to speed the development of new drugs? Explain briefly.

6. Good news! Gold has just been discovered in your backyard. Mining engineers tell you that you can extract five ounces of gold per year forever. Gold is currently selling for $400 per ounce, and that price is not expected to change. If the discount rate is 5 percent per year, estimate the total value of your gold mine.

7. One year ago, you bought a two-year bond for $900. The bond has a face value of $1,000 and has one year left until maturity. It promises one additional interest payment of $50 at the maturity date. If the interest rate is 5 percent per year, what capital gain (or loss) would you get if you sell the bond today?

8. Suppose a risk-free bond has a face value of $100,000 with a maturity date three years from now. The bond also gives coupon payments of $5,000 at the end of each of the next three years. What will this bond sell for if the annual interest rate for risk-free lending in the economy is
   a. 5 percent?
   b. 10 percent?

9. Suppose a risk-free bond has a face value of $250,000 with a maturity date four years from now. The bond also gives coupon payments of $8,000 at the end of each of the next four years.
   a. What will this bond sell for if the risk-free lending rate in the economy is 4 percent?
   b. What will this bond sell for if the risk-free lending rate is 5 percent?
   c. What is the relationship between the bond's price and the level of interest rates in the economy in this exercise?

10. Suppose that people are sure that a firm will earn annual profit of $10 per share forever. If the interest rate is 10 percent, how much will people pay for a share of this firm's stock? Suppose that people become uncertain about future profit. What would

happen to the price they would be willing to pay? (Your answer will be descriptive only.)

11. In the market for Amazon.com *bonds,* explain how each of the following events, *ceteris paribus,* would affect (1) the demand curve for the bonds, (2) the price and (3) the yield?
    a. Fitch upgrades the bond from AA to AAA.
    b. The interest rate on U.S. government bonds decreases.
    c. People *expect* the interest rate on U.S. government bonds to decrease, but it hasn't yet happened.

12. In the market for Amazon.com *stock,* explain how each of the following events, *ceteris paribus,* would affect the demand curve for the stock and the stock's price.
    a. The interest rate on U.S. government bonds rises.
    b. People *expect* the interest rate on U.S. government bonds to rise, but it hasn't yet risen.
    c. Google announces that it will soon start competing with Amazon in the market for books, DVDs, and everything else that Amazon sells.

13. Suppose that 1,000,000 people select which stocks to buy and hold for the year by throwing darts at the stock page. Suppose, too, that in any given year:
    - The average stock price rises by 7 percent.
    - In any given year, 50 percent of the dart throwers will have a return that is average or better.
    - In any given year and by luck alone, 20 percent of the dart throwers will "beat the average" by 5 percentage points or more.
    - In any given year and by luck alone, 10 percent of the dart throwers will beat the average by 10 percentage points or more.

    a. After five years, how many people will report that they've earned 7 percent (the market average) or more on their stocks in every one of the previous five years?
    b. After five years, how many people will report that they've earned 12 percent (5 percent above the average) or more on their stocks in every one of the previous five years?
    c. After five years, how many people will report that they've earned 17 percent or more on their stocks in every one of the previous five years?

14. State whether each of the following, with no other change, would *increase* or *decrease* the economic attractiveness of going to college, and give a brief explanation for each.
    a. A decrease in estimated working life.
    b. An increase in the earnings of the average high school student.
    c. Permanently higher interest rates in the economy.

15. In the Using the Theory section, we calculated the present value of attending a private college, under the assumption that the costs remain the same for each of the four years. Recalculate the present value of these costs under the assumption that while implicit costs remain the same, tuition and other *explicit* costs rise by 8 percent per year, starting with the second year's tuition and expenses. (Continue to assume that all costs are paid at the end of each year.)

16. In the Using the Theory section, we calculated the present value of attending a private college, under the assumption that each year's costs are paid at the *end* of the year. Recalculate the present value of these costs under the assumption that all costs are paid at the *beginning* of each year.

17. In the Using the Theory section, we calculated the present value of the costs of attending a private college, with no financial aid.
    a. Using Table 1 in Chapter 1, what would have been the present value of the cost of college if we had done the analysis for a four-year public institution, with no financial aid? (Hint: Assume that future income does not depend on the type of college.)
    b. What is the percentage difference between the present value of costs at a private versus a public institution?
    c. Private college tuition is about four times higher than public college tuition. Explain briefly why the (present value) cost figures do not differ this widely.

More Challenging

18. In the chapter, you learned that when future income from an asset is uncertain, the asset's value will be less than the present value of its future payments, because present value is calculated assuming these payments are certain. A different way to think about uncertainty is to calculate a *modified* present value, which uses a higher discount rate, equal to the *safe* lending interest rate *plus a risk premium*. In Problem 5, answer a through d again, assuming that (a) each interest rate provided in the problem is for riskless lending; and (b) the risk premium for developing new drugs is 20 percent.

$$\text{Value} = \frac{Y}{(1 + r_1)} + \frac{Y_2}{(1 + r_1)(1 + r_2)} + \frac{Y_3}{(1 + r_1)(1 + r_2)(1 + r_3)}$$

CHAPTER 14

Economic Efficiency and the Competitive Ideal

Throughout this book you've learned about the different types of markets in which products and resources are traded. Most of our discussion has been *descriptive:* For each type of market structure, we reached conclusions about price, quantity, firm profit, and how changes of various kinds affect the market equilibrium.

In this chapter, we will look at markets in a different way—*assessing* them in terms of their *economic efficiency.* As you'll soon see, there is more to the concept of efficiency than you might think.

We'll look first at the idea of economic efficiency itself: what it is and what it is not. Next, you'll learn how economists measure the gains from trading, and use this measure to gauge whether or not a market is efficient. You'll see why perfectly competitive markets are considered economically efficient, and why other market structures fall short of this ideal. We'll also explore how some government actions can *prevent* a market from achieving an efficient outcome.

Of course, the market—left on its own—is sometimes *not* efficient, and government action can be an appropriate remedy. We'll take up these situations, and government's role in fostering efficiency, in the next chapter.

The Meaning of Economic Efficiency

What, exactly, do we mean by the word *efficiency*? We all use this word, or its opposite, in our everyday conversation: "I wish I could organize my time more efficiently," "He's such an inefficient worker," "Our office is organized very efficiently," and so on. In each of these cases, we use the word *inefficient* to mean "wasteful" and *efficient* to mean "the absence of waste."

In economics, too, efficiency means the absence of waste, although a very specific kind of waste: *the waste of an opportunity to make someone better off without harming anyone else*. More specifically,

> *economic efficiency is achieved when all activities that can make at least one person better off without making anybody else worse off are taking place.*

Notice that economic efficiency is a limited concept. Even though it is an important goal for a society, it is not the only goal. Most of us would list fairness as another important social goal. But an efficient economy is not necessarily a fair economy. For example, an economy could, in theory, be efficient even if 99 percent of its income went to a single person—a situation that almost everyone would regard as unfair.

Economists are concerned about both efficiency and fairness. But they generally spend more time and energy on efficiency. Why? Largely because it is so much easier for people to agree about efficiency. We all define fairness differently, depending on our different ethical and moral views. Issues of fairness must therefore be resolved politically.

But virtually all of us would agree that if we fail to take actions that would make some people in our society better off *without harming anyone*—that is, if we fail to achieve economic efficiency—we have wasted a valuable opportunity. Economics—by helping us understand the preconditions for economic efficiency and teaching us how we can bring about those preconditions—can make a major contribution to our material well-being.

Pareto Improvements

Imagine the following scenario: A boy and a girl are having lunch in elementary school. The boy frowns at a peanut butter and jelly sandwich, which, on this particular day, makes the girl's mouth water. She says, "Wanna trade?" The boy looks at her chicken sandwich, considers a moment, and says, "Okay."

This little scene, which is played out thousands of times every day in schools around the country, is an example of a trade that makes people better off without harming anyone. And as simple as it seems, such trading is at the core of the concept of economic efficiency. It is an example of a *Pareto* (pronounced puh-RAY-toe) *improvement,* named after the Italian economist, Vilfredo Pareto (1848–1923), who first systematically explored the issue of economic efficiency.

*A **Pareto improvement** is any action that makes at least one person better off, and harms no one.*

Pareto improvement An action that makes at least one person better off, and harms no one.

In a market economy such as that in the United States, where trading is voluntary, literally hundreds of millions of Pareto improvements take place every day. Almost every purchase is an example of a Pareto improvement. If you pay $30 for a pair of jeans, then the jeans must be worth more to you than the $30 that you parted with or you wouldn't have bought them. Thus, you are better off after making the purchase. On the other side, the owner of the store must have valued your $30 more highly than he valued the jeans or he wouldn't have sold them to you. So he is better off, too. Your purchase of the jeans, like virtually every purchase made by every consumer every day, is an example of a Pareto improvement.

The notion of a Pareto improvement helps us arrive at a formal definition of economic efficiency:

***Economic efficiency** is a situation in which every possible Pareto improvement is being exploited.*

Economic efficiency A situation in which every possible Pareto improvement is being exploited.

This definition can be applied to any part of the economy, or to the economy as a whole. For example, if we want to know whether the U.S. market for airline travel is efficient, we would ask: Are there any possible Pareto improvements that are *not* happening in this market? If the answer is yes, then the market is *in*efficient. If the answer is no, then all Pareto-improving changes in this market *are* being exploited, so we would call it efficient.

When thinking about the economy as a whole, we realize that it is not possible to exploit *every* possible Pareto improvement. But, *ceteris paribus,* achieving something close to economic efficiency is desirable.

Side Payments and Pareto Improvements

So far, the Pareto improvements we've considered are easily arranged transactions. Because both parties come out ahead, they have every incentive to find each other and trade.

But there are more complicated situations in which a Pareto improvement will come about only if one side makes a special kind of payment to the other, which we call a *side payment*. These are situations in which an action, without the side payment, would benefit one group and harm another.

Here's a simple example. The owner of an empty lot wants to build a movie theater on her property. Many people might gain from the theater: the owner of the lot, moviegoers, the theater's employees, and more. But the residents in the immediate vicinity might be harmed, because the theater will bring noise and traffic congestion.

Imagine that we can measure the gains and losses for each person in the town in dollars. When we sum them up, the total benefits to the gainers are valued at $100,000 while the total harm to the losers is valued at $70,000.

Building the theater—by itself—would *not* be a Pareto improvement, because even though some would benefit, others would be hurt. But suppose we can arrange for a *side payment*. Specifically, we collect $80,000 from those who benefit and pay it to those who are harmed. Those who make the payment will come out ahead: They initially gained $100,000 worth of benefits, so after paying $80,000, they are left with net benefits of $20,000. Those who initially didn't want the theater come out ahead as well: They suffer harm worth $70,000, but the $80,000 payment gives them a net benefit of $10,000.

If you experiment around a bit, you'll see that *any* side payment greater than $70,000 and less than $100,000 would make building the theater a Pareto improvement.

More generally,

> *if any action creates greater total gains for some than total losses to others, then a side payment exists which, if transferred from the gainers to the losers, would make the action a Pareto improvement.*

Any side payment with a value between the total benefits to the gainers and the total losses to the losers will do the trick.

This has an important implication for economic efficiency. Suppose we find some transaction that could benefit some more than it harms others, and suppose *an appropriate side payment can be easily arranged*. Then *not* doing that transaction would be a waste of an opportunity to make people better off. Economic efficiency requires that we find, and exploit, opportunities that—with side payments—would be Pareto improvements.

But reread the italicized words in the paragraph above. The appropriate side payment—as you'll see in the next chapter—is *not* always easy to arrange. In many instances, arranging a side payment to ensure that everyone benefits has high costs.

It may be that, after deducting these costs, too little would be left to adequately compensate the losers while still leaving the gainers better off.

When someone is harmed by an action and a side payment *cannot* be made—or for any reason is *not* made—then even though the action might create greater gain than harm, it might not be considered fair. Achieving the efficient outcome then becomes *one* consideration, but not the only one.

Competitive Markets and Economic Efficiency

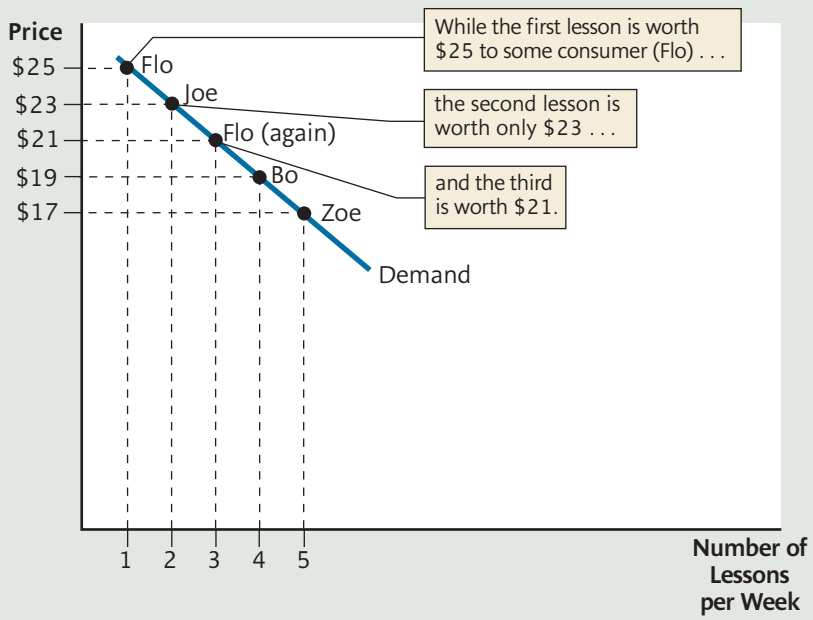
In a market system, firms and consumers are largely free to produce and consume as they wish, without anyone orchestrating the process from above. Can we expect such unsupervised trading to be economically efficient? That is, when markets are in equilibrium, will we discover that we are exploiting all possible Pareto improvements, so we are not wasting opportunities to make people better off?

In this section, you'll see that the answer is usually yes . . . *if* trading takes place in perfectly competitive markets. To see this, we'll return to two familiar tools—the demand curve and the supply curve—but we'll be interpreting them in a new way.

Reinterpreting the Demand Curve

Figure 1 shows a market demand curve for guitar lessons. It also indicates the specific person who would be taking each lesson along the curve. For example, at a price of $25, only Flo—who values guitar lessons the most—takes a lesson, so weekly quantity demanded is one. If the price drops to $23, Joe will take one lesson, so quantity demanded is two. At $21, Flo will decide to take a second lesson each

FIGURE 1 The Value of Another Guitar Lesson



week, so quantity demanded rises to three. This is the standard way of thinking about a market demand curve: It tells us quantity demanded at each price.

But we can also view the curve in a different way: It tells us the maximum price someone would be willing to pay for each unit. Therefore, it tells us how much that unit is *worth* to the person who buys it. In Figure 1, for example, the maximum value of the first lesson to some consumer in the market is just a tiny bit greater than $25. How do we know this? Because Flo, who values this lesson more highly than anyone else, will not buy it at any price greater than $25. But if the price falls to $25, she will buy it. When she decides to buy it, she must be getting at least a tiny bit more in value than the $25 she is giving up. So the value of that first lesson must be just a tiny bit more than $25. Ignoring for the moment the phrase "tiny bit more," we can say that the first lesson in the market is worth $25 to some consumer (Flo), the second is worth $23 to some consumer (Joe), and the third is worth $21 (Flo again).

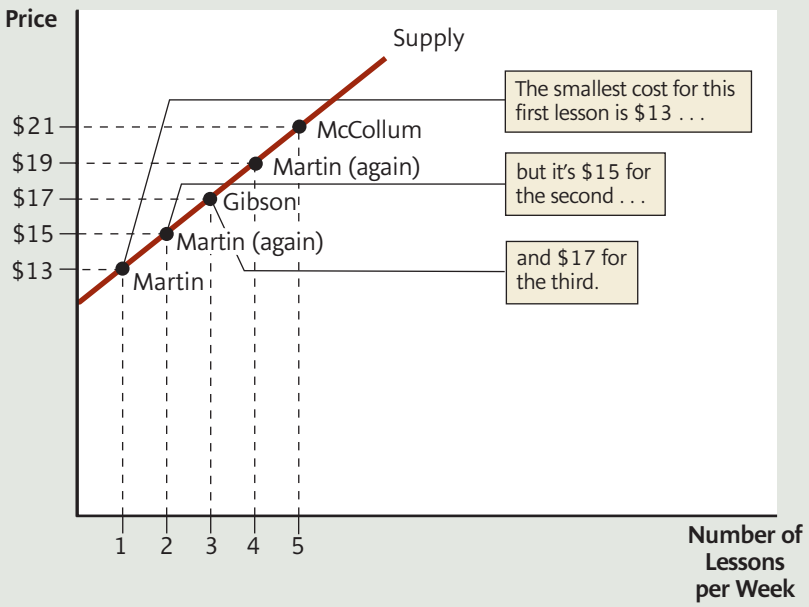
Of course, in Figure 1, we've simplified by assuming there are very few consumers in the market for guitar lessons. This makes the graph easier to read. But the point is the same whether there are five consumers in the market, or 500, or 50,000. In general,

> *the height of the market demand curve at any quantity shows us the value—to someone—of the last unit of the good consumed.*

Reinterpreting the Supply Curve

Now let's look at the other side of the market: those who *supply* guitar lessons. Figure 2 shows us a supply curve for guitar lessons. The figure also indicates who would be supplying each lesson. For example, at a price of $13, Martin would offer

FIGURE 2 The Cost of Another Guitar Lesson



one lesson each week. If the price rose to $15, Martin would offer two lessons per week, and at $17, another teacher—Gibson—would enter the market and offer a third. This is the standard way to view the supply curve: It tells us the quantity supplied at each price.

But we can also interpret the supply curve this way: It tells us the minimum price a seller must get in order to supply that lesson. For example, for the first lesson, the price would have to be at least $13. At any price less than that, no one will offer it. However, when the price reaches $13, one supplier (Martin) will decide to offer it. Similarly, $15 is the minimum price it would take to get some producer in this market (Martin again) to supply the second lesson, and $17 is what it would take for the third lesson to be supplied (Gibson this time).

Why does it take higher prices to get more lessons? Because offering lessons is *costly* to guitar teachers. Not only do they have to rent studio space, but they must also use their time, which comes at an opportunity cost. In order to get someone to supply a guitar lesson, the price must *at least* cover the additional costs of giving that lesson. We can expect this additional cost to rise as more lessons are given. (For example, the opportunity cost of a guitar teacher's time will rise as he gives more lessons and has less free time to do other things.)

The minimum price that would convince Martin, Gibson, or any other teacher to supply a lesson will be the amount that just barely compensates for the additional cost of that lesson—and a tiny bit more. Ignoring the phrase "a tiny bit more," we can say that

> *the height of the market supply curve at any quantity shows us the additional cost—to some producer—of the last unit of the good supplied.*[1]

The Efficient Quantity of a Good

Figure 3 combines the supply and demand curves for guitar lessons. Remember that the demand curve shows us the *value* of each lesson to some *consumer* and the supply curve shows us the additional *cost* of each lesson to some producer. We can then find the efficient quantity of weekly guitar lessons by using the following logical principle:

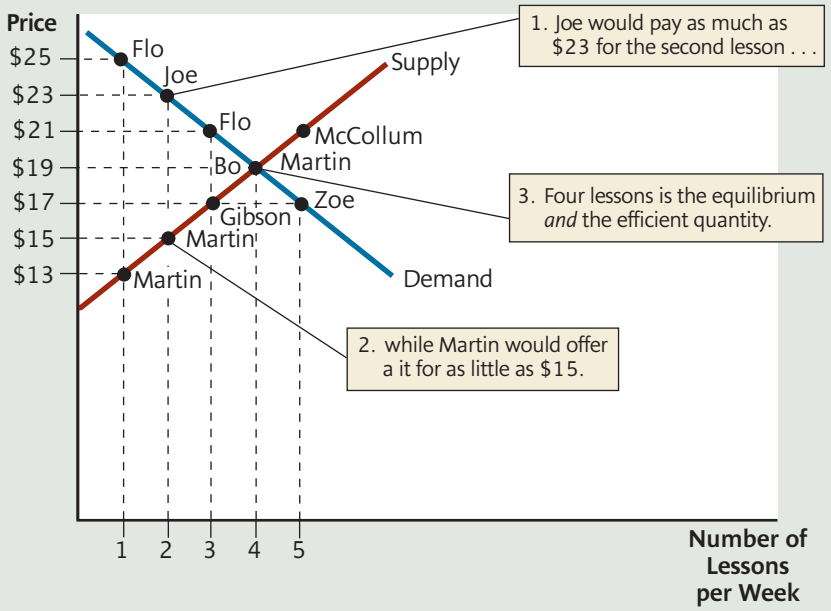
> *Whenever—at some quantity—the demand curve is higher than the supply curve, the value of one more unit to some consumer is greater than its additional cost to some producer.*

This means that when the demand curve lies above the supply curve, we can always find a price for one more unit that makes both the consumer and the producer better off—a Pareto improvement.

Here's an example: Look at the *second* lesson in the figure. Tracing up vertically, we see that the demand curve (with a height of $23) lies above the supply curve (with a height of $15). That tells us that some consumer—Joe—values this lesson more than it would cost some teacher—Martin—to provide it. If Joe can *buy* the lesson at any price *less than $23*, he comes out ahead; if Martin can

[1] If you've been reading the chapters in order, you'll recognize *additional cost* as *marginal cost*, first introduced in Chapter 7 and discussed in later chapters. That is, the height of the supply curve tells us the lowest marginal cost at which each unit could be supplied in the market.

FIGURE 3 **Efficiency in the Market for Guitar Lessons**



sell it for any price *greater than $15,* he comes out ahead. So, at any price *between $15 and $23,* both will come out ahead and no one is harmed: a Pareto improvement.

Continuing in this way, we find that the third lesson, and even the fourth, could be offered as Pareto improvements. (The fourth would be only a *slight* Pareto improvement, because its value is just a tiny bit greater than $19 and its cost a tiny bit less than $19.)

What about lessons for which the demand curve is *lower* than the supply curve—such as the fifth? Then there is *no* price at which both buyer and seller could come out ahead. To Zoe, the consumer who values the fifth weekly lesson the most, it's worth only $17. And for McCollum, the one who could provide it at the lowest additional cost, that cost would be $21. Producing this lesson could not be a Pareto improvement—no matter what the price—since the lowest cost of providing it is greater than its highest value to anyone in the market.

Let's recap: Whenever the demand curve lies *above* the supply curve, producing and selling another lesson (at a price between its value and its cost) is a Pareto improvement. Whenever the demand curve lies *below* the supply curve, producing another lesson can*not* be a Pareto improvement. This tells us that the efficient quantity of guitar lessons—the quantity at which all Pareto improvements are being exploited—is where the demand and supply curves intersect. At this quantity, the value of the last unit produced will be equal to (or possibly a tiny bit greater than) the cost of providing it.

The efficient quantity of a good is the quantity at which the market demand curve and market supply curve intersect.

The Efficiency of Perfect Competition

As you learned in Chapter 3 (and again in Chapter 9), when markets behave as the model of perfect competition predicts, the price adjusts until the market quantity reaches its *equilibrium:* where the market demand curve and market supply curve intersect. But we've just seen that this quantity is also the *economically efficient* quantity—the one that exploits all possible Pareto improvements. This gives us a very important and powerful result:

> *A well-functioning, perfectly competitive market will automatically achieve the efficient quantity.*

If the market for guitar lessons is perfectly competitive, the equilibrium quantity will be the efficient quantity.

Let's consider this statement carefully. It tells us that, if we leave producers and consumers alone to trade with each other as they wish, then—as long as the market is working well and it's perfectly competitive—the market will exploit every opportunity to make someone better off that doesn't harm anyone else. No special side payments need to be arranged, because the price paid for the good *is itself* the side payment.

The notion that perfect competition—where many buyers and sellers each try to do the best for themselves—actually delivers efficient markets is one of the most important ideas in economics. The great British economist of the 18th century, Adam Smith, coined the term *invisible hand* to describe the force that leads a competitive economy toward economic efficiency:

> *[The 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 not part of his intention.*[2]

We can recognize the *end* promoted by the invisible hand as the economically efficient outcome.

Measuring Market Gains

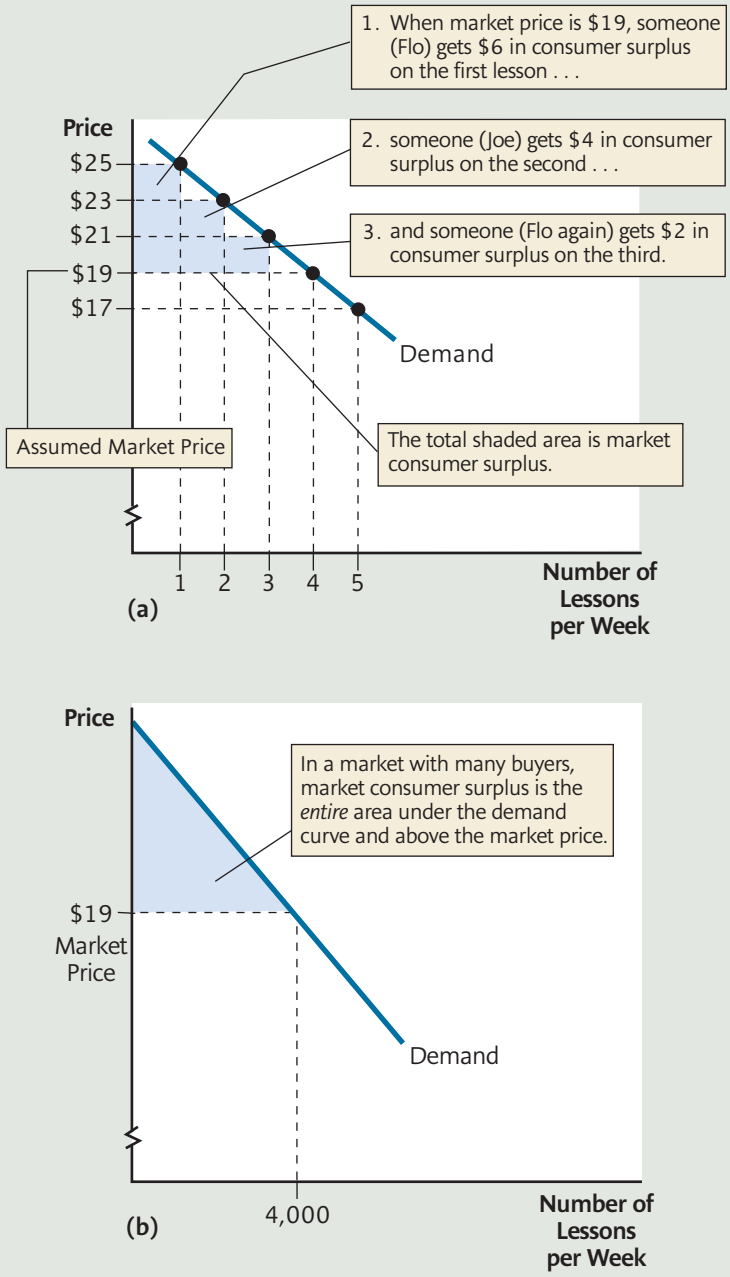
When markets are *not* perfectly competitive, or when they fail to function in other ways, they are *in*efficient. By comparing the actual benefits a market delivers with the benefits it *could* provide if it were operating efficiently, we can estimate what we lose from the inefficiency. In this section, you'll learn how economists measure the benefits traders receive in a market—and the potential benefits not realized.

Consumer Surplus

Let's start with the benefits enjoyed by consumers in a market. Consumers rarely have to pay what a unit of a good is actually worth to them. Regardless of how much they value the good, they can usually buy all they choose at the market price. For example, in panel (a) of Figure 4, the equilibrium price of guitar lessons is $19.

[2] Adam Smith, *The Wealth of Nations* (Modern Library Classics edition, 2000), p. 423.

FIGURE 4 **Consumer Surplus in a Small and a Large Market for Guitar Lessons**



So Flo is able to buy her first weekly lesson at $19, even though that lesson is *worth* $25 to her. By being able to purchase the lesson for less than its value to her, Flo gets a net benefit—called *consumer surplus*—on that lesson.

Consumer surplus The difference between the value of a unit of a good to the buyer and what the buyer actually pays for it.

*A buyer's **consumer surplus** on a unit of a good is the difference between its value to the buyer and what the buyer actually pays for that unit.*

Flo's consumer surplus on the first lesson is equal to \$25 − \$19 = \$6. It can be represented graphically by the *shaded area* of the first (leftmost) rectangle in the upper panel, which has a base of one unit and height of \$6 (from \$19 to \$25).

Continuing down the demand curve, the consumer surplus on the *second* guitar lesson purchased in this market (by Joe) would be what that lesson is worth to him (\$23) minus what he actually pays (\$19), or \$4. This is represented by the shaded area of the second, smaller rectangle. In similar fashion, the area of the third rectangle gives us a \$2 consumer surplus on the third lesson (Flo again). The fourth lesson is purchased by Bo. Since the most he'd be willing to pay for that lesson is \$19, its value to him is at most a tiny bit more than \$19—say, \$19.01. When the market price is \$19, Bo will buy the lesson, but he hardly gets any consumer surplus at all—so little that we can safely ignore it.

The total consumer surplus enjoyed by *all* consumers in a market is called **market consumer surplus,** the sum of the consumer surplus on all units (the areas of the shaded rectangles). In the figure, with the price of guitar lessons at \$19, market consumer surplus is \$6 + \$4 + \$2 = \$12. Notice that this is *roughly* equal to the entire area under the demand curve and above the market price of \$19. We say *roughly* because there are some unshaded triangles in the upper panel that are *not* part of anyone's consumer surplus. With only five consumers in the market, these unshaded triangles are significant, and we should not include them when we measure market consumer surplus.

Market consumer surplus The total consumer surplus enjoyed by all consumers in a market.

Panel (b) of Figure 4 shows a larger market—one we might find in a large city, with thousands of potential guitar students. In such a market, a one-unit width for each rectangle would be very small, and the unshaded triangles would be so insignificant that including them as part of consumer surplus hardly makes a difference in our measure. This is why, in the lower panel, we've indicated the market consumer surplus as the *entire* shaded area under the demand curve and above the market price.

We measure market consumer surplus, in dollars per period, as the total area under *the market demand curve and* above *the market price.*

Producer Surplus

Now let's turn to the supply side of the market. Only in the rarest of situations would a supplier have to sell each unit at the lowest acceptable price. As long as there are *many* sellers and each is too small to influence the price, each can sell all the units he chooses at the market price.

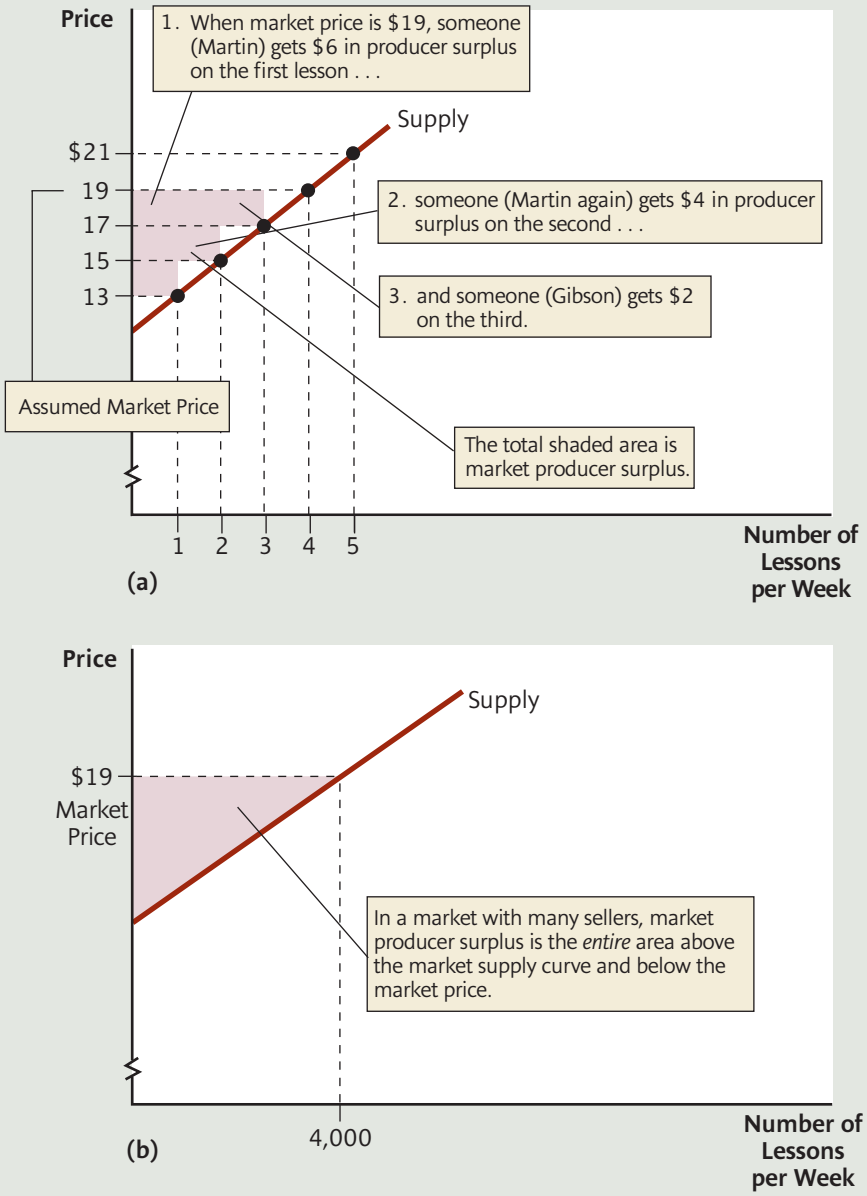
For example, in panel (a) of Figure 5, the market price of guitar lessons is \$19. Martin is able to sell the first lesson at \$19 even though he'd be *willing* to sell it for as little as \$13, which would just barely cover the additional costs of supplying it (studio rental, opportunity cost of time, and so on). By being able to sell the lesson for *more* than \$13, Martin gets a net benefit—called *producer surplus*—on that lesson.

A seller's ***producer surplus*** *on a unit of a good is the difference between the price the seller gets and the additional cost of providing it.*

Producer surplus The difference between what the seller actually gets for a unit of a good and the cost of providing it.

Martin's producer surplus on the first lesson is equal to \$19 − \$13 = \$6. It can be represented graphically by the *shaded area* of the first (leftmost) rectangle in the upper panel, which has a base of one unit and height of \$6 (from \$13 to \$19).

FIGURE 5 **Producer Surplus from Selling Guitar Lessons**



Continuing up the supply curve, the producer surplus on the *second* guitar lesson (Martin again) is the price for that lesson ($19) minus the lowest amount that would get Martin to supply it ($15), or $4. This is represented by the shaded area of the second, smaller rectangle. In similar fashion, the area of the third rectangle gives us the producer surplus on the third lesson (Gibson), equal to $2. The fourth lesson would be provided by Martin once again. Since the lowest price he'd be willing to accept in exchange for that lesson is $19, the additional cost to him is at most a tiny bit less than $19—say, $18.99. When the market price is $19, Martin will sell that lesson, but he hardly gets any producer surplus at all—so little that we can safely ignore it.

The total producer surplus gained by *all* sellers in a market is called **market producer surplus,** found by adding up the shaded rectangles gained by *all* sellers in the market. In the figure, with the market price for guitar lessons at $19, market producer surplus is $6 + $4 + $2 = $12. Notice that this is *roughly* equal to the entire shaded area *above* the supply curve and *below* the market price of $19—except for the little unshaded triangles in panel (a).

Market producer surplus The total producer surplus gained by all sellers in a market.

Panel (b) of Figure 5 shows a larger market for guitar lessons, which might have hundreds of potential teachers, each capable of offering a dozen or more lessons every week. In such a market, the unshaded triangles are insignificant, so market producer surplus would essentially equal the *entire* shaded area above the supply curve and below the market price.

We measure market producer surplus, in dollars per period, as the total area above *the market supply curve and* below *the market price.*

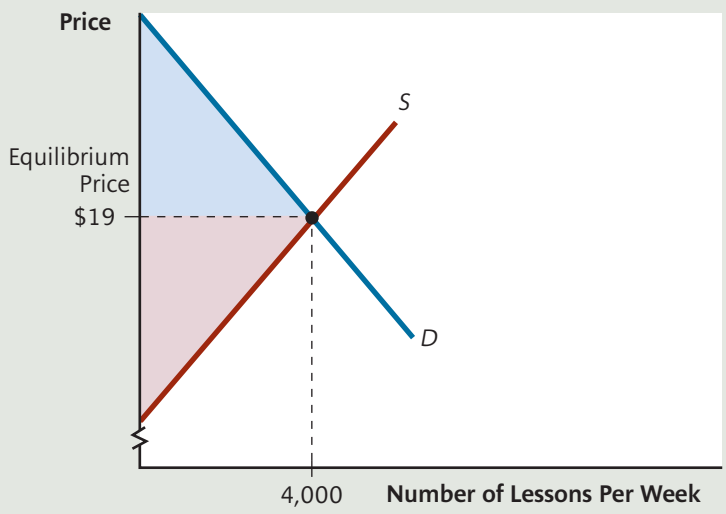
Total Benefits and Efficiency

Figure 6 combines the supply and demand curves for the large market shown in the previous figures. It shows consumer and producer surplus when this perfectly competitive market is in *equilibrium,* with price equal to $19 and quantity equal to 4,000. Market consumer surplus is the area under the demand curve and above the market price, or the blue-shaded area. Market producer surplus is the (red-shaded) area under the market price and above the supply curve. The total shaded area (both blue and red) represents the total benefits that consumers and producers receive from participating in this market each week.

We measure the ***total benefits*** *gained in a market as the sum of consumer and producer surplus in that market.*

Total benefits The sum of consumer and producer surplus in a particular market.

FIGURE 6 Total Benefits in a Competitive Market for Guitar Lessons



When a competitive market reaches equilibrium, the sum of market consumer surplus and market producer surplus is maximized. At any quantity less than 4,000 or greater than 4,000, total benefits will be smaller.

There is an important relationship between efficiency and total benefits. Each time we make a Pareto improvement in a market (making at least one party better off and harming no one), total benefits increase. But efficiency means that all such Pareto improvements are exploited. Therefore, efficiency means that we are exploiting all opportunities to increase total benefits. Or, to put it more succinctly:

A market is efficient when total benefits—the sum of consumer and producer surplus—are maximized in that market.

Perfect Competition: The Total Benefits View

Total benefits The sum of consumer and producer surplus in a particular market.

Look again at Figure 6, which shows the perfectly competitive market for guitar lessons at the equilibrium price and quantity. When this market is in equilibrium, the **total benefits**—the sum of the blue- and red-shaded areas—are the maximum total benefits achievable in this market. How do we know this? Because any change in either the quantity or the price away from their equilibrium values will shrink the total benefits.

Let's first see what would happen if we arbitrarily moved the *quantity* of guitar lessons to some value other than 4,000. Suppose the market were forced to supply *more* than 4,000—say, 5,000. For lessons 4,001 to 5,000, the supply curve lies above the demand curve, so their cost to the seller would be greater than their value to the buyer. Providing these additional lessons cannot add to total benefits in the market. In fact, no matter what price is charged for them, each additional lesson would *reduce* total benefits. For example, if a lesson has a cost of $21 but a value of $17, then producing that lesson *reduces* total benefits by $4, regardless of the price charged. (Charging $21 or more puts the entire loss on the buyer; charging $17 or less puts the entire loss on the seller.) Therefore, as we increase quantity above 4,000, total benefits fall.

What about a quantity *less* than 4,000—say, 3,000? For lessons 3,001 to 4,000, the demand curve lies above the supply curve, so their value to the buyer would be greater than their cost to the seller. These potential gains are lost when quantity is reduced to 3,000. Therefore, as we decrease quantity below 4,000, we reduce total benefits.

Because increasing the quantity above or below 4,000 *reduces* total benefits, we know that at 4,000, total benefits are maximized. More generally,

in a well-functioning, perfectly competitive market, the equilibrium quantity maximizes total benefits. This is another illustration that the equilibrium under perfect competition is efficient.

Inefficiency and Deadweight Loss

Our example of the perfectly competitive market for guitar lessons can be regarded as a benchmark (and ideal) case. The market, left to itself, exploits every possible Pareto improvement involving buyers and sellers.

But markets don't always achieve this efficient outcome. In some cases government intervention may be preventing efficiency. In other cases the nature of the market itself is responsible. In this section, we'll explore some examples of each of these cases of *in*efficiency.

A Price Ceiling

What happens in an otherwise efficient market if we impose a price ceiling below the equilibrium price?

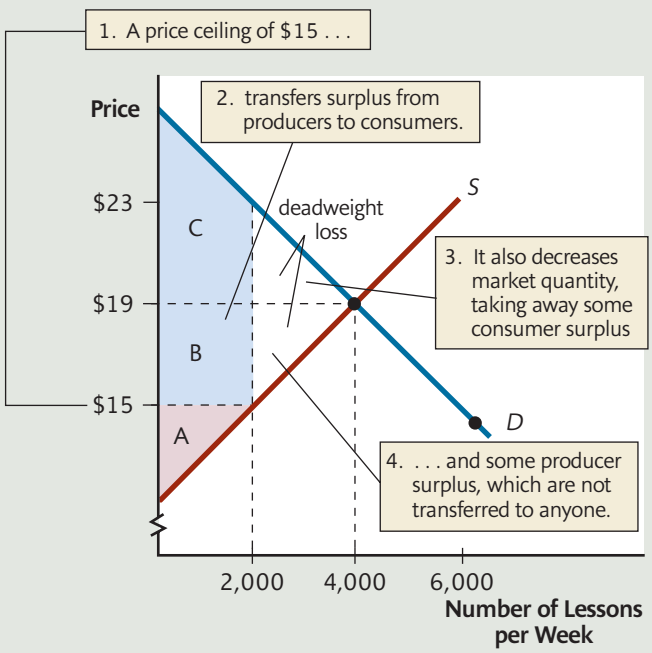
If the quantity remained the same, the change in price would change the way benefits are allocated, but not their total. For example, consider a guitar lesson that is initially produced and sold for $19. If we force the price down to $17, and if that guitar lesson were *still* produced, the seller would enjoy $2 less in benefits, and the buyer would enjoy $2 more. By merely transferring benefits from one party to the other, we would not affect the total.

However, as you learned in Chapter 4, a price ceiling *does* change quantity. In a market economy we cannot force people to sell more than they want to. So a price ceiling that lowers the price will *reduce* the quantity sold.

Figure 7 shows the impact of a $15 price ceiling imposed on the market for guitar lessons. At that price, quantity supplied of 2,000 lessons per week is smaller than quantity demanded of 6,000. Since sellers are the short side of this market, and they can't be forced to offer more than 2,000 lessons, that will be the market quantity bought and sold.

The figure also shows how buyers' and sellers' market benefits are affected by the price ceiling, under two assumptions. First, we assume that no black market develops. Second, we assume that the smaller number of available lessons will be purchased by those who value them the most (and who would therefore get the greatest possible surplus from them). (In the problem set, you'll be asked to analyze price ceilings with different assumptions.) Producer surplus (shaded in red) is the area above the supply curve and below the new market price of $15, up to 2,000 units. Consumer surplus (in blue) is the area below the demand curve and above $15 also

FIGURE 7 The Inefficiency of a Price Ceiling



up to 2,000 units. Even though consumers would like to buy *more* than 2,000 units at a price of $15, we must stop at 2,000 when measuring their surplus: The lessons beyond 2,000 are no longer provided, so no consumer surplus is gained from them.

Comparing the market with the price ceiling (Figure 7) to the market *without* the price ceiling (Figure 6), it looks like the blue area measuring consumer surplus (marked B and C) has increased. (A price ceiling doesn't always increase consumer surplus, but it does in our example.) But the price ceiling *decreases* the red area (marked A) measuring producer surplus. (An effective price ceiling will *always* reduce producer surplus.)

What has happened to the *total* of producer and consumer surplus? If you compare total benefits in Figure 6 with total benefits in Figure 7, you'll see that the price ceiling causes total benefits to fall. Specifically, in Figure 7, total benefits have been reduced by the area of the unshaded triangle. This area represents benefits that *could* be achieved in the market but are *not* achieved, because the market quantity drops from its efficient level of 4,000 to the inefficient level of 2,000.

The unshaded triangle is an example of what economists call a *deadweight loss* (sometimes also called a *welfare loss* or *excess burden*).

Deadweight loss The dollar value of potential benefits not achieved due to inefficiency in a particular market.

*The **deadweight loss** in a market is the loss of potential benefits (measured in dollars) due to a deviation from the efficient outcome.*

As you can see in the figure, a price ceiling causes a deadweight loss because it moves the market quantity away from its efficient level.

Calculating the Deadweight Loss

To make this more concrete, let's calculate the *dollar value* of the deadweight loss caused by the price ceiling—the area of the unshaded triangle. From high school algebra, the area of any triangle is $\frac{1}{2} \times$ base $\times$ height. Imagine that our triangle has been tipped on its side, so that its *vertical* side is the *base*. This side goes from $15 to $23, so the base has a length of $23 − $15 = $8. The horizontal dashed line cutting through the middle of the triangle would be its *height*. This line goes from 2,000 to 4,000, so its length is 2,000. Applying the formula, we find that the loss = $\frac{1}{2} \times$ base $\times$ height = $\frac{1}{2} \times \$8 \times 2{,}000 = \$8{,}000$.

In words, when this market is delivering only 2,000 lessons per week instead of the efficient 4,000, guitar teachers and students together lose $8,000 in potential benefits—per *week*. If measured yearly, the deadweight loss would be 52 weeks $\times$ $8,000 per week = $416,000 per year.

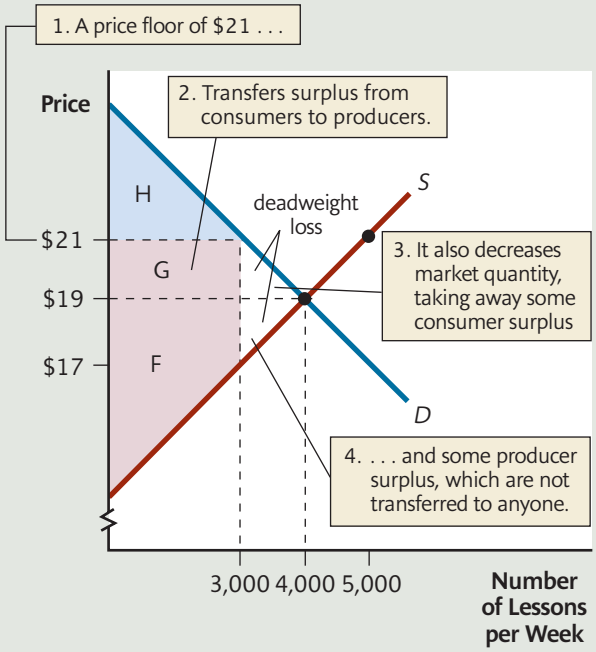
A Price Floor

What happens in an otherwise efficient market if we raise the price *above* its equilibrium value? That is, how does a price *floor* affect total benefits? The analysis is very similar to that for a price ceiling.

Figure 8 shows the impact of a price *floor* in our market, set at $21 per lesson. At that price, quantity demanded of 3,000 lessons per week is smaller than quantity supplied of 5,000. Now *buyers* are the short side of this market, so 3,000 will be the market quantity.

Let's assume that the 3,000 lessons supplied are provided by those sellers who would get the most surplus from them, i.e., those willing to supply them at the lowest prices. Producer surplus after the price floor is shaded in red and consumer surplus in blue. Notice that both surpluses are measured only up to 3,000 lessons—the

FIGURE 8 The Inefficiency of a Price Floor



quantity actually provided. Comparing the market with the price floor (Figure 8) to the market *without* the price floor (Figure 6), we find that the red area measuring producer surplus (marked F and G) has increased. (A price floor won't always increase producer surplus, but it does in our example.) But the price floor *decreases* the blue area (marked H) measuring consumer surplus. (An effective price floor will always reduce consumer surplus.)

What has happened to the *total* of producer and consumer surplus? If you compare total benefits in Figure 6 with total benefits in Figure 8, you'll see that the price floor causes total benefits to fall. In Figure 8, the *deadweight loss* from the floor is equal to the area of the unshaded triangle. (In the problem set, you'll be asked to calculate the dollar value of this deadweight loss.)

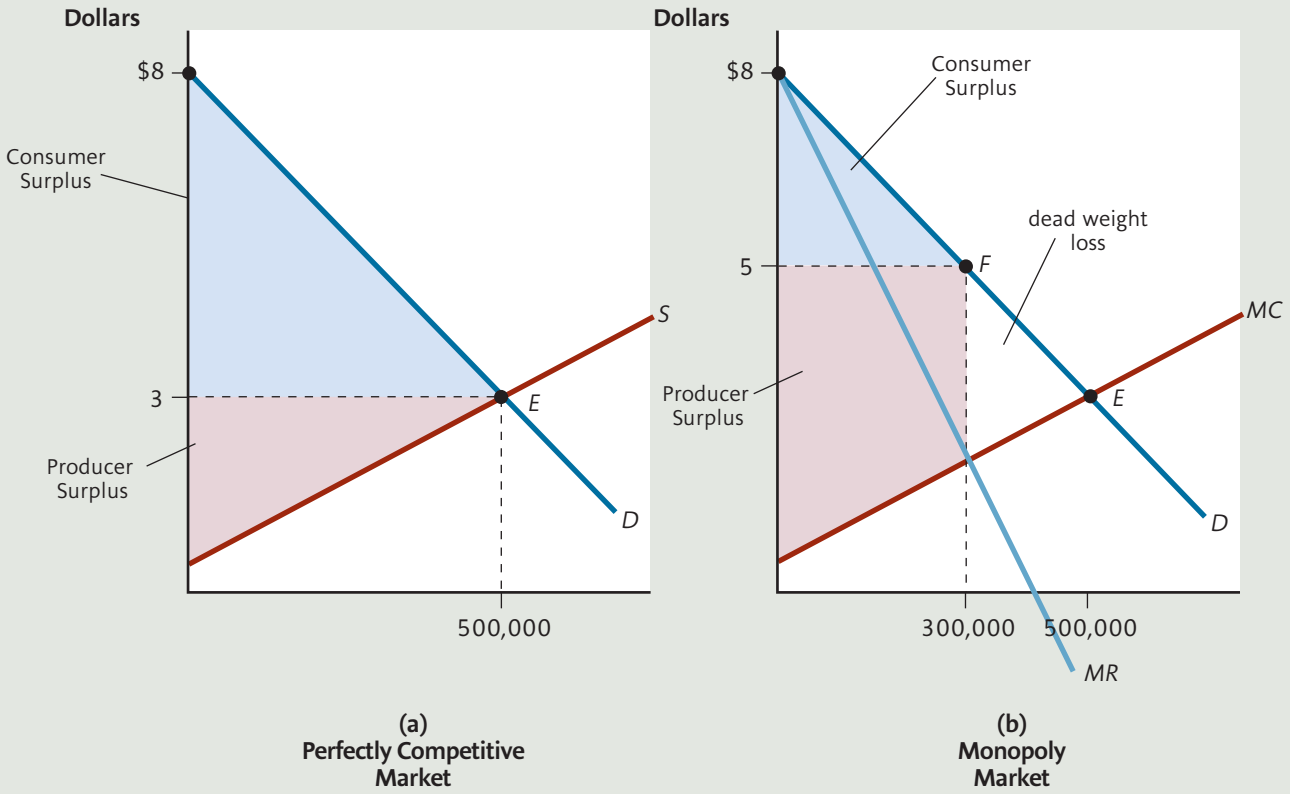
Market Power

A firm has *market power* when it can influence the price that it charges for its product. Monopolists, oligopolists, and monopolistic competitors all have some degree of market power because they *set* their price in order to maximize profit, rather than take the price as given.

In monopolistic competition, the presence of many competitors generally limits market power and helps keep prices low. But when a market has just one seller, or a few oligopolists who collude, the price can be significantly higher than in an otherwise similar competitive market. With a higher price, a lower quantity will be produced and sold—a quantity that is *less* than the efficient quantity.

To see how this works, look at Figure 9(a). The left panel shows a perfectly competitive market for wheat, in equilibrium at point *E*, with a market price of $3 per bushel and market output of 500,000 bushels per period.

FIGURE 9 The Deadweight Loss from Monopoly



In panel (a), the market for wheat is perfectly competitive. Price is $3 per bushel, and output is 500,000 bushels per period. Total benefits are equal to the area of the two shaded triangles.

In panel (b), the market for wheat is monopolized. The competitive market supply curve becomes the monopoly's MC *curve. The monopoly, at point* F*, supplies fewer bushels (300,000) at a higher price ($5). Potential benefits on bushels 300,001 to 500,000—for which the value to some buyer would be greater than the marginal cost—are not realized. The result is a dead-weight loss equal to the area of the unshaded triangle.*

Now imagine that a single company buys up all the farms in this market. We'll assume this new monopoly treats each of the old farms as an independent operation—with one exception: The monopoly owner will now set the price of wheat in the market so as to maximize its total profit.

Figure 9(b) shows the market from the perspective of the new monopoly. Each time the monopoly wants to sell an additional bushel of wheat, it will produce it on the individual farm that can do so at the lowest additional cost—that is, by the supplier who provided that bushel before, when the market was perfectly competitive. Thus, the monopoly's marginal cost curve in panel (b)—showing the additional cost of another bushel—is the same as the old market supply curve in panel (a).

The demand curves in the two panels are identical: *At any given price,* buyers in the market would choose to buy the same quantity of wheat before and after monopolization.

The monopoly maximizes profit by choosing the number of bushels at which marginal revenue and marginal cost are equal. However, the monopoly—unlike the competitive suppliers—must drop the price on *all* bushels in order to sell one more. That's why the monopoly's marginal revenue curve lies below the demand curve: Marginal revenue is less than the price of the last bushel.

In panel (b), you can see that the monopoly's profit-maximizing output level is 300,000 bushels, and it charges the highest price—$5—at which it can sell that quantity.

The Deadweight Loss from Monopoly

Now we'll calculate the deadweight loss—the extent to which total benefits fall short of their maximum—caused by the monopolization of this market. Our first step is to identify the new consumer surplus in panel (b). This is the blue-shaded area—below the demand curve and above the market price of $5.

Next, we identify the monopoly's producer surplus as the purple-shaded area (above its new marginal cost curve and below the market price of $5). Why? Because producer surplus on each unit for the monopoly is the difference between the added cost of that unit (given by the *MC* curve) and what it actually gets for it ($5). When we sum up this surplus for all 300,000 units produced, the result is the purple-shaded area.

What has happened to the *total* of producer and consumer surplus? Compare total benefits in the perfectly competitive market (the shaded areas in panel (a)) with total benefits under monopoly (the shaded areas of panel (b)). You'll see that the monopoly takeover has caused total benefits to fall. Specifically, total benefits have been reduced by the area of the *unshaded triangle*. This area is the *deadweight loss* from monopolization of the industry.

To understand the deadweight loss caused by monopolization, remember that 500,000 is the efficient quantity—the quantity supplied and demanded under perfect competition. But the monopoly provides only 300,000. Providing each bushel from number 300,001 to number 500,000 would be a Pareto improvement, but none of these is provided. Why not? Because *the monopoly charges a price that is greater than marginal cost*. Thus, buyers will choose *not* to buy some bushels even though their value exceeds the cost of providing them.

Note that if the monopoly could *price discriminate*—say, continuing to charge $5 on bushels 1 through 300,000, and then charge $3 for bushels 300,001 through 500,000—it *would* choose to supply the efficient quantity. (You may want to review price discrimination in Chapter 10; in this case, the *MR* curve would be a horizontal line at $3 for all quantities from 300,001 to 500,000.) But as you've learned, not all firms can price discriminate. A *single-price* monopoly, like the one in panel (b), will be inefficient.

Our example generalizes to any firm facing a downward-sloping demand curve—that is, any firm with market power—that cannot price discriminate.

> *Monopoly and imperfectly competitive markets—in which firms charge a single price on all units that is greater than marginal cost—are generally inefficient. Price is too high, and output is too low, to maximize the total benefits in the market.*

In the next chapter, we'll explore some options for government policy in dealing with monopoly.

Using the Theory

TAXES AND DEADWEIGHT LOSSES

© TIM BOYLE/GETTY IMAGES NEWS

In Chapter 4, you learned that imposing a tax on a competitive market changes the equilibrium quantity. Based on what you've learned in *this* chapter, you won't be surprised that this change in quantity—in an otherwise efficient market—creates a deadweight loss.

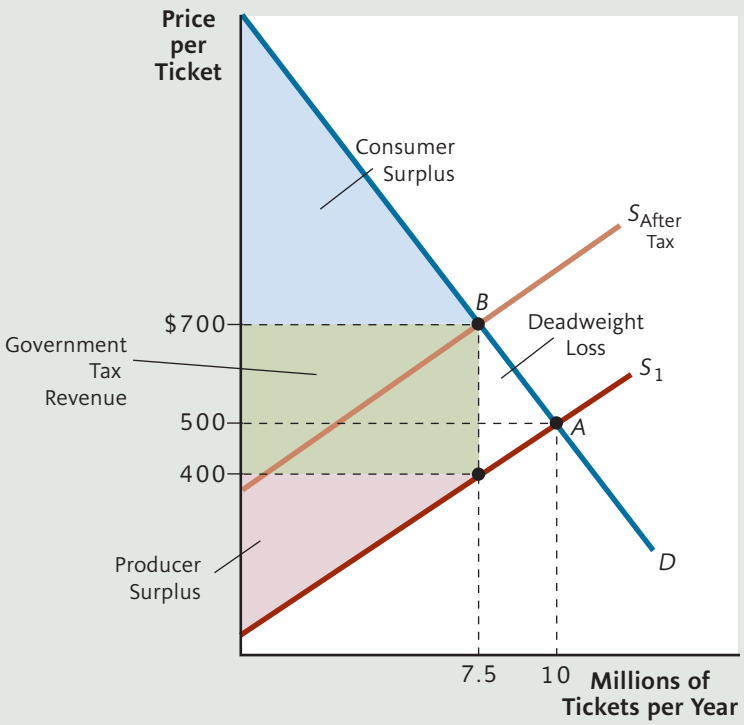
Before we get to that result, let's briefly revisit how a tax affects a competitive market.

How a Tax affects Price and Quantity: A Review

Figure 10 illustrates the impact of an excise on a competitive market for airline travel. The initial market equilibrium is at point *A*, with 10 million tickets sold per year at a price of $500 each. Then, the government imposes a tax on the market, collected from sellers. We'll make that tax $300 per ticket, which is much larger than actual airline taxes, but makes the graph easier to read.

Remember that a tax on sellers shifts the supply curve upward. To supply any given number of tickets, the price required by sellers is $300 more than before. As a result, the supply curve shifts upward by $300, to $S_{\text{After Tax}}$. The market equilibrium moves from point *A* to point *B*, with the new quantity equal to 7.5 million tickets

FIGURE 10 Deadweight Loss from an Excise Tax



per year. Buyers now pay $700 per ticket, and sellers receive $400 (after we deduct the $300 tax they must pay).

This part of the analysis should be familiar to you, based on what you learned about excise taxes in Chapter 4. But now, let's look at this tax from the point of view of economic efficiency.

Measuring the Deadweight Loss

Before the tax was imposed, consumer surplus was equal to the area of the entire large triangle under the demand curve and above the price of $500. (Make sure you can identify this on the graph.) Producer surplus was equal to the area of the entire triangle above the supply curve and below the price of $500. Total benefits were equal to those two areas summed together.

But after the tax is imposed, both producer and consumer surplus shrink. Consumer surplus shrinks to the (blue shaded) area under the demand curve and above the price of $700 that consumers pay. Producer surplus shrinks to the (red shaded) area above the supply curve and below the price of $400 that sellers now receive. The sum of these two shaded areas represents the total benefits to consumers and producers after the tax.

However, this sum is not *total* benefits, because there is one more benefit to consider: the government revenue from the tax. After all, this revenue can be used to reduce other taxes (a benefit to other taxpayers) or to provide government services (a benefit to those who receive them).

In the figure, the government's revenue is equal to the area of the green-shaded rectangle. To see why, note that the tax per ticket is $300, the height of the green rectangle. It is collected on each of 7.5 million tickets per year, the base of the rectangle. When we multiply the tax per ticket times the number of tickets, we get $300 × 7.5 million = $2,250 million, which is the area of the green rectangle.

If we now add together consumer surplus (blue), producer surplus (red), and the government's tax revenue (green), we get the total benefits after the tax is imposed on this market. Comparing this total to what benefits would be with *no* tax (all of the shaded areas together plus the unshaded triangle), we see that total benefits have been reduced by the area of the unshaded triangle—the deadweight loss from the tax.

As always, the deadweight loss occurs because quantity has changed. If the quantity had remained at the efficient level, the tax would have merely redistributed benefits from buyers and sellers to the government, with no change in the total. But the tax—by raising the price buyers pay and lowering the price sellers receive—results in fewer tickets being sold. The "missing tickets" (from 7,500,001 to 10,000,000) previously provided benefits to both buyers and sellers, but now they provide benefits to no one—not even the government.

The result we obtained for an excise tax applies to other types of taxes as well. In general,

> *a tax imposed on an otherwise efficient market creates a deadweight loss: the loss in benefits to buyers and sellers is greater than the gain in revenue to the government.*

For example, the payroll tax on wages creates a deadweight loss in *labor* markets. The loss in benefits to those who buy labor (firms) and sell it (households) is greater than the gain in revenue to the government, so total benefits decrease. (You'll be asked to analyze a labor market tax in the problem set.)

The personal income tax, which is a tax on wage income together with other sources of income, has a similar effect. And taxes on interest, dividends, and capital gains create deadweight losses in otherwise efficient markets for capital.

How large are these deadweight losses from taxes? They can be substantial. Estimates of the deadweight loss from various taxes in the United States range from 20 to 60 cents on each dollar of revenue collected.[3] A deadweight loss of this size means that each dollar collected by the government reduces total consumer and producer surplus by between $1.20 and $1.60.

Does knowledge about deadweight losses help us? After all, doesn't the government need revenue? Doesn't it have to impose taxes? Indeed it does. But recognizing *why* a tax causes a deadweight loss can help us design more efficient tax policies.

Deadweight Loss and Elasticity

A deadweight loss arises because the *quantity* of the good or resource being traded decreases below the efficient quantity. And, in turn, the quantity falls because the tax changes the *price* paid by buyers and received by sellers.

When demand or supply is very elastic (sensitive to price), even a small tax will result in a relatively large change in quantity, and deadweight losses will be relatively large. But the opposite applies as well: the more *inelastic* demand or supply, the smaller the deadweight loss.

Therefore,

> *all else equal, taxes create smaller deadweight losses when they are imposed on markets in which demand or supply is relatively inelastic.*

In the extreme case, if supply or demand were *completely* inelastic, a tax would cause no change in quantity, and no deadweight loss at all.

A Tax on Land

Let's consider a tax on a market with a completely inelastic supply: the market for land. In Figure 11, we assume for simplicity that all land in this market is rented. The demand curve tells us the quantity of land people would want to rent at each price (each monthly rent). The supply curve tells us the quantity of land that owners make available to renters.

Notice that the supply curve for land is vertical. The price may rise or fall, but the same quantity (3 million acres) will always be available. Why? In many cases, the total quantity of land available is fixed. If you are a landowner in one of these markets, it is always better to rent out your land, even at a very low price, rather than *not* rent it and earn nothing. So no matter how low the price, every acre of land owned by every landowner will still be available for rent. That is what we've assumed in the figure.

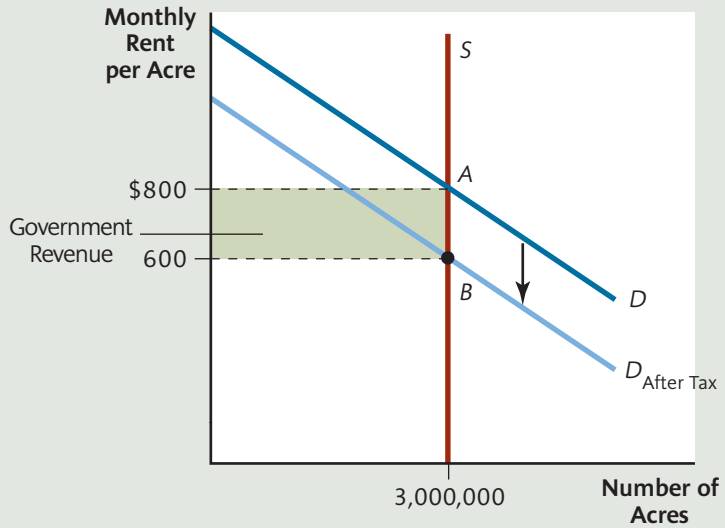
Initially, the market is in equilibrium at point *A*, with rent of $800 per acre per month. Then a tax of $200 per acre per month is imposed on land rentals. We'll first imagine that the tax is collected from the demand side of the market—those who rent land from landowners. The tax causes the demand curve to shift downward to $D_{\text{After Tax}}$. Why? To get renters to choose any given quantity, they must be charged $200 less than before, because now they also have to pay the tax.

The tax moves the market equilibrium from *A* to *B*, and rent drops by the full amount of the tax, from $800 to $600. This tells us that renters neither gain nor lose: Including the tax, they still pay $800 per acre and still rent 3 million acres. *Landowners*, however, receive only $600 per acre, so they lose $200 per acre. They end up paying the entire tax. The total loss for landowners is $200 × 3 million = $600 million per month, equal to the green-shaded area.

But notice that the landowners' loss is precisely equal to the government's gain in revenue, which is also $600 million (the same green-shaded area). The tax causes

[3] Congressional Budget Office, *Budget Options,* February 2001, p. 381.

FIGURE 11 A Tax on Land



The market for land is initially in equilibrium at point A, with monthly rent equal to $800 per acre, and all 3 million acres rented. When a tax of $200 is imposed, the demand curve shifts downward to $D_{After\ Tax}$, and the new market equilibrium is at point B. Rent decreases from $800 to $600 (by the full amount of the tax), and the government's dollar revenue gain is equal to the dollar loss by landowners. Renters pay none of the tax. Since the equilibrium quantity of rented land remains unchanged, there is no deadweight loss.

no deadweight loss at all; it has merely transferred dollars from landowners to the government.

What if the tax were imposed on *landowners* instead of renters? The result would be exactly the same. We can't shift the supply curve to illustrate that case, because a vertical curve can't be shifted upward or downward. But we *can* use logic to see what would happen. Suppose landowners *tried* to pass some or all of the tax on to renters by raising the rent they charge. As you can see in the figure, any rent greater than $800 creates an excess supply of land, so rent would fall back to $800. Because rent must remain at $800 per acre, but the tax is collected from the owners, they are left getting $600 per acre. Once again, the entire burden of the tax falls on the owners.

Henry George and the Single-Tax Movement

Economists have long debated the idea of a land tax. In 1879, economist and social philosopher Henry George proposed that other taxes should be abolished, and that government should raise *all* of its revenue by taxing land. The "single-tax" movement that was formed to advocate this idea was very influential for a time and still has some prominent adherents. While Henry George's proposal was based more on concerns for equity than efficiency, there is clearly an efficiency argument in favor of shifting more of the tax burden to land.

But many economists see problems in a major shift from current revenue sources to a land tax. First, in order to avoid deadweight loss, the tax would have to be on the value of the *land only*—excluding the value of any improvements made to it. Such improvements include homes, factories, irrigation systems, private roads, and other infrastructure. If the value of such improvements is taxed, their quantity will fall in the long run, and we are back to the problem of deadweight losses on the improvements themselves. But once we limit ourselves to taxing the value of just land, minus the value of any improvements, the tax base shrinks significantly. It would be too small to substitute completely for other sources of revenue.

There are also equity considerations. A shift to a land tax would—by lowering rent—reduce the rate of return on land. This harm would occur *after* people had

bought land at a price based on the current tax system. Those who had already bought land before the tax was imposed would be forced to either (1) suffer a lower-than-expected rate of return on their investment, or (2) suffer a capital loss if they try to sell their land. (Once rents fall, the price any new buyer would pay for land would drop as well.) Thus, current landowners—whether wealthy or not—would be harmed by this change in the rules.

In spite of these objections, the idea of shifting *some* of the tax burden toward gifts of nature like land, as well as toward other resources with a completely inelastic supply (such as radio spectrum or airline corridors), is alive and well.[4]

One final word about taxes and efficiency. If you've read this section carefully, you've seen that we've often used the phrase "otherwise efficient market" in discussing the deadweight loss from a tax. But in some cases, when a market is *not* otherwise efficient, imposing a tax can actually *make* it efficient. We'll explore these cases, as well as other situations in which government action can help to foster efficiency, in the next chapter.

SUMMARY

A *Pareto improvement* is an action that makes at least one person better off, and harms no one. A market or an economy is *economically efficient* when all Pareto improvements have been exploited. When well-functioning, perfectly competitive markets are left free to reach equilibrium, they produce the economically efficient quantity.

Economic efficiency can also be viewed as the outcome that maximizes *total benefits* in a market. In a market without taxes, total benefits are the sum of *consumer surplus* (the dollar value to consumers minus what they actually pay) and *producer surplus* (the dollar value producers receive minus the minimum payment necessary to get them to produce). In a competitive market, the equilibrium quantity, which is also the efficient quantity, maximizes total benefits.

When market quantity deviates from the efficient quantity, the result is a *deadweight loss:* the value of potential benefits not achieved due to inefficiency. A price *ceiling* or a price *floor* in an otherwise efficient market may or may not increase the surplus for one side of the market. But it changes the equilibrium quantity, and therefore decreases *total* benefits and creates a deadweight loss.

A deadweight loss also occurs when a competitive market is monopolized, or when a tax is imposed in an otherwise efficient market. The more elastic is demand or supply, the greater is the deadweight loss from any given tax.

PROBLEM SET

Answers to even-numbered Questions and Problems can be found on the text website at www.cengage.com/economics/hall.

1. In Figure 3, suppose that, initially, McCollum is providing the fifth guitar lesson to Zoe for a price of $16. Who would gain and who would lose from this lesson and how much?
2. Figure 8 shows a price floor of $21 in the market for guitar lessons. Calculate the dollar value of the deadweight loss caused by the price floor, using the numbers on the graph.
3. The following table shows the quantities of bottled water demanded and supplied per week at different prices in a particular city:

Price	Quantity Demanded	Quantity Supplied
$1.10	8,000	0
$1.15	7,000	1,000
$1.20	6,000	2,000
$1.25	5,000	3,000
$1.30	4,000	4,000
$1.35	3,000	5,000
$1.40	2,000	6,000
$1.45	1,000	7,000
$1.50	0	8,000

[4] For a well-argued defense of such a shift, see Fred Foldvary, "Geo-Rent: A Plea to Public Economists," *Econ Journal Watch,* April 2005, pp. 106–132.

a. Draw the supply and demand curves for this market, and identify the equilibrium price and quantity.
b. Identify on your graph areas for market consumer surplus and market producer surplus when the market is in equilibrium.
c. Using your graph, calculate the dollar value of market consumer surplus, market producer surplus, and the total net benefits in the market at equilibrium.

4. Suppose the government imposes a price *ceiling* of $1.20 in the market for bottled water in problem 3. Calculate the dollar value of each of the following:
   a. Market consumer surplus
   b. Market producer surplus
   c. Total net benefits in the market
   d. The deadweight loss from the price ceiling

5. Suppose the government imposes a price *floor* of $1.40 in the market for bottled water in problem 3. Calculate the dollar value of each of the following:
   a. Market consumer surplus
   b. Market producer surplus
   c. Total benefits in the market
   d. The deadweight loss from the price floor

6. Calculate the deadweight loss caused by the monopolization of the wheat industry in Figure 9. (Note: For marginal revenue at 300,000 bushels, use $2.00.)

7. The Using the Theory section pointed out that a tax on labor income can cause a deadweight loss, just like an excise tax on a good.
   a. Draw a diagram of a labor market in which the equilibrium wage is $20 per hour and total employment is 10,000 workers. On the graph, identify an area that represents total benefits to workers. (Hint: This area will be analogous to producer surplus in a goods market. Think about each point on the labor supply curve, and ask: What is the lowest wage at which this worker would supply labor, compared to the wage they are actually being paid?)
   b. On the same graph, identify an area that represents total benefits to firms from hiring labor. (Hint: This area will be analogous to consumer surplus in a goods market. Think about each point on the labor demand curve, and ask: What is the highest wage the firm would pay to hire *this* particular worker, compared to the wage it is actually paying?)
   c. Draw a second diagram showing the impact of a $10 per hour tax on labor income, collected from workers. On this diagram, identify areas that represent, after the tax, each of the following: (1) the total benefits to workers, (2) the total benefits to firms, (3) the government's tax revenue, and (4) the deadweight loss from the tax.

8. Suppose equilibrium price in a market is $5, and then a price ceiling of $3 is imposed. Assume (as in the chapter) that those who value the product the most are able to buy whatever quantity is available, and there is no black market.
   a. If *supply* is completely price inelastic between $3 and $5, is there a deadweight loss? Briefly, why or why not?
   b. If *demand* is completely price inelastic between $3 and $5, is there a deadweight loss? Briefly, why or why not?

More Challenging

9. In Figure 7, we assumed that the 2,000 lessons available would be purchased by those who value them most (i.e., those who would get the most surplus from them). But one problem with price ceilings is that available supplies are sometimes allocated haphazardly, so that some consumers who value the good less are able to buy it, while others who value it more are not. Re-do the analysis of the price ceiling of $15 in Figure 7, this time under the (extreme) assumption that the 2,000 consumers who value lessons the *least* (but are willing to pay $15 or more) end up getting them. Specifically:
   a. Identify the *new* area representing consumer surplus after the price ceiling.
   b. Identify the *new* area representing the deadweight loss after the price ceiling.
   c. Evaluate the following statement: "The unshaded triangles in the original Figure 7 show the *maximum* deadweight loss we would expect from a price ceiling in that market." True or false? Explain.

10. Figure 7 shows how a price ceiling affects consumer and producer surplus in the competitive market for guitar lessons. Suppose instead that Figure 7 (and all of the numbers in it) depicts the competitive market for tickets to rock concerts by local bands. Further, when the city imposes a $15 price ceiling, a black market develops in which ticket scalpers buy up all of the tickets available, and sell them all at the highest single price that the market will bear. Draw a graph and identify areas for each of the following after the price ceiling is imposed:
    a. Consumer surplus
    b. Producer surplus
    c. Ticket scalpers' revenue.
    d. Deadweight loss

11. Suppose the weekly quantity demanded (Q^D) for a good is given by the equation $Q^D = 10{,}000 - 80P$, and the weekly quantity supplied (Q^S) is given by $Q^S = 20P$, where P is the price per unit.
    a. What is the equilibrium price and quantity?
    b. When the market is in equilibrium, what are the values of consumer surplus, producer surplus, and total benefits? (Hint: Sketch a rough graph first.)
    c. Find the value of the deadweight loss (dollars per week) if a price ceiling of $80 is imposed on this market.
    d. Find the value of the deadweight loss (dollars per week) if a price floor of $110 is imposed on this market.

CHAPTER 15

Government's Role in Economic Efficiency

In nations around the world, virtually every disagreement about the economy ultimately leads to government. And some disagreements start there as well.

In the United States, for example, hardly a day goes by without a speech in Congress attacking or applauding the government's spending on defense, education, environmental programs, and more.

Underlying many of these speeches are sharp disagreements about the *role* that government should play in economic life. Should it help people send their children to private schools by giving them vouchers? Should it discourage the merger of two large airlines? Should local governments collect trash and run prisons, or should these services be contracted out to private firms? How far should the government go in regulating the activities of private businesses? Similar controversies exist in other developed market economies, such as the nations of the European Union or Japan.

But these disagreements tend to obscure a remarkable degree of *agreement* about the economy and the government's role in it. For example, in the United States and most other countries, the vast majority of goods and services you buy are provided by private firms. Almost everyone agrees that's how it should be. Hardly anyone proposes that the government should provide the economy's books, jeans, computers, entertainment, or soft drinks. At the same time, there is widespread agreement that certain goods and services should be provided by government alone, such as general police protection, the court system, and national defense. Much of this agreement is based on ideas about economic efficiency.

In the previous chapter, we looked at ways in which government intervention in an otherwise efficient markets *reduced* total benefits and made them inefficient. In this chapter, by contrast, we'll be looking at how government can *increase* total benefits in a market, and create efficiency where it would not otherwise exist.

The Legal and Regulatory Infrastructure

The word *infrastructure* in this section's title suggests roads, bridges, airports, waterways, and the like. Indeed, this sort of *physical* infrastructure, often provided by government, is vital for a well-functioning market system.

But equally important is the government-provided *institutional* infrastructure: the legal and regulatory framework without which markets could function only primitively, if at all.

The Legal System

Laws are important for reasons that go far beyond economics. The law protects us from many kinds of physical and emotional harm, guarantees us freedom of speech and other vital civil liberties, and helps provide a sense of security and dignity in our lives. But people often overlook the purely *economic* role of law—ways in which it supports markets and helps us achieve economic efficiency.

Criminal law, for example, makes it illegal to engage in many types of involuntary exchange—such as robbery—in which one party is harmed. In this way, criminal law encourages people to channel their efforts into mutually beneficial, voluntary exchanges—that is, into Pareto improvements.

Property law enables society to assign ownership to assets (such as land, housing, capital equipment, or financial assets) and determine who is entitled to the rewards. This, in turn, encourages people to find the most productive uses for their property, rather than spend time trying to capture the property of others, or prevent others from taking the property they have.

Contract law enables parties to exchange promises involving future activities. It specifies what sorts of promises can be made, and establishes procedures for compensation if one party breaks the promise. Without well-enforced contract law, much productive activity would come to a halt, because people would be wary of fulfilling their side of a bargain first. For example, you would not be able to hire someone to fix your roof because the roofer would insist on payment first, and you would insist he do the work first. Similarly, you would not invest in a company, because you wouldn't trust that you'd receive your share of the future profits.

Tort law defines obligations among people who are *not* linked by contracts. It defines the types of harm for which someone can seek remedy, and the procedures for injured persons to collect reasonable compensation. Tort law helps protect consumers from unsafe products (such as automobiles with brakes that may fail). It also helps protect businesses from unreasonable liabilities (you can't sue Spalding if you trip on a tennis ball). U.S. tort law has become especially controversial in recent years, with one side claiming that costly, frivolous lawsuits are restraining beneficial production, and the other stressing that only the threat of lawsuits keeps a business's eye on consumer safety. But without some form of tort law, many Pareto improvements would never take place; virtually all production carries the risk of harm to someone.

Antitrust law is designed to prevent businesses from engaging in behavior that limits competition and harms consumers. For example, the *Sherman Act* of 1890 prohibits collusion to fix prices, as well as certain types of competitive behavior that can lead to a monopoly. The *Clayton Act* of 1914 gave the federal government the power to prevent mergers and acquisitions that would harm competition. Because (as you've learned) monopoly behavior is inefficient, antitrust enforcement that preserves competition can contribute to economic efficiency.

Regulation

Regulation is fundamentally different from legal procedures. The legal system imposes fines or other penalties when a business violates the law. But regulation goes further. It directs businesses to take some specific actions and prohibits other actions, often on a case-by-case basis. This can contribute to efficiency by protecting buyers, sellers, or third parties from some of the potential harm that market exchanges might otherwise cause.

For example, in the United States, the Food and Drug Administration (FDA) can prohibit a pharmaceutical company from selling a particular drug, or order that further research be done to prove its safety or effectiveness. The Environmental Protection Agency (EPA) has detailed control over what substances a business can release into the atmosphere or into the water.

Regulation is controversial. Tighter regulations can help protect the public from harm, but if they are too strict, they can stifle innovation. The FDA, for example, is criticized for delays in approving new drugs, and for approving too quickly when a drug later proves harmful.

After the financial crisis of 2008, the regulation of financial institutions took center stage. Some economists blamed the crisis itself on weak and inconsistent financial-sector regulations, or the failure of government officials to broadly apply existing regulations. We'll have more to say about the crisis, and the new regulatory proposals, in the Using the Theory section at the end of this chapter.

The Importance of Infrastructure

Almost every Pareto improvement we can think of relies on the legal and regulatory components of our institutional infrastructure. Recall the last time you bought a meal in a restaurant. If you paid cash, the criminal law against counterfeiting enabled the restaurant to more readily accept your paper currency. If you paid by credit card, contract law assured the restaurant that it would eventually be paid by the credit card company. The restaurant itself couldn't function without contracts with its suppliers, landlord, and employees. And you could be reasonably confident that the food was not contaminated, in part because of inspections by local regulatory agencies and also because tort law provides legal disincentives for harmful products.

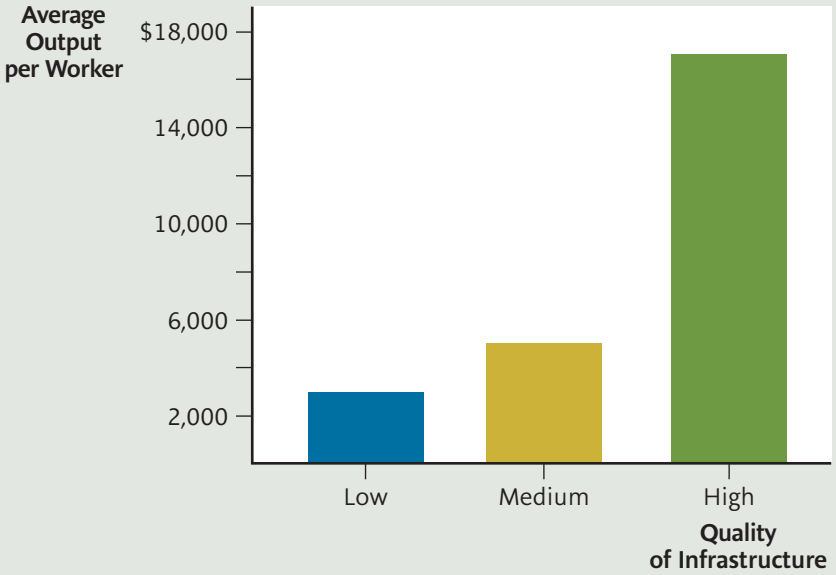
Americans take their institutional infrastructure almost completely for granted. The best way to appreciate the infrastructure of the United States is to visit a country that has a poor one. In many countries, the police are more likely to steal from citizens than to protect them from thievery. Many local economies, and a few national ones, are dominated by powerful mafias that extort protection money from businesses and threaten government officials when they try to enforce the law. The result is serious economic inefficiency that reduces living standards.

Figure 1 shows that when countries are divided into three groups, according to the quality of their institutional infrastructure, there is a strong relation between infrastructure and output per worker. The countries on the left—the ones with the lowest quality infrastructures—were able to produce only about $3,000 in output per worker per year in 1988. These are the nations where property rights were weak, contracts were not enforced, and the government was more often predator than protector of economic activity. In the middle of the figure are countries with medium-quality infrastructures, averaging about $5,500 in output per worker per year. On the right are the best-organized countries, averaging $17,000 in output per worker. Within this group, nations with the very best infrastructures—such as the United States, Sweden, and Japan—achieved output levels more than double that average.

More recent research has confirmed results like those in the figure. And a recent World Bank[1] study estimates that national legal and regulatory infrastructures

[1] *Where Is the Wealth of Nations? Measuring Capital for the 21st Century,* The World Bank (Washington, DC), 2006.

FIGURE 1 Government Infrastructure and Output per Worker



Countries with low-quality infrastructures produced an average of only $3,000 per worker per year in 1988. These countries tend to have corrupt governments, poor enforcement of contracts, and weak property rights. Countries with higher quality infrastructures, including the United States, produced an average of $17,000 per worker per year.

Source: Robert E. Hall and Charles I. Jones, "The Productivity of Nations," National Bureau of Economic Research Working Paper 5812, November 1997.

(measured using a "rule of law" index for each country) accounts for about 44% of the world's productive wealth—about the same percentage as the world's physical and human capital combined.

Market Failures

So far, we've been discussing how the legal and regulatory infrastructure operates in the background, creating fertile ground for markets to operate and generate Pareto improvements. But this infrastructure also enables government to operate more aggressively in specific situations in which a market, left to itself, remains inefficient. These situations are called *market failures*.

*A **market failure** occurs when a market—even with the proper institutional support—is economically inefficient.*

Market failure A market that operates inefficiently without government intervention.

In the remainder of this chapter, we'll focus on four general types of market failures to which economists have devoted a lot of attention. These are:

- monopoly markets
- externalities
- public goods
- information asymmetry

Although economists agree in theory on what causes each of these market failures, addressing them in the real world can create new problems. Dealing with market failures remains one of the most controversial aspects of microeconomic policy.

Monopoly

In Chapter 14, you learned that a monopoly market, or a market in which firms have significant market power, is inefficient. In such markets price is too high, and output is too low, to maximize the total benefits achievable: a market failure.

Is there a proper role for government to make such markets more efficient, and thus cure a market failure?

That depends on the nature of the market, and the reason for the monopoly power.

Potential Remedies for Monopoly Power

The easiest monopoly market for government to address is one that would function very well under more competitive conditions. An example was presented earlier, in Chapter 14 (Figure 9). In that example, a perfectly competitive market for wheat was taken over by a monopoly, which then raised the price and reduced output. In such a case, the government could use *antitrust law* to break the monopoly into several competing firms.

But in the real world, monopolies tend to arise in markets that would *not* perform well under competitive conditions. In these cases, simply breaking up the monopoly could make things worse.

For example, monopolies that arise from patents and copyrights, as discussed in Chapter 10, provide an incentive for artistic creations and scientific discovery. Breaking up a monopoly in, say, a particular drug—by removing its patent before it expired—would lead to a greater and closer-to-efficient quantity of *that* drug. But it would also reduce incentives to develop *future* drugs. Over a long period of time, the benefits from the drug industry as a whole could be reduced. Drug prices are controversial: There are hot debates about the duration of drug patents and what should qualify as a patentable drug. But no one seriously proposes destroying temporary drug monopolies by eliminating patents or their equivalent entirely, and turning the market into anything resembling perfect competition.

Similarly, market power that arises from *network externalities*—discussed in Chapter 10—offers benefits that would be hard to achieve under more competitive conditions. Microsoft, for example, takes advantage of its market power in several ways. But the Windows network, which provides substantial benefits, is possible because a single firm produces the operating system used by most personal computers.

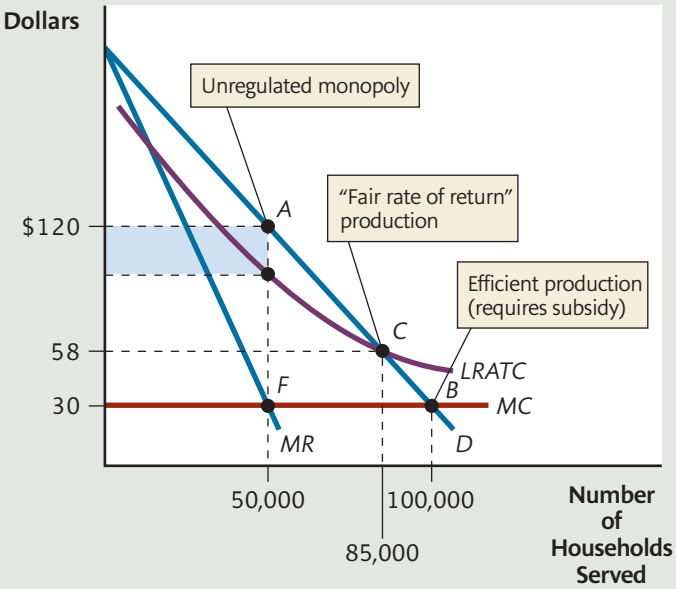
Finally, when a monopoly arises as a *natural monopoly,* using antitrust law to break it up or even to prevent its formation in the first place is a poor remedy. Because this type of monopoly presents special challenges, it's worth its own discussion.

The Special Case of Natural Monopoly

In Chapter 10, you learned that a *natural monopoly* exists when, due to economies of scale, one firm can produce for the entire market at a lower cost per unit than could two or more firms. If the government steps aside, such a market will naturally evolve toward monopoly.

Figure 2 presents an example of a natural monopoly: a local cable company in a small city. The company has important *lumpy inputs*, including costly underground cable, a legal department to negotiate contracts with entertainment

FIGURE 2 Regulating a Natural Monopoly



Left unregulated, the cable monopoly would serve 50,000 households, where MC = MR. *This is inefficient, because units 50,001 to 100,000 have value to some consumers greater than their marginal cost.*

By mandating a price of $30, government regulators could achieve the efficient outcome—100,000 households—at point B. *But with price less than* LRATC, *the monopoly would suffer a loss, so it would have to be subsidized or go out of business.*

The alternative, which is typically chosen, is to set price at $58—the lowest achievable average cost in this market, which includes a "fair rate of return." At this price, the monopoly serves 85,000 households—not quite efficient, but closer than without regulation.

providers, and more. Spreading these costs over more subscribers creates economies of scale. In the figure, these economies of scale continue until the entire market is served: The *LRATC* curve slopes ever downward.

To serve an *additional* household, however, is *not* very costly to this cable company: just the installation appointment, some additional above-ground cable, and occasional customer service. Therefore, *marginal* cost is relatively low. In Figure 2, we assume for simplicity that marginal cost is a constant $30 per additional household—per month, no matter how many households are served. Accordingly, the marginal cost curve is a horizontal line at $30.

If the market for cable service in the city is left to itself, one firm would become the sole supplier. It would sign up the profit-maximizing number of households, where the *MR* and *MC* curves intersect. In the figure, the cable monopoly will sign up 50,000 households. The price, at point *A* on the demand curve, is $120 per month. Profit is the area of the shaded rectangle.

But point *A*—with an output of 50,000—is *inefficient*. The 50,001st through the 100,000th households still value cable service more than the additional cost of providing it to them, so total net benefits in the market would rise if they acquired service. But when the monopoly is left at liberty to charge $120 for service, these households do not buy it. In fact, the efficient level of output is 100,000 households, at point *B*, where the *MC* curve crosses the demand curve. The monopoly—by failing to provide this quantity—is a market failure. What can the government do?

Using antitrust law to break this natural monopoly into several competing firms would not make sense. With several firms—each supplying to only a part of the market—each firm's cost per unit would be even higher than the monopoly's cost per unit. Therefore, the price in a more competitive market could never be $30, so we would not get the efficient quantity of 100,000. In fact, competition, by raising cost per unit, might result in an even *higher* price and *lower* quantity than under monopoly.

But if breaking up a natural monopoly is not advisable, what *can* government do? One option is public *ownership* and *operation* of cable service, as is done with the post office, another natural monopoly. Public takeover of private business is rare, except when certain conditions are present (to be discussed later in this chapter). That leaves one other option, and the one local governments actually choose for the cable industry: *regulation*.

Regulation of Natural Monopoly

At the beginning of this chapter, you learned that under regulation, a government agency digs deep into the operations of a business and takes some of the firm's decisions under its own control. In the case of a natural monopoly, regulators would tell the firm what price it can charge.

Marginal Cost Pricing

Marginal cost pricing Setting a monopoly's regulated price equal to marginal cost where the marginal cost curve crosses the market demand curve.

One option for regulators is **marginal cost pricing:** setting the price equal to the firm's marginal cost of production.[2] In the figure, where marginal cost is constant at $30, the regulators would set price equal to $30. This will automatically bring the market to the efficient level of output: At a price of $30 for service, 100,000 households will sign up.

But there's a problem. If you look again at Figure 2, you'll notice that the *MC* curve lies everywhere *below* the *LRATC* curve. This must be the case for a natural monopoly, since economies of scale—the reason for the natural monopoly—means that the *LRATC* curve slopes downward, and this can occur only when marginal cost is less than average cost. (See Chapter 7 on the marginal–average relationship if you've forgotten why. Here, both marginal cost and average cost refer to the long run.)

Now you can see the problem for regulators: If they set the efficient price of $30, then cost per unit (on the *LRATC* curve) will be greater than the price. The firm suffers a loss, and in the long run, it would go out of business. Therefore, with marginal cost pricing, the government must also *subsidize* the natural monopoly—to cover its losses with government funds. This is often controversial because it requires taxpayers in general, rather than just the monopoly's customers, to help pay for the product.

Average Cost Pricing

Average cost pricing Setting a monopoly's regulated price equal to long-run average cost where the *LRATC* curve crosses the market demand curve.

Because of the problems with marginal cost pricing, regulators around the world more commonly choose an alternative, called **average cost pricing.** With this method, the price is set as low as possible—so serve as many customers as possible—while

[2] In the figure, marginal cost is constant. When marginal cost rises with output, efficiency requires the regulator to set price equal to the value of marginal cost *where the marginal cost curve crosses the demand curve*. At this price, every unit valued more than its marginal cost would be purchased.

still covering the firm's cost per unit. Such a price is found where the *LRATC* curve crosses the demand curve ($58 in the figure).

Average cost pricing is sometimes called *fair rate of return* pricing. To see why, remember that average total cost incorporates *all* costs, including the opportunity cost of owners' funds. Thus, a fair rate of return to owners is already built into the *LRATC* curve. When the firm charges $58 per unit, it is covering *all* of its production costs—including a fair rate of return for owners. Moreover, $58 is the lowest price the natural monopoly could charge without suffering a loss. (Confirm for yourself that if it charged any less, it would operate at a point on the demand curve *below* the *LRATC* curve, creating a loss.)

More generally,

with average cost pricing, regulators strive to set the price equal to cost per unit where the LRATC *curve crosses the demand curve. The natural monopoly makes zero economic profit, which provides its owners with a fair rate of return and keeps the monopoly in business.*

Average cost pricing does not quite make the market efficient. For example, in Figure 2, only 85,000 units are produced, instead of the efficient quantity of 100,000. But compared to no regulation at all, average cost pricing lowers the price to consumers and increases the quantity they buy, bringing us closer to the efficient level.

Another problem with average cost pricing is that it provides little or no incentive for the natural monopoly to control costs. The monopoly can grow larger and larger, buying more machinery, office buildings, and other forms of capital—confident that the regulators will always adjust the price upward to ensure a fair rate of return. This can lead to rising costs for customers, and a further movement away from the efficient output level.

Externalities

If you live in a dormitory, you have no doubt had the unpleasant experience of trying to study while the stereo in the next room is blasting through your walls—and usually not your choice of music. This may not sound like an economic problem, but it is one. The problem is that your neighbor, in deciding to listen to loud music, is considering only the private costs (the sacrifice of his own time) and private benefits (the enjoyment of music) of his action. He is not considering the harm it causes to you. Indeed, the harm you suffer might be greater than the benefit he gets from blasting his music.

When a private action has side effects that affect other people in important ways, we have the problem of externalities:

An ***externality*** *is a by-product of consumption or production that affects someone other than the buyer or seller.*

Externality A by-product of consuming or producing a good that affects someone other than the buyer or seller.

A *negative externality* is one that causes harm to others, while a *positive externality* creates benefits for others. As you'll see, both types of externalities can create economic inefficiency. We'll consider the case of negative externalities first.

Suppose a negative externality from some activity or transaction causes more harm than it provides in benefits. Then *total* benefits to society (including those

harmed) are reduced. Because total benefits are not maximized, the situation is economically inefficient. However, under certain conditions, private parties can correct the inefficiency on their own, without the government's help.

The Private Solution

Here's a simple example. Every morning at 6 A.M., a noisy, heavy truck takes a shortcut off the main highway, drives along an isolated road past the single house there, and wakes up the resident. The truck driver benefits by using the road—it saves time. The resident is harmed by the loss of sleep.

Suppose the value of harm to the resident is $10 per day (that's what he'd be willing to pay for undisturbed sleep). And suppose that the shortcut saves the truck driver just a few minutes, which he values at $4 (that's the highest toll the driver would be willing to pay if the short-cut were a toll road). In this case, the harm to the resident ($10) is greater than the gain to the driver ($4). Total benefits for "society" (the two people affected) would be $6 greater if the truck driver stayed on the highway. So staying on the highway is the efficient outcome.

In a case like this, if *either party* has clear legal rights over the activity, the outcome will be efficient: the truck driver will stay on the highway. To see why, let's first suppose that it's the *resident* who possesses the legal rights. That is, suppose the shortcut is a private road, and the driver needs the resident's permission to use it. Then the truck will stay on the highway, because the resident will certainly not grant permission. After all, the largest side payment the truck driver would be willing to make to the resident ($4 per day) is less than the minimum side payment ($10) that would make the resident change his mind.

But what if the shortcut is a public road, so the *truck driver* has the legal rights? Then the resident will find it worth his while to make a side payment to the driver for staying on the highway. In fact, any side payment between $4 and $10 would more than compensate the truck driver for the extra travel time of staying on the highway, and still leave the resident with some left-over benefit. For example, a side payment of $7 would make the truck driver $3 better off than when he used the shortcut, and also leave the resident $3 better off than when he was awakened at 6 A.M.

As you can see, regardless of who has the rights, the efficient outcome—the truck stays on the highway—will occur. No government intervention is required, other than the initial establishment of the legal rights.

What if the numbers in this example were changed so that using the shortcut created more gains to the driver than harm to the resident? Then—as you'll see in an end-of-chapter question—the driver will take the shortcut, regardless of which party possesses the legal rights. Once again, the outcome will be the efficient one.

The Coase Theorem

Our example's conclusion—that the outcome will be efficient regardless of which party holds the rights—is a rather surprising result. It is an example of the *Coase theorem,* named after economist Ronald Coase:

Coase theorem When a side payment can be arranged without cost, the market will solve an externality problem—and create the efficient outcome—on its own.

> *The* ***Coase theorem*** *states that—when side payments can be negotiated and arranged without cost—the private market will solve the externality problem on its own, always arriving at the efficient outcome. The allocation of legal rights determines gains and losses among the parties, but does not affect the action taken.*

The Coase theorem points out that externalities do not always create market failures, and that private parties—with proper institutional backing—can sometimes work things out themselves.

However, note that the Coase theorem requires that side payments can be arranged *without cost*—or, in practice, that the cost is so low relative to the gains or losses at stake that it doesn't matter. This requirement is most likely to be satisfied when all three of the following conditions are present: (1) legal rights are clearly established; (2) legal rights can be easily transferred; and (3) the number of people involved is very small.

However, many real-world situations do not satisfy these conditions. Legal rights are often in dispute. If the rights are vaguely defined in our example, the truck driver and the resident might very well end up in court. Since courts are often more concerned about fairness or legal interpretation than efficiency, the outcome may not be the efficient one. (This is not necessarily bad; fairness, as we've stressed, is a concern as well as efficiency.)

Furthermore, once a court decides the issue, legal rights may not be transferable. For example, suppose using the shortcut *was* efficient, but the resident won the case. Then the court may forbid the truck from taking the shortcut, and prohibit the resident from granting permission if the driver later offers him a side payment.

But the biggest problem in applying the Coase theorem is the third condition. In most real-world situations, a large number of people are involved. In our truck example, there could be hundreds of trucks traveling along the shortcut and hundreds of residents disturbed by the noise. Determining the gains and losses for each one, getting them all together, and coming up with a mutually agreeable solution would be very costly. Moreover, when many people are involved, achieving efficiency through side payments is plagued by an often insoluble problem, to which we turn now.

The Free Rider Problem

Suppose that there are 100 trucks and 100 residents, and that the efficient outcome is for the trucks to stay on the highway. But because the shortcut is a public road, the drivers have the legal right. Representatives for each side make an arrangement: the residents will pay a total of $600 per day to the truck drivers (so $6 per day for each driver) who, in turn, will stay on the highway. This arrangement makes *everyone* better off. Each resident is asked to contribute $6 per day to the side-payment fund.

Now, we face a problem: A resident may try to get a *free ride,* refusing to pay. After all, his own part of the side payment is so small—just a "drop in the bucket"—that the arrangement will work with or without him. He may claim that the noise doesn't really bother him that much, or just laugh off anyone who comes to collect. (Remember, the government is not involved at this point, so there is no legal body to enforce the agreement.) If *many* of those who benefit from the agreement reason this way and refuse to contribute, we have the *free rider problem*.

*The **free rider problem** occurs when the efficient outcome requires a side payment but some or all individuals—each obligated to pay a small share of the total—will not contribute.*

Free rider problem When the efficient outcome requires a side payment but some individuals will not contribute.

The free rider problem, if extensive enough, can shrink the total side payment until it is too small to compensate those harmed. In that case, the private

arrangement—based on voluntary participation rather than government coercion—will break down and the efficient outcome will not be achieved. Indeed, the free rider problem stands in the way of many Pareto improvements. And it is one of the main reasons we typically turn to government to deal with important externalities that affect many people.

Government and Negative Externalities

Competitive markets, by definition, involve many buyers and sellers. And as we've just seen, with so many people involved, the private (Coase theorem) solution to a negative externality may not work. This is why economists frequently advocate for *government* involvement in competitive markets with negative externalities.

Consider, for example, the negative externality from driving. When people drive their cars, they create local pollution, global greenhouse gas emissions, contribute to traffic congestion, and increase the probability of an accident. Because driving is so closely correlated with gasoline use, we can view these as negative externalities that arise in the market for gasoline. While estimates vary, research suggests that the costs imposed on third parties from driving in the United States can be reasonably valued at about $1 per gallon of gasoline consumed.[3] Additional accidents and road congestion are the two largest factors.

Figure 3 shows the market for gasoline in the United States, with hypothetical supply and demand curves. (Ignore the curve labeled *MSC* for now.) Before the government gets involved, the market reaches equilibrium at point *A*, where supply curve *S* and demand curve *D* intersect. But the equilibrium quantity, 400 million gallons per day, is *not* efficient. Why not? Recall (from Chapter 14) that the height of the supply curve at any output level tells us the cost to *producers* of providing gasoline. It does *not* reflect any of the costs to the general public from accidents, pollution, etc. because—without government involvement—producers do not *pay* these costs.

Let's suppose that each gallon of gasoline involves a negative externality of $1. If we add this to the marginal cost actually paid by gasoline producers, we get the **marginal social cost** (*MSC*) of another unit of gasoline. *MSC* includes *all* costs of producing another unit of gas: the resources used up and paid for by the industry, *and* the costs imposed on third parties. This is why the *MSC* curve in Figure 3 lies *above* the market supply curve.

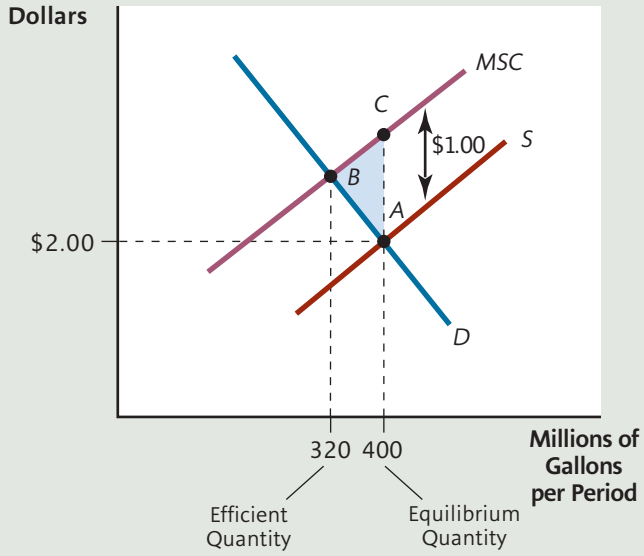
Marginal social cost (MSC) The full cost of producing another unit of a good, including the marginal cost to the producer *and* any harm caused to third parties.

Now let's find the efficient quantity of gasoline. Efficiency requires that the market provide only those units that are valued more highly by consumers than their cost. The *MSC* curve tells us the true, full cost of gasoline, when *all* costs are considered. For all units up to 320 million, the demand curve lies above the *MSC* curve, so those units are more highly valued than their costs. Total benefits in the market are increased by providing the first 320 million gallons.

But for all units *beyond* 320 million, the *MSC* curve lies above the demand curve. These units should *not* be produced, because they cost more—in the fullest sense—than their value to consumers. Each time a unit beyond 320 million is produced, total benefits—to producers, consumers, *and* society—shrink. Thus, the efficient quantity of gasoline—at which total benefits are maximized—is 320 million gallons

[3] See, for example, Ian W. H. Parry and Kenneth A. Small, "Does Britain or the United States Have the Right Gasoline Tax?" *American Economic Review*, Vol. 95, No. 4, September 2005.

FIGURE 3 Inefficiency from a Negative Externality



In the competitive market for gasoline the market equilibrium is at point A *with quantity at 400 million. But the supply curve includes only marginal costs to private suppliers, ignoring the negative externality of $1.00 per gallon. When the negative externality is added to other costs, the result is the marginal social cost curve* (MSC). *The efficient output level (320 million) is at point* B. *Without government intervention, the units from 320 million to 400 million are produced, even though their full cost (given by the* MSC *curve) is greater than their value (on the demand curve). The result is a deadweight loss equal to the area of the shaded triangle.*

per day. Alternatively, because each gallon of gasoline creates $1 in harmful externalities, we can identify the "efficient amount of harm"—the amount of harm that maximizes total benefits in the market—as $1 × 320 million = $320 million per period. The labeled triangle *ABC* shows the *deadweight loss* for this market in equilibrium—the loss in benefits from producing too much output.

A market with a negative externality will produce more than the efficient quantity of the good (and therefore, more than the efficient quantity of the harmful byproduct), creating a deadweight loss.

How can we deal with this inefficiency?

Regulation

Until recently, the most common method used in the United States for dealing with negative externalities—especially pollution—has been regulation. Federal, state, and local governments limit the amounts of certain pollutants released into the water and air by individual firms, and mandate pollution control devices on automobiles. If we wanted to take a purely regulatory approach to the gasoline market, we could limit the amount of gasoline each producer could produce, or that each consumer could use.

For example, if there are 100 million gasoline consumers, we could limit each one individually to 3.2 gallons of gas per day, achieving the efficient quantity of

320 million gallons. However, while this approach would give us the efficient total *quantity,* it would still leave significant inefficiency. That's because the total quantity would not be *allocated* efficiently among consumers.

Suppose Consumer A values gasoline very highly and would be willing to pay $10 for another gallon each day beyond the 3.2 gallons allowed. And suppose Consumer B values gasoline very little and would be willing to sell a gallon each day for $1. At any price between $1 and $10, a trade between consumers A and B would be a Pareto improvement. But the regulation—3.2 gallons per day for each consumer and no more—prevent this trade from occurring. We are leaving Pareto improvements unexploited.

A similar problem would occur if we used regulation to dictate gasoline *production* for each refining company. Suppose high-cost producers and low-cost producers are each told how much they are allowed to produce. Then they will not have to compete for business in the marketplace, so high-cost producers will end up producing gallons that could be produced more cheaply by low-cost producers. As a result, we would not be producing the given quantity of gasoline at the lowest possible cost.

Ordinarily, in a competitive market, we don't need to worry about efficient allocation of goods among consumers and producers because the market does this automatically. In equilibrium, consumers can buy all the units that are worth more to them than the market price, and no consumers will buy any units that are worth less to them than the market price. Similarly, any firm that can produce units at a cost less than the market price will produce them, and any units that would cost more than the market price will not be produced. But with regulation, there is no such guarantee.

This inefficiency in allocating production and consumption among market participants is why economists often favor a different type of solution to negative externalities. With a *market-based approach,* we can achieve the efficient market quantity of a good (and therefore, the efficient market quantity of the externality), but with an important advantage: We can also ensure that we're exploiting all Pareto improvements in *allocating* that quantity among market participants.

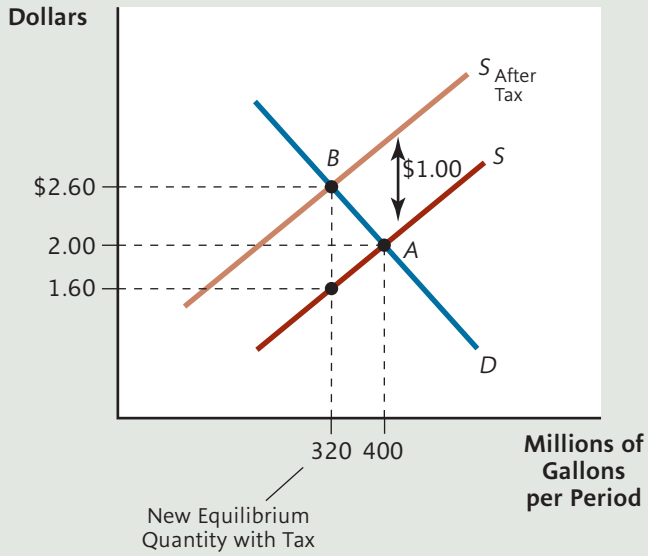
Let's consider the two leading market-based approaches to negative externalities favored by economists.

Market Based Approach #1: Taxes

One market-based approach is to impose a tax per unit equal to the negative externality created by each unit of the good. In the market for gasoline, for example, we've assumed that each gallon creates $1 of harm, so we could impose a tax on gasoline of $1 per gallon. As you learned in Chapter 4, it makes no difference whether we impose the tax on gasoline sellers or buyers—the effect on equilibrium quantity, and on the price for each side of the market, is the same in either case. But for negative externalities, imagining that the tax is imposed on sellers allows for a more straightforward discussion.

Figure 4 illustrates the impact of a $1 per gallon tax on gasoline sellers. As you've learned (Chapter 4), the tax shifts the supply curve upward by $1. If you compare Figure 4 with Figure 3, you'll see that the tax shifts the supply curve until it is the same as the *MSC* curve. Once the tax is in place, the market reaches a new equilibrium at point *B*. The efficient quantity of gasoline (320 million gallons per day) is produced, resulting in the efficient amount of harm from the externality ($1 × 320 million = $320 million).

FIGURE 4 A Tax on Producers to Correct a Negative Externality



The negative externality is corrected with a tax on suppliers of $1.00—equal to the negative externality per unit. The supply curve shifts upward by the amount of the tax, moving the market equilibrium to the efficient point B. *Only units valued at least as much as their full cost—the first 320 million units—are produced.*

So far, the tax achieves just what regulation did. But notice one further result when we use the tax: Every gallon of gasoline valued by any consumer at more than $2.60 per gallon is purchased by that consumer. And no one ends up with gasoline that they value less than $2.60 (because no one would pay $2.60 in that case). So there are no further Pareto-improving trades among gasoline consumers. Moreover, every gallon of gasoline that can be produced and sold by any firm at less than $2.60 per gallon is being produced and sold. And no firm ends up selling any gallons that cost more than $2.60 to produce, because it would not be profitable to do so. So there is no reallocation of production among firms that could reduce costs for society.

A tax on each unit of a good equal to the external harm it causes can correct a negative externality, and bring the market to the efficient quantity of the good. Moreover, using a tax (rather than limits on individual firms and consumers) assures that the total market quantity of the good is allocated efficiently among consumers and producers.

Notice that, in the new equilibrium after the tax, each consumer of gasoline is paying 60 cents more than initially, and each seller is receiving 40 cents less than initially for each unit. In this sense, both buyers and sellers share the burden of paying for the harm to society caused by each unit of gasoline sold. And society—represented by the government—receives the tax revenue, equal to $1 × 320 million = $320 million per day. This revenue can be used for environmental cleanup, reducing other taxes, or providing other government services valued by society. In effect, if used properly, the tax can serve as a kind of side payment from gasoline buyers and sellers to those (society in general) harmed by its use.

In our example so far, we've imposed a tax on a *good*—gasoline—whose use is associated with a negative externality. But a tax can also be imposed directly on the

harmful externality itself—say, on each unit of pollution emitted by each firm. A tax more narrowly focused on the harm itself creates additional incentives: to alter technology in order to reduce the harm from each unit. Gasoline producers, for example, would have an incentive to change the chemical composition of gasoline so it creates less pollution. Only those producers that can reduce pollution for less than the cost of paying the tax would have an incentive to do so. In this way, the tax ensures that any given amount of emissions reduction is achieved at the lowest possible cost.

A direct tax on the negative externality itself (like a tax on the associated good) can bring the market to the efficient quantity of the good. But taxing the externality has an added advantage: It encourages technological change that reduces the harm from each unit of the good produced or consumed.

Many countries around the world have taxed a variety of negative externalities, including the pollution caused by discarded plastic bags and batteries, and atmospheric pollutants like nitrous oxide, and carbon dioxide emissions associated with global warming. However, another market-based approach has become increasingly popular, especially for atmospheric pollution associated with climate change.

Market Based Approach #2: Tradable Permits

Tradable permit A license that allows a company to release a unit of pollution into the environment over some period of time.

A **tradable permit** is a license that allows a company to release a unit of pollution into the environment over some period of time. By issuing a fixed number of permits, the government determines the total level of pollution that can be legally emitted each period. However, firms can sell their government-issued permits to other firms in an organized market.

Because the permits are tradable and sell for a price, a firm faces an opportunity cost for each unit of pollution that it creates: Either it must buy a permit, or it must forgo the revenue it *could* earn by selling the permit to some other firm. Thus, the price of the permit becomes part of the marginal cost of producing the polluting good. With higher marginal cost, the market supply curve shifts upward, as in Figure 4, and the market equilibrium price of the polluting good rises as well. This can move the quantity of a polluting good toward its efficient level, similar to the effects of a tax.

And, also like a tax, tradable permits allow consumers and firms to respond to prices when determining how much of the negative externality to create individually. Overall efficiency is ensured by the limit on the total number of permits issued.

Let's take a closer look at how this works for producers.

A firm whose technology would make it very costly to reduce pollution generally *buys* permits in the market. By buying a permit at a price lower than its cost of reducing pollution by another unit, the high-cost firm comes out ahead. At the same time, a firm whose technology enables it to reduce pollution rather cheaply will *sell* permits. By giving up permits, the low-cost firm takes on the obligation to reduce its pollution further. But by selling the permit at a price greater than its pollution-control cost, the low-cost firm gains as well.

dangerous curves

Taxes, Externalities, and Deadweight Losses In the previous chapter, our examples showed how taxes *create* a deadweight loss. In this chapter, you've seen an example in which taxes *eliminate* a deadweight loss. So which is it?

The answer is: A tax can do either, depending on the nature of the market in which the tax is imposed. In the previous chapter, we dealt with competitive markets that were *otherwise economically efficient*. Those markets were not characterized by any negative externalities. In such markets, a tax will *reduce* total benefits and create a deadweight loss. In this chapter, we're looking at competitive markets that are *not* otherwise efficient, because of a negative externality. In such markets, a tax of the right amount can *increase* total benefits and eliminate the deadweight loss.

The trading of permits shifts the costs of any given level of environmental improvement toward those firms that can do so more cheaply. The general public, however, is not affected by the trade because total pollution remains unchanged. Therefore, for any given level of pollution, allowing firms to buy and sell licenses generates Pareto improvements. Viewed another way, tradable permits—by making it cheaper to lower pollution—enable the government to impose stricter environmental standards with the same total burden on producers.

A system of tradable permits for pollution, like a tax, can make the market quantity of polluting goods efficient. Moreover, unlike limits on individual consumers and producers it allows the total market quantity of the good, and of pollution reduction efforts, to be allocated efficiently among buyers and sellers.

Tradable permits (often called "cap and trade") have been used since the early 1980s to reduce several types of pollution in the U.S. Permits for adding lead to gasoline virtually eliminated leaded gasoline from the market within 5 years. And a system of tradable permits begun in 1990 for sulfur dioxide (the pollutant that causes acid rain) cut emissions in half within 5 years—well ahead of schedule and at much lower cost than anticipated. In 2005, the European Union established a system of tradable permits for several greenhouse gases. Many countries also participate in a global system of credits that functions much like tradable permits. The Dutch government, for example, has constructed a modern refuse plant in Brazil to reduce methane emissions from more primitive trash dumps. In this way, the Dutch obtain credits for reducing emissions at a fraction of what it would cost to reduce them at home.

In June, 2009, the U.S House of Representatives—in consultation with the Obama administration—passed the first tradable permits program to limit greenhouse gas emissions in the United States. The bill called for a system of tradable permits to bring about a 17% cut in greenhouse gas emissions by 2020, and an 83% cut by 2050. If passed by the Senate, tradable permits would become the policy of choice to reduce greenhouse gas emissions in the United States.

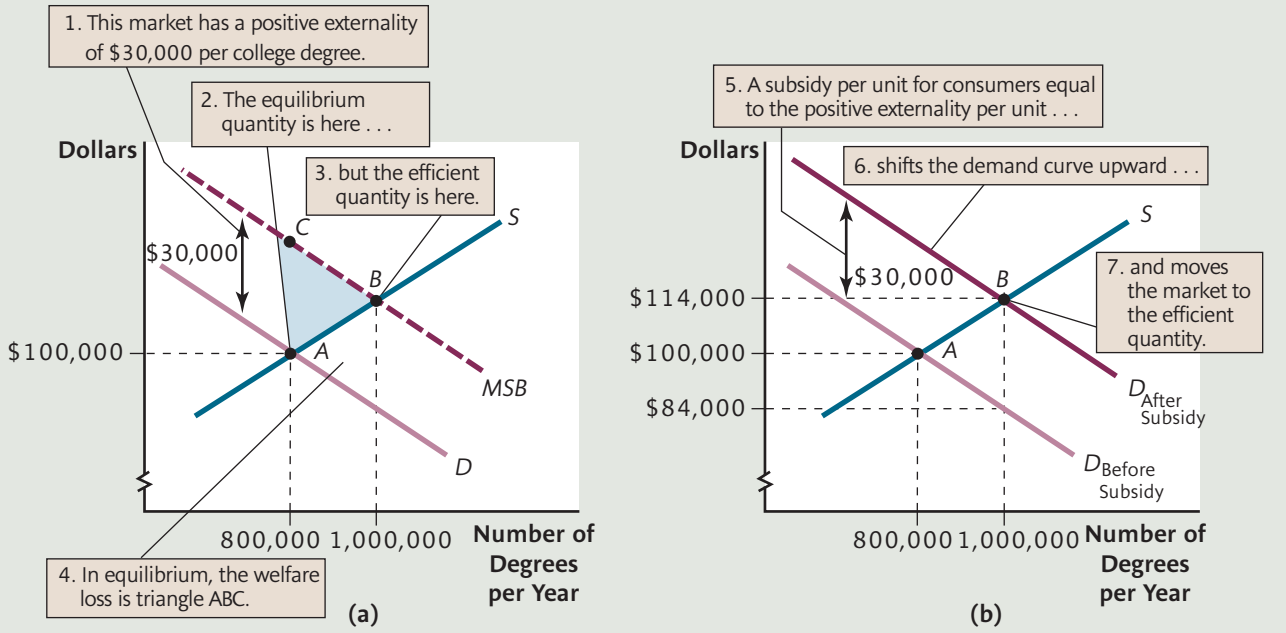
Positive Externalities

Now we move away from negative externalities to examine positive ones. For a positive externality, the by-product of the good or service *benefits* other parties, rather than harms them. A homeowner who plants flowers in the front yard creates enjoyment for everyone who walks by and helps to raise property values in the neighborhood. People who purchase hidden tracking devices for their cars that enable the police to arrest car thieves help to reduce the rate of auto theft. This benefits all car owners—not just those who bought the devices.

Why a Positive Externality Is a Market Failure

It may seem strange to think of activities that benefit other people as "failures," but remember that a market failure is an *inefficient* market—one that leaves Pareto improvements unexploited, and thus wastes the opportunity to make people better off. In the case of a positive externality, the market—left to itself—will produce *less* than the efficient quantity.

FIGURE 5 A Subsidy for Consumers to Correct a Positive Externality



To see why, consider the market for a college education. In deciding whether to attend college, each of us takes into account only the costs to *us* (including tuition, and room and board) and the benefits to *us* (such as a higher-paying and more interesting job later on, or enjoyment of learning for its own sake). But when you become educated, you also create a number of benefits for other people. For example, you will be a more-informed voter and thereby help to steer the government in directions that benefit many people other than you. Or your education may make you more likely to make a scientific discovery or start a business that creates benefits for society in general. Thus, the market for college education involves a positive externality.

Let's see why a competitive market in college education, with no government interference, would be inefficient. Figure 5(a) shows the market for bachelor's degrees. The horizontal axis measures the number of students acquiring bachelor's degrees each year, and the vertical axis measures the price the college charges for each degree.[4] The height of the supply curve *S* reflects the costs to a college or university to provide each degree at colleges and universities, and the height of the demand curve measures the value of each degree *to the student who gets it.* But the demand curve does not reflect any of the benefits that another college degree provides to the general public. Without government intervention, the market reaches equilibrium at point *A*, with 800,000 bachelor's degrees awarded each year, and a four-year price of $100,000.

But 800,000 is *not* the efficient number of degrees. We can see this by incorporating the positive externality into our diagram. Let's suppose that each degree gives

[4] We could also measure price and quantity "per year of education" (as we did in Chapter 4) rather than "per degree" (as we do here). Either way, we come to the same conclusions.

the general public $30,000 in benefits. When we add this benefit to the benefit of the degree holders themselves, we get the **marginal social benefit** (*MSB*) of another degree. *MSB* includes *all* the benefits of acquiring another college degree—the benefits to the student *and* the benefits to society at large. This is why the *MSB* curve in Figure 5 lies above the market demand curve. The distance between the curves is $30,000, the value of the positive externality.

Marginal social benefit (MSB) The full benefit provided by another unit of a good, including the benefit to the consumer *and* any benefits enjoyed by third parties.

Once we draw the *MSB* curve in Figure 5(a), we discover that the efficient output level in this market is 1 million college degrees, where the *MSB* curve intersects the supply curve at point *B*. Why? Efficiency requires that the market provide any unit that has more value than it would cost to produce. The *MSB* curve tells us what the value of each degree *really* is, when *all* benefits are considered. For all units up to 1 million, the *MSB* curve lies above the supply curve, so those units provide greater value than their costs. Total benefits in the market are increased by providing them. Because the market, left to itself, fails to produce any of the degrees between 800,000 and 1 million, we are wasting opportunities to increase total benefits in the market. In fact, we can measure the benefits given up for any degree *not* provided as the distance between the *MSB* curve and the supply curve for that degree. For example, the 800,001st degree provides benefits of about $130,000 to both the student and society in general, but its cost is only $100,000, so we lose $30,000 in total benefits by not providing that degree. The total deadweight loss from *all* the degrees not provided is the shaded triangle *ABC*.

A market with a positive externality associated with producing or consuming a good will produce less than the efficient quantity, creating a deadweight loss.

How can we achieve the efficient quantity of 1 million college degrees each year? One way is to *change* the market demand curve so it is the same as the *MSB* curve in Panel (a) of Figure 5. If those affected could cheaply negotiate and enforce an agreement, then the Coase theorem would apply and private parties could solve the problem themselves. People who benefit from others' degrees would make a side payment to those in college. This side payment would be included as part of the value of the degree for those considering college, shifting the demand curve upward. With a side payment of $30,000 per degree, the market demand curve would rise to the position of the *MSB* curve, and the market would provide 1 million degrees—the efficient number.

As you might imagine, the time, trouble, and expense of arranging private side payments in such a large market would be prohibitive. And the free rider problem would be unmanageable. (Imagine someone passing the hat asking for contributions to support strangers in college!)

Government Subsidies for Positive Externalities

In a large market, such as the market for college degrees, achieving efficiency may require government involvement. And in such cases, economists once again favor a market-based approach, which changes the price of the product but then allows individual consumers and producers to respond to those prices and make their own decisions. An effective market-based approach—and one used in the United States and many other countries for higher education—is a *subsidy*.

In Figure 5(b) we imagine that a subsidy of $30,000 per degree—the value of the positive externality—is paid to each student. (As you learned in Chapter 4, the impact on price and quantity for each side of the market is the same regardless of

which side initially pays a tax or receives a subsidy.) The subsidy causes each student to add $30,000 to the value that he or she places on a degree, thereby shifting up the demand curve by that amount. The market equilibrium now moves to 1 million degrees per year—the efficient number.

Notice how the subsidy causes buyers and sellers to take account of the positive externality in making their decisions. In the new equilibrium, the price of a degree rises—from $100,000 to $114,000 in the figure. This encourages colleges to increase enrollment. But students, after accounting for the subsidy, pay only $84,000 ($114,000 minus the $30,000 subsidy). This is why more of them choose to acquire degrees.

A subsidy on each unit of a good, equal to the external benefits it creates, can correct a positive externality and bring the market to an efficient output level.

Public Goods

So far, all of our examples of market failures have been goods that are left to the market to provide. The market failure arises because some government manipulation of the price is needed (via a tax, subsidy, or regulation) to induce firms to move the quantity closer to the efficient level.

But some types of goods, by their nature, are especially difficult for the market to provide, or to provide even close to efficiently. Such goods require a more *direct* form of involvement, in which the government itself provides the goods. These are called *public goods*. (A more formal definition follows a bit later.)

Private Goods

The best way to understand *public* goods is to consider their opposite: Private goods. These are goods that share two characteristics: rivalry and excludability.

Rivalry

Rivalry A situation in which one person's consumption of a unit of a good or service means that no one else can consume that unit.

A private good is characterized by **rivalry** in consumption—if one person consumes a unit someone else cannot consume that unit. If you rent an apartment, then someone else will *not* be able to rent that apartment. The same applies to most goods that you buy—food, computers, air travel, and so on. Rivalry also applies to privately provided services: the time you spend with your doctor, lawyer, or career counselor is time that someone else will *not* spend with that professional.

By allowing the market to provide rival goods at a *price,* we ensure that people take account of the opportunity costs to society of their decisions to use these goods. If they were provided free of charge, people would tend to use them even if their value were less than the value of the resources used to produce them. Moreover, offering a rival good free of charge would enable some people who don't value it very highly to grab up all available supplies, depriving others who might value the goods even more. Thus, leaving such goods to the market—where a price reflecting marginal cost is charged—tends to promote economic efficiency.

Excludability

A second feature of a private good is **excludability**, the ability to exclude those who do not pay for a good from enjoying it. Excludability is what makes it *possible* for a private business to provide a good. After all, without excludability people would never pay, so any private firm that tried to provide the good would quickly go out of business.

Excludability The ability to exclude those who do not pay for a good from consuming it.

Let's sum up the discussion so far: Excludability means that firms *can* provide a good. Rivalry means that a price *should* be charged for a good (something that private firms do automatically). If a good has both of these characteristics, it is called a *pure private good*.

A good that is both rivalrous *and* excludable *is a* ***pure private good****. In the absence of any significant market failure, private firms will provide these goods at close to efficient levels.*

Pure private good A good that is both rivalrous and excludable.

But not all goods or services have these characteristics.

Pure Public Goods

Consider a small, urban park whose chief benefit is that people find it nice to look at when they walk by. To provide and maintain the park is costly. But its benefits are *nonexcludable*—there is no practical way to limit enjoyment just to those who pay.

What's wrong with this picture?

© CREATAS (BASED ON AN IMAGE IN ECONOCLASS.COM BY LORI ALDEN, 2005)

In general, when goods are nonexcludable, private firms will not be able to provide them. Because everyone will be able to enjoy the benefits without paying, no one will pay. This should sound familiar: It's another instance of the *free-rider problem* we discussed earlier, in the context of externalities. Here, the free rider problem prevents private firms from producing and selling a nonexcludable good. The only way to make the park excludable would be to construct a giant fence with a gate and charge admission. But then it's a different good—no longer a walk-by park, but rather a private club that you must go out of your way to enter and spend time in to get the benefit.

The walk-by park is also *nonrival*. One person's enjoyment while passing does not prevent or lessen the enjoyment of anyone else. Moreover, no more of society's resources are used up when another person views it. For this reason, even if a firm *could* figure out a way to charge for the park, charging would be inefficient. Each time an additional person sees the park, a Pareto improvement takes place: That person gains and no one loses. Thus, to be economically efficient, everyone who places *any value at all* on seeing the park should be able to see it. This can only happen if the price is *zero*.

Generalizing from this example: If a good is nonrival, a private firm *should* not provide it. The firm would have to charge a price, while efficiency requires a price of zero. And if the good is also nonexcludable, the private market is usually *unable* to provide it anyway. A good with both of these characteristics is called a *pure public good*.

A good that is both nonrival and nonexcludable is a ***pure public good****. These goods are generally provided by government without charge, because the private market cannot supply them at all, or cannot supply them efficiently.*

Pure public good A good that is both nonrival and nonexcludable.

FIGURE 6 **Pure Private, Pure Public, and Mixed Goods**

Figure 6 illustrates the classification of goods into pure private and pure public, and provides some examples. (Ignore the other two categories for now.) Pure private goods are in the upper left corner. There is general agreement that private firms should provide such goods in the market and charge for them. While government may intervene in these markets, it does not generally provide the goods itself.

In the lower right corner are pure public goods. It is generally agreed that government should provide these goods without charge. A classic example is national defense. It would be virtually impossible to exclude those who did not pay from enjoying most of the benefits of national defense, as long as they remain in the country. And once a given quantity of national defense is provided, extending its benefits to an additional person requires no additional resources, nor does it lessen anyone else's defense.

Mixed Goods

Goods that appear in the other two corners of Figure 6 can be called *mixed goods* because they share features of both public and private goods. These goods are becoming increasingly important in our society.

Marketable public good An excludable and nonrival good. Generally provided by the market for a price, though efficiency would require a price of zero.

Marketable Public Goods

In the upper right corner are goods that are excludable, but nonrival. We'll call these **marketable public goods.** Because their benefits are excludable, the market gener-

ally can, and generally does, provide them at a price. But because they are nonrival, the quantity is less than efficient. Some people decide not to pay the price and don't consume, even though their consumption would not use up resources or lessen anyone else's enjoyment.

An example is a downloadable music or video file. Recent court cases and other efforts have eliminated most unauthorized downloading at no charge. As a result, these files are now excludable: private firms (such as Apple and RealNetworks) offer them for sale, and only those who pay can download them. But they could provide an *additional* music file—or allow customers to share with anyone they wanted—without decreasing anyone else's enjoyment. The same is true of any other form of digital information or entertainment: software, movies, and more. All are provided privately at less than efficient quantities.

In some cases, if firms can rely on advertising for their revenue, we can get closer to the efficient quantity. Broadcast television and Google searches are examples of marketable public goods that are offered privately, at no charge, because they are supported entirely by advertising.

But in most cases, a price is charged, and we accept the resulting inefficiency. Most economists would not even describe a marketable public good as a market failure. It is true that, once such goods are created and we desire them, public provision at no charge would generate the efficient quantity. But having the government deeply involved in providing our music or software, and making decisions about which products to provide, would be unacceptable to most of us.

dangerous curves

Public *Provision* versus Public *Production* Don't confuse public *provision* of a good with public *production*. The government must *provide* a pure public good in order to correct the market failure problem. But it can provide it by either producing the good itself (public production) or contracting production out to private firms (private production). Many local governments, for example, pay private firms to collect trash or run prisons. These services are privately produced, but purchased by government and *publicly provided* without charge to residents.

Common Resources

In the lower left corner of Figure 6 is another category of mixed good: nonexcludable, but rivalrous. Crowded city streets and some important natural resources fall into this category. They are often called **common resources** because they are available to everyone in common—no one can be excluded from consuming them. But their consumption lessens the benefits available for others.

Common resource A nonexcludable and rival good. Generally available free of charge, though efficiency would require a positive price.

Economists use the term *tragedy of the commons* to describe the problem caused by many of these goods. In a traditional English village, the commons was an area freely available to all families for grazing their animals. Grazing rights are a rivalrous good: If one cow eats the grass, another can't. But the commons had no method of exclusion, so it was overgrazed, causing harm to *all* families.

*The **tragedy of the commons** occurs when rivalrous but nonexcludable goods are overused, to the detriment of all.*

Tragedy of the commons The problem of overuse when a good is rivalrous but nonexcludable.

When common resources are used within a single country, that country's government can sometimes solve the market failure by passing laws or regulations to limit consumption. But when a nonexcludable resource is shared internationally, it is much harder to find a solution, and the tragedy of the commons often results.

An example is fishing in international waters, generally beyond 200 miles from national coastlines. No one owns these areas of the ocean, and no single government can tell people not to fish in them. And since no one charges for the fish removed,

dangerous curves

Categorizing Market Failures. Don't be troubled if a particular market seems to fit more than one type of market failure. The categories are not mutually exclusive. For example, a walk-by park—which we've discussed in this section as a pure public good—could also be viewed as a good with an important positive externality. After all, if the park were provided by the private market, many people who did not produce it or pay to enjoy it would get external benefits. In general, when a positive externality becomes important enough, economists begin to think of the good as a pure public good.

Similarly, when a negative externality becomes large enough—such as with greenhouse gas emissions or other types of pollution—it can be useful to shift our thinking to the common resource model. The atmosphere is a common resource, because the private market cannot establish excludability in using it or polluting it.

fishing boats use huge nets that catch just about every source of protein (most of which is ground up to be used as cattle feed). These methods are also changing the ocean's ecosystem, threatening species of ocean life further down the food chain. Although scientists have recommended specific limits on national governments since 1987, virtually every country—concerned about the welfare of its *own* fishing industry—has chosen to ignore the recommendations. Each government is a free rider in the international community, reasoning that its own fish catch is just a "drop in the bucket," and that whatever other governments do, it is better off allowing its *own* crews to continue depleting the fish.

Another example of a common resource is the earth's atmosphere. The planet is warming. To the extent that this trend is exacerbated by emissions of carbon dioxide and other greenhouse gases, the world has an interest in reducing this activity. But reducing emissions is costly. And it would benefit everyone on the planet, whether they helped pay these costs or not. Therefore, each country—and even each person—can be a free rider, with the result that nothing is done. This is the logic behind international efforts to assign quotas or tradable permits for emissions, as the European Union has done, and the United States was moving toward in 2009.

Some Important Provisos

Classifying goods into the four categories in Figure 6 is not always cut-and-dried. Consider a newspaper. It is mostly rival: The newspaper I buy at the newsstand and take home can't be bought and taken home by you. It is also largely excludable: You can't take it from the newsstand unless you pay.

But an important aspect of a newspaper is the *information inside it*. This information is largely nonrival (I can tell you what I've read without diminishing anyone else's knowledge) and largely nonexcludable (once I tell you the news, you've gotten it without having to pay the newspaper company).

Another complication is that the same good can be rival at some times and nonrival at others. Think about a highway. On a Sunday afternoon, with few cars on the road, it is nonrival: Another vehicle does not lessen anyone else's benefits. But when the highway is congested, it becomes rivalrous: Each additional driver subtracts benefits for everyone else on the road. The more congested the road, the more rivalrous it becomes.

Finally, excludability is actually more a spectrum than a binary trait. At one end are goods that are easily excludable at very low cost. For example, alarm tags and a single security guard can make most items in a large store excludable at little cost. At the other extreme, the cost is so high as to be impractical. (Imagine trying to charge pedestrians for their use of the sidewalk.) A further complication is that technological progress can change the cost of excluding nonpayers. For example, new electronic monitoring devices are making it possible to charge people who drive in congested areas during peak times.

For all these reasons, not all goods will fall cleanly into just one category in Figure 6, or remain in the same category over time. Nevertheless, the categories help us understand the role of rivalry and excludability in creating market failures and creating an efficiency role for government.

Asymmetric Information

In the previous chapter, we saw that a well-functioning competitive market provides the efficient quantity of a good. In this chapter, you've already learned some of the features behind the "well functioning" condition. For example, the good must be a private good, characterized by excludability and rivalry. There should be no positive or negative externalities in production or consumption of the good. And remember that we've always been assuming a *competitive* market (ideally, a *perfectly* competitive market).

When you learned about perfect competition in Chapter 9, one of the conditions was *easily obtained information*. In perfect competition, we assume that each party knows what it is getting and what it is giving up. We also assume that obtaining this information has no significant costs. But this is not always the case.

Suppose you buy a bottle of sugar pills for $50 because you believe they will help you shed pounds without diet or exercise. The seller is better off, but you are worse off. Moreover, this transaction likely *decreases* total benefits. You lost $50 and also suffered the humiliation of being fooled; the seller gained *less* than $50 because he has to deduct from his revenue the time and expense of making and marketing the pills. And a buyer, in order to find out what is in the pills, would have to go through the costly process of having them analyzed in a lab.

This transaction is an example of a problem known as **asymmetric information**—when one party to a transaction knows something about its value that the other party does not know. Asymmetric information can cause markets to fail in numerous ways, depending on the type of information involved

Asymmetric information A situation in which one party to a transaction has relevant information not known by the other party.

Adverse Selection

One type of asymmetric information concerns the *quality* of a good. The classic example is the used-car market. The owner of a car knows much more about its quality than anyone else. From general knowledge, or consulting Web sites, or life experience, buyers may know the quality of the *average* used car of that model and with that mileage, but not the particular one being offered to them. The buyer's willingness to pay for any particular car will thus be based on the average quality of used cars of the same category.

In this sort of market, anyone who sells a "lemon"—an unusually poor car—stands to gain, because the car can be sold for more than its value. By contrast, anyone with a great used car would lose out in this market, because buyers will believe it to be of average quality. As a result, people with great used cars are reluctant to sell them to strangers in the market. But lemons are offered in large numbers.

Of course, once the market becomes awash in lemons, the average quality drops further, reducing used-car prices even more. Eventually, the market may offer nothing but lemons. The good-quality cars disappear: Used-car prices are so low that the good cars are kept by their owners or sold only to friends.

In this example, asymmetric information about quality leads to the problem of **adverse selection.** The market acts as if it is "selecting" only the worst cars to offer for sale. Pareto improvements—in which people sell good used cars to strangers at mutually beneficial prices—remain unexploited.

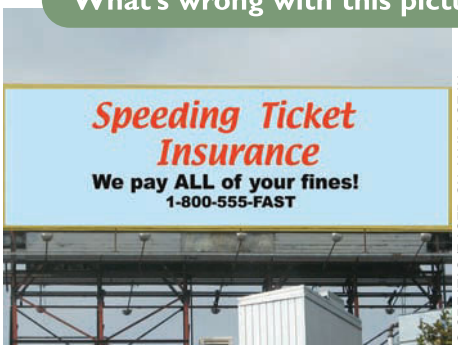
Adverse selection A situation in which asymmetric information about quality eliminates high-quality goods from a market.

Adverse selection occurs in many markets. In labor markets, employers are more likely to let go of their "lemons" and retain their highest-quality workers, so well qualified but unemployed workers may have a harder time finding jobs. In the

market for health insurance, those with the poorest health—who would get the best deal—are most likely to want to sign up for health insurance at any given price. This can drive up the cost of insurance until people of average or good health are priced out of the market.

Moral Hazard

What's wrong with this picture?



© CREATAS (BASED ON AN IMAGE IN ECONOCLASS.COM BY LORI ALDEN, 2005)

Another information problem—especially for insurance markets—arises from lack of information about someone's *future behavior.* It is called **moral hazard**—the tendency for people to change their behavior and act less responsibly when they are protected from the harmful consequences of their behavior. For example, suppose you are away from home and realize you may have forgotten to lock your front door. Without theft insurance, you would think about the full value of what you'd lose if you were robbed. You might then go to considerable trouble to return home and check the lock. But if you have theft insurance you might only consider *part* of the cost—the part that *you* would pay. The rest of the cost—covered by your insurance company—would not matter to you. You would be less likely to return home to check the lock. In this way, theft insurance leads to fewer locked doors, a greater incidence of theft, and higher insurance costs for everyone.

Moral hazard When someone is protected from paying the full costs of their harmful actions and acts irresponsibly, making the harmful consequences more likely.

Moral hazard is a problem, to some degree, for every kind of insurance in which the insured has some influence over the likelihood or amount of their future loss. Insuring people for floods or earthquakes not only makes them more likely to live in areas prone to those natural disasters but reduces their incentive to make their homes less vulnerable to damage. And though few people would choose to become ill just because they have health insurance, there is still a moral hazard problem: The more of my costs that are covered by the insurance company, the less I care whether my doctor charges excessive fees or uses inefficient and costly procedures as part of my health care.

The Principal–Agent Problem

Principal–agent problem When one party (the principal) hires another (the agent), who in turn can pursue goals that conflict with the principal's because of asymmetric information.

Finally, another potential market failure that arises from information asymmetry is the **principal–agent problem.** A principal is someone who hires someone else—an agent—to act in the principal's interest. The problem arises when the principal does not have full information about the agent's performance. This enables the agent to act in his *own* interest, at the expense of the principal.

A classic example is when you hire someone to fix your car. Unless you stand and watch (and also have expertise in car repair), you won't know whether the mechanic has done all he promised, and done it well. Some of your payment may just be a transfer of benefits from you to the mechanic, leaving you worse off. Knowing this, you may not want to take your car to a mechanic at all, and may do so only when the car stops running.

Principal–agent problems plague transactions involving individual contractors (roofers, plumbers, and so forth) and also many labor market relationships. In a corporation, managers are hired to be agents of the stockholders. But managers have their own goals that may conflict with those of the stockholders (higher-than-competitive salaries for themselves, larger-than-efficient offices, padded expense accounts, and more). Similarly, the managers act as principals when they hire hourly workers as agents. But hourly workers can find hidden ways of shirking that managers

cannot monitor, such as spending time on Facebook, or not paying attention to the task at hand when the boss isn't looking.

Market and Government Solutions

Information asymmetry—and the problems that arise from it—can reduce the total benefits in a market by preventing some transactions from taking place at all. In some cases, though, the information asymmetry can be addressed by the market itself. The problem of adverse selection in the used-car market, for example, can be partly corrected through *reputation*. A used-car dealer can hire mechanics to carefully screen the cars it buys and develop a reputation for selling only high-quality used cars. It could also signal to buyers that its cars are above-average quality by offering a warranty. The business could then charge a premium—which consumers would be willing to pay—to cover the additional costs of the mechanics and the warranty.

Or consider the problem of moral hazard. In some insurance markets, an insurance company could determine which consumers are most likely to exhibit costly behavior (for example, those who have made more theft insurance claims in the past). These people can be charged higher rates. In this way, those who behave more responsibly will not be priced out of the market.

The principal–agent problem can be partly addressed by offering agents incentives for good performance. One example is a *contingent contract*, which spells out rewards and penalties based on the agent's future behavior. (If I hire you to fix my roof, the contract requires you to fix it again if it leaks within the next six months.) Within business firms, the goals of employees and owners can be more closely aligned by long-term employment contracts, profit-sharing, and bonuses, for example.

All of these market-based methods are often used by buyers and sellers to *reduce* the inefficiency caused by information asymmetry. But they cannot eliminate it. The market's corrective efforts always entail a cost—either to obtain the additional information or to deal with its absence. This additional cost is like a tax on the product. And, like a tax, it creates its own deadweight loss.

When the market failure is significant, and when government solutions have a lower cost than market solutions, the government's direct involvement can be justified on efficiency grounds.

Regulation—discussed earlier in this chapter—is one way that government attempts to correct problems of information asymmetry. For example, left to the market, pharmaceutical companies would know more about their drugs, and the nature of the research done on them, than those who buy or prescribe the drugs. The Food and Drug Administration, by imposing standards for research, safety, and effectiveness, helps to correct this asymmetry. The Federal Trade Commission (FTC) and a variety of state agencies regulate a variety of markets, including the used-car market, to prevent misrepresentations and deceptive practices.

An Example: Market and Government Solutions in Health Insurance

One of the major policy debates of recent years concerns the proper role for government in the market for health insurance. As discussed earlier, this market is plagued by both moral hazard and adverse selection. But there are also special ethical and humanitarian concerns.

The private market can partially resolve some of the potential market failures. Consider, for example, one of the moral hazards: Once insured, patients need not be as careful about using costly medical expenses of dubious value because they are protected from paying the full costs themselves. Health insurance companies can—and to some degree, do—mitigate this problem by deciding which tests and procedures they will cover, and which they won't. In theory, if there were adequate competition among health insurance companies, their coverage restrictions would have to strike the right balance between costs and benefits or they would lose customers to competitors.

But in some localities, the health insurance market is *not* competitive: one or a few large insurers dominate the market. In any case, people may not trust a private insurance company—whose profits rise when it limits total payouts—to make life-or-death decisions on future medical procedures.

The adverse selection problem, too, can be solved, at least partially, by the private market. Recall that adverse selection arises because those who apply for insurance might know more about their health condition (and about likely future claims) than do the insurance companies. But if companies can obtain information about applicants' health at reasonable cost, they can charge more to someone with a preexisting condition, or offer coverage that excludes a prior condition, or refuse to insure the person at all if they believe they cannot do so profitably. This is analogous to someone trying to obtain information about a particular used car, to help them screen out lemons. Viewed from the narrow perspective of efficiency, it helps to solve the adverse selection problem: It prevents the applicant pool from becoming dominated by those with the most costly future health claims, so the companies can continue to offer health insurance at reasonable rates to those of average or better health.

But cars are not people. When car buyers have information, we don't object if an unlucky owner of a lemon cannot unload it. But many people *do* object when—because of information obtained by an insurance company—someone with an unlucky medical condition cannot get health insurance, or has to pay significantly more than others. In addition, some economists argue that the information costs are too high, and that insurance companies waste society's resources investigating applicants and rejecting claims that might be due to a prior medical condition.

In many countries in Europe and elsewhere, the government has chosen to solve the adverse selection problem by providing its own, universal health insurance to all citizens. The government then deals with the moral hazard problem by regulating the fees that doctors and hospitals can charge, and rationing health care among competing demands. In the United States, government health insurance has been more controversial and has been provided only to veterans, the elderly, and the very poor. The rest of the population is either insured by private insurance companies (through their employer or on a policy purchased individually) or has no health insurance coverage at all.

In 2009, the Obama administration began floating proposals to increase government's role in the health insurance market. The twin goals were to increase the percentage of insured Americans, and to involve the government—directly or indirectly—in controlling costs.

Efficiency and Government in Perspective

In this chapter, you've seen that an economy with *well-functioning, perfectly competitive markets* tends to be economically efficient. But notice the italicized

words. As you've seen in this chapter, many types of government involvement are needed to ensure that markets function well and to deal with market failures. The government helps markets to function by providing a legal and regulatory infrastructure. And the government frequently intervenes in markets directly, to correct market failures.

These cases of government involvement are not without controversy. In fact, most of the controversies that pit Democrats against Republicans in the United States (or Conservatives against Labourites in Britain, or Social Democrats against Christian Democrats in Germany) relate to when, and to what extent, the government should be involved in the economy. Debates about public education, Social Security, international trade, health care, and immigration all center on questions of the proper role for government.

Those who tend to be skeptical of government solutions point out several potential problems.

Government Failure

Government itself can be plagued by the same types of problems that cause market failures in the private economy. One example is the principal–agent problem. Government officials are the agents of the general public and are supposed to serve the public interest. But these officials may have their own incentives, such as expanding the size of their departmental budgets or maximizing contributions to their campaign funds. Although there are checks and balances to monitor and limit this type of behavior, they are not always effective. Even in the most lawful democracies, lobbying by corporations, unions, and other interest-groups can lead politicians and government officials astray. In recent years, the pharmaceutical industry, the financial sector, teachers unions, and many others have each spent large sums—hundreds of millions of dollars annually—to influence government policy.

Thus, even when a market failure could, *in theory,* be solved by government, we cannot be sure that government will do so effectively. And once government has the power to intervene, it could even make things worse. Those most skeptical of government intervention stress examples in which, arguably, government intervention has created even more inefficiency. Those who view intervention more favorably stress examples where, arguably, markets are functioning poorly because the government has *not* intervened.

Deadweight Loss from Taxes

In order for government to have the funds it needs to support markets and do other things, it must raise revenue through taxes. As you learned in Chapter 14, these taxes—when they are imposed on otherwise-efficient markets—introduce deadweight losses of their own. This cost of government activity should be considered along with the benefits in evaluating any government action.

Equity

Fostering efficiency—the focus of this chapter—is just one of the government's roles in the economy. We also want our government to be concerned with equity, fairness, justice, and more. These are not issues about which people easily agree.

Almost every government policy designed to improve efficiency also has consequences for equity. Taxing gasoline to correct an externality would hit the poor harder than the rich, while taxing airline travel would do the opposite. Eliminating price floors for agricultural goods might move the economy toward efficiency, but it would cause harm to many farmers and their families. Almost every change in the tax code designed to improve efficiency will raise a firestorm of protest because of the way it might affect equity. This limits the government's ability to focus on efficiency when dealing with market failures.

The controversies we've discussed can be so heated and so varied that it is easy to forget how much agreement there is about the role of government. Anyone studying the role of government in the economies of the United States, Canada, Mexico, France, Germany, Britain, Japan, and the vast majority of other developed economies, is struck by one glaring fact: Most economic activity is carried out among private individuals. In all of these countries, there is widespread agreement that although government intervention is often necessary, the most powerful forces that exploit Pareto improvements and drive the economy toward efficiency are the actions of individual producers and consumers. But there are areas of disagreement as well. In the next Using the Theory section, we discuss market failure in financial markets, and the ongoing controversy about what government should do about it.

Using the Theory

MORAL HAZARD AND THE FINANCIAL CRISIS OF 2008

NAJLAH FEANNY/CORBIS

Private financial institutions play an important role in the economy by channeling funds from savers to borrowers. Some financial institutions (such as commercial banks and savings and loans) take in customer deposits from savers. Others (including investment banks and hedge funds) start with their owners' funds, and then borrow more themselves by issuing bonds or taking out bank loans. In addition, there are insurance companies, pension funds, money market funds, and more, each gathering funds from savers in a different way.

All of these institutions create substantial benefits for both savers and borrowers. Financial institutions develop expertise in lending, and can diversify their loans. As a result, savers enjoy less risk and/or higher returns than they could hope to achieve by lending out funds on their own. And credit-worthy borrowers have easier access to loans than if they had to contract individually with savers. The financial institution (and its owners) comes out ahead too, earning a higher rate of return on its investments than it pays out to savers.

In sum, financial institutions help the economy achieve Pareto improvements, and contribute to economic efficiency . . . *when they work well.*

In 2008 and 2009, however, it became apparent that banks and other financial institutions around the world were *not* working so well. For several years they had been making huge investments in securities backed by housing and other assets. In 2008, they revealed that they had lost (in total) several trillion dollars on these investments. *Creditors* (those who had loaned funds to these institutions) suddenly realized the institutions might go bankrupt and they might not be paid back. They stopped lending to the financial institutions, which, in turn, stopped making new loans throughout the economy. A full-fledged financial crisis developed. And at the root of this crisis was *moral hazard,* which arose in several different ways.

Moral Hazard from Deposit Insurance

In most developed countries, governments (or agencies backed by the government) typically guarantee most of the deposits held in banks and other depository institutions. For example, in the United States, the Federal Deposit Insurance Corporation (FDIC) has traditionally guaranteed deposits at most commercial banks up to $100,000, and in early 2009, raised the limit to $250,000 until 2013. Deposit insurance helps to make the banking system more stable: In the face of rumors or fears about the safety of their deposits, people have no need to rush to the bank and withdraw their insured funds. Federal deposit insurance has been very successful in preventing "banking runs" in the United States, a potential problem you will learn more about in macroeconomics.

But deposit insurance also creates a problem: When your deposits are insured, you needn't be concerned if your bank makes irresponsible loans or even loses all of your deposits in an online poker game. If the bank, for any reason, can't return your deposits when you want them, the government will do so. As a result, depositors, once insured, no longer watch over their banks, leaving them free to engage in costly, irresponsible behavior. This is moral hazard.

Governments that insure deposits have long recognized this moral hazard, which is why they have imposed strict *regulations* on depository institutions. Government provides the discipline that depositors would otherwise provide. Government regulators tell banks what types of risks they are allowed to take, and also require bank owners to have minimum amounts of their own funds at risk. For many decades, this system—in the U.S. and many other countries—has worked well. In the decade leading up to 2008, however, many of the largest banks found ways around the intent of the regulations, enabling them to take increasingly-risky gambles with depositor funds.

Moral Hazard from "Too Big to Fail"

Another type of moral hazard arises more informally in the largest financial institutions (or their subsidiaries) that do *not* take deposits, whose funds are *not* formally guaranteed by the government, and which are thus *not* regulated as strictly. Because they are so heavily intertwined with the broader economy, and have financial relationships with so many businesses and households around the world, governments have been loath to let them fail or allow their creditors to suffer serious losses. If the creditors of one such institution suffered losses, people might fear similar losses at the others, causing funding—and lending—to suddenly dry up for all of them. The

expression "too big to fail" sums up the attitude governments have taken toward these large, complex institutions.

But once everyone believes that a financial institution is too big to fail, its creditors are affected by moral hazard: They don't have to monitor the institution as carefully. If the firm gets into trouble, the government (using taxpayer funds) will have no choice but to come to the rescue. While the government may not rescue the firm's owners (stockholders), it would certainly rescue the creditors. As a result, the careful monitoring and discipline that nervous creditors would ordinarily provide is missing. This leaves the stockholders as the main watchdogs over these firms. But stockholder discipline is weakened by another form of moral hazard, which we discuss next.

Moral Hazard from Limited Liability

In the United States, Europe, and many other countries, stockholders of corporations enjoy a legal protection known as "limited liability." As a shareholder in the corporation, you cannot lose more than your entire investment. If employees of the firm take excessive risks or even engage in criminal behavior, you—as a stockholder—might lose your entire investment in the firm's shares, but the rest of your personal wealth is safe from lawsuits or prosecution.

Limited liability has made it much easier for corporations to raise funds by issuing new shares, and helped corporations grow so they can exploit economies of scale. Both investors and the economy in general have benefited. But it does have a downside. A financial institution's shareholders know that if their firm's high risk gambles turn out badly—losing even more than the shareholders have invested—the shareholders will not have to cover all of the loss. In effect, when financial firms gamble with other peoples' money (borrowed or deposited), their shareholders get all the winnings, but their potential losses are limited to the value of the shares they own.

Moral Hazard from Employment Practices

Although limited liability does not formally extend to individual employees, they typically enjoy a similar protection. Suppose you work for a financial institution, and you take big risks with the firm's funds and lose billions of dollars. In most cases, the only cost to you is losing your job. As long as you haven't engaged in criminal behavior or clearly violated your employment contract, your personal wealth is safe—including any wealth you've accumulated in bonuses from earlier gambles that paid off.

How Moral Hazard Can Skew Investment Decisions

To see how these moral hazard problems can distort investment decisions, let's consider a very simple example. Imagine that you have decided to gamble $10,000 of your own money on a simple coin toss: Heads you win, tails you lose. But you can choose between two different gambles. In Gamble A, if the coin comes up heads, you double your money. If it comes up tails, you lose only *half* of your money. So Gamble A looks like this:

Gamble A:	Heads:	$10,000 gain
	Tails:	$ 5,000 loss

Now consider Gamble B: If the coin comes up heads, you will gain $12,000. But if it comes up tails, you lose your entire bet. So Gamble B looks like this:

Gamble B: Heads: $12,000 gain
Tails: $10,000 loss.

Which gamble would you choose? If you are like most people, you'd choose Gamble A. While B pays off a bit more if you win, the potential losses are much greater for B than for A.

Now let's change the situation slightly. Instead of making the choice yourself, suppose you hire a representative to make the choice for you—someone with special training and ability to assess gambles and choose the best one. This is what financial institutions do. We'll call this representative a "trader," the informal title used by financial institutions for such employees. To make sure your trader has something at stake, you make the following arrangement: If you win, your trader gets a bonus equal to one-tenth of your winnings. But if you lose, the trader will be fired.

From the *trader's* point of view, looking at the personal consequences only, the choice now looks like this:

Gamble A: Heads: $1,000 bonus
Tails: Lose job

Gamble B: Heads: $1,200 bonus
Tails: Lose job

For the trader, Gamble B is the better choice. By protecting the trader from the full consequences of risky behavior, the trader has an incentive to take excessive risks. This is moral hazard.

But wait . . . wouldn't you *realize* this, and refuse to hire the trader to make the choice for you? Perhaps—if the gambles are as easy to understand as those shown here. But even if you do understand the gambles, it might still be in your interest to hire the trader . . . *if* we make one more change.

Suppose that, before you gamble any money, you form a corporation. Your corporation issues one share of stock, which you purchase for $10,000 and become the sole owner. Moreover, you decide to *leverage* 20 times (see the appendix to Chapter 4 on leverage). That is, your corporation will borrow $190,000 from others. When combined with your original $10,000, your corporation has $200,000 to gamble with. Now that you are leveraged, the gains and losses are 20 times what they were when you gambled with only your own funds. However, due to limited liability, the maximum loss for your corporation (and for *you* as its owner) is what you've paid for your share of stock: $10,000. Let's look at the payoffs now:[5]

Gamble A: Heads: $200,000 gain
Tails: ~~$100,000 loss~~ **$10,000 loss**

Gamble B: Heads: $240,000 gain
Tails: ~~$200,000 loss~~ **$10,000 loss**

Notice that, in the payoffs listed above, we've first listed the *total* gains and losses from each leveraged gamble, and then replaced the total loss with the actual loss borne by your corporation. For example, Gamble A loses $200,000 if the coin comes up tails, but your corporation can only lose $10,000 due to limited liability.

[5] To keep the example simple, we're ignoring the interest payments you'd make on the borrowed funds, because they'd be small relative to the gains and losses and would not change our conclusions.

Which gamble do you, as owner of the corporation, prefer now? Clearly B. And so does your trader: As before, the potential bonus under B is greater than the potential bonus under A.

This raises a question: If gamble B is chosen, and the coin comes up tails, and your corporation bears only $10,000 of the total $200,000 loss, then who bears the rest? The answer is: the outsiders—the people who lent you $190,000.

But even *they* may not suffer. Suppose your corporation is somehow deemed "too big to fail" by the government. Or suppose it is a bank, so the $190,000 in outside funds were customer deposits guaranteed by the FDIC. Then those who provided the additional funds will not lose either. Instead, the loss will ultimately fall on the taxpayers. Although our example is simple, and involves small sums relative to the trillions of dollars that were lost by large financial institutions around the world, it illustrates how a chain of moral hazard can lead to excessive risk taking, with taxpayers eventually bearing the cost.

Moral Hazard and the Principal-Agent Problem

In our simple examples, moral hazard allows traders, owners, and creditors to consciously exploit taxpayers. But during the financial crisis of 2008, stockholders and creditors suffered huge losses as well. Moral hazard—in addition to creating problems of its own—also created a principal-agent problem. Financial trading had become so complex that even the top management of the firm had difficulty understanding the risks being taken by its traders. Stockholders were even further removed, and did not have information that would have enabled them to assess the risks.

For creditors of financial firms, the situation was a bit different. To some extent, they relied on three large government-sanctioned *credit rating agencies* (Standard & Poor's, Moody's, and Fitch), which were supposed to monitor the financial institutions and warn creditors about any excessive risk taking. But the rating agencies failed to do so. While there is some debate about the reasons, it is clear that the ratings agencies themselves were plagued with their own moral hazard and principal-agent problems.

The Aftermath of the Financial Crisis

In the decade leading up to 2008, many of the real-world risks taken by financial institutions were highly leveraged bets, mostly bets that housing prices would continue to rise. For example, Lehman Brothers, Bear Stearns, Merrill Lynch, and American International Group (AIG) were leveraged more than 30 times. When the housing boom collapsed (see Chapter 4) and the bets were lost, the largest financial institutions lost hundreds of billions of dollars each.

When the huge losses came to light, Lehman Brothers—which had always been viewed as too big to fail—was actually allowed to fail. Lehman's creditors—to their surprise—bore much of the cost. Two other large financial institutions (Bear Stearns and Merrill Lynch) were acquired by other firms in distress sales, with the government's help. Their creditors were largely rescued. And AIG was simply taken over by the government, which continued to honor virtually all of its obligations. Such events were not limited to the United States. In England, Ireland, France, Germany, Switzerland, Iceland, and many other countries, large financial institutions went bankrupt or were rescued by their governments.

In the aftermath of the financial crisis, economists agreed that the moral hazard problem had to be addressed. In 2009, the United States and several European

governments took the lead in proposing new, stricter regulations for financial institutions. These included requiring that bonuses be withheld for several years, to be paid only after financial gambles had proved profitable in the *long-run*. New legal procedures would enable large financial institutions to fail without disrupting the broader economy. There was some discussion of limiting how large and complex any one financial institution could grow. And almost everyone agreed that financial firms would have to report more information to regulators, stockholders, and creditors so there could be better monitoring of the risks being taken with outsiders' funds.

But a few of the proposals were controversial. Some economists warned that the pendulum might be swinging too far in the direction of *over*-regulation, resulting in too little financial innovation and risk-taking. And some pointed to the dangers of *government* failure that we discussed earlier in this chapter. But in 2009, the public debate was dominated by those (including many economists) who believed that financial markets had failed to do their job, and that new regulations were needed to make another, similar failure less likely in the future.

SUMMARY

Government contributes to economic efficiency in two general ways. First, it provides the legal and regulatory system that enables the market system to function. Second, it often steps in to correct specific *market failures*—situations in which a specific market, left to itself, is inefficient.

One solution for the market failure of monopoly or monopoly power (used when the market is not a natural monopoly) is antitrust action to create more competition. For a *natural monopoly,* a common solution is a regulated price that includes a fair rate of return for owners.

Externalities are unpriced by-products of economic transactions that affect outsiders. With an externality, a market may not be efficient. The *Coase theorem* tells us that under certain conditions, the market can solve the externality problem with private action. When private action is not possible, the market will not produce the efficient quantity on its own, creating a deadweight loss. Negative externalities can be corrected through regulations, but in many cases, market-based approaches (taxes or tradable permits) are more efficient. For positive externalities, government typically use subsidies to raise output to the efficient level.

Pure public goods—those that are nonrival and nonexcludable—are a market failure because private firms generally will not provide them, and the efficient price in any case is zero. These goods are typically provided by government at no charge.

Asymmetric information can create market failures when it leads to *adverse selection, moral hazard,* or the *principal–agent problem*. The market can address some of these problems itself, but government often uses regulation to help solve them.

Government solutions to market failures are often imperfect and can introduce their own inefficiencies. And virtually all government policies have implications for equity, which is another important issue for government. For this and other reasons, government solutions to market failures are often controversial.

PROBLEM SET

Answers to even-numbered Questions and Problems can be found on the text Web site at www.cengage.com/economics/hall.

1. Review the section of the chapter titled, "The Private Solution" for negative externalities. Suppose that the truck driver gains $12 by using the shortcut, and the harm to the resident is $7.
   a. Is it efficient for the truck driver to stay on the highway and not use the shortcut? Briefly, why or why not?
   b. If the truck driver has the legal right to use the shortcut, would you expect the resident to offer him a sufficient side payment to get him to stay on the highway? Why or why not?
   c. If the resident has the legal right to block the truck driver from using the shortcut, would you expect the truck driver to stay on the highway? Why or why not?

2. Figure 3 shows the market for gasoline with a negative externality from pollution. Using the information in figure 3 and 4, calculate the dollar value of the deadweight loss per period before any government intervention.
3. Figure 5 shows the market for college degrees with a positive externality. Using the information in both panels, calculate the dollar value of the deadweight loss per period before any government intervention.
4. Last year, Pat and Chris occupied separate apartments. Each consumed 400 gallons of hot water monthly. This year, they are sharing an apartment. To their surprise, they find that they are using a total of 1,000 gallons per month between them. Why? What concept discussed in this chapter is illustrated by this example?
5. Some have argued that the music industry is by nature inefficient because once a piece of music is produced, the firm that owns it has a monopoly and charges the monopoly price. Yet, the marginal cost of making the music available to one more member of the public (via the Internet) is zero. Draw a diagram, similar to Figure 2, to represent this situation. Identify on your diagram:
   a. The efficient level of production
   b. The level of production a government-regulated music industry would earn if it were permitted to charge just enough for a "fair rate of return"
   c. The level of production provided by the (currently unregulated) industry
6. In Figure 4, a negative externality was corrected with a $1.00 per gallon tax on gasoline producers. Draw a diagram to show that the total price paid by consumers, the total price received by firms, and the equilibrium quantity would have been exactly the same if the same tax had been imposed on gasoline consumers instead of producers.
7. In Figure 5(b), a positive externality was corrected with a $30,000 subsidy paid to students. Draw a diagram to show that the total price paid by students, the total price received by colleges, and the equilibrium quantity of degrees would have been exactly the same if the $30,000 subsidy per student had been given to colleges instead.
8. Each of the following is an example of (or would lead to) a particular type of market failure arising from information asymmetry. Identify the type of market failure and justify your answer briefly.
   a. A woman in the "dating market" complains, "All the good ones are taken."
   b. A college announces a new policy: Any senior with a GPA less than 2.0 will, upon graduation, have all grades of C− or lower retroactively raised to a grade of C.
   c. A restaurant in New York hires workers to pass out fliers to pedestrians all over the city, and pays them based on how many fliers they get into people's hands.

More Challenging

9. The following table shows the quantities of car alarms demanded and supplied per year in a town:

Price	Quantity Demanded	Quantity Supplied
$ 75	800	0
$100	750	150
$125	700	300
$150	650	450
$175	600	600
$200	550	750
$225	500	900
$250	450	1,050

   Without drawing a graph, determine the efficient quantity in this market under each of the following assumptions:
   a. Each car alarm sold creates a negative externality (noise pollution) that causes $100 in harm to the public.
   b. Each car alarm creates a *positive* externality (reduced law enforcement costs) that provides $100 in benefits to the public.
10. Suppose Douglas and Ziffel have properties that adjoin the farm of Mr. Haney. The current zoning law permits Haney to use the farm for any purpose. Haney has decided to raise pigs (the best use of the land). A pig farm will earn $50,000 per year, forever.
    a. Assume the interest rate is 10 percent per year. What is Haney's pig farm worth? (Hint: Use a special formula from Chapter 13.)
    b. Suppose the next best use of Haney's property is residential, where it could earn $20,000 per year. What is the minimum one-time payment Haney would accept to agree to restrict his land for residential use forever?
    c. Suppose Douglas is willing to pay $200,000 for an end to pig farming on Haney's land, while Ziffel is willing to pay no more than $150,000. (For some reason, Ziffel does not mind pig farming as much as Douglas does.) If Douglas pays Haney $200,000 and Ziffel pays Haney $150,000, and Haney converts his land to residential use, is this a Pareto improvement? Who benefits, who loses, and by how much?
    d. Suppose instead that Douglas pays $150,000 and Ziffel pays $150,000. Is this move a Pareto improvement? Who benefits, who loses, and by how much?

CHAPTER 16

Comparative Advantage and the Gains from International Trade

Consumers love bargains. And the rest of the world offers U.S. consumers bargains galore: cars from Japan, computer memory chips from Korea, shoes and clothing from China, tomatoes from Mexico, lumber from Canada, and sugar from the Caribbean. But Americans' purchases of foreign-made goods have always been a controversial subject. Should we let these bargain goods into the country? Consumers certainly benefit when we do so. But don't cheap foreign goods threaten the jobs of American workers and the profits of American producers? How do we balance the interests of specific workers and producers on the one hand with the interests of consumers in general? These questions are important not just in the United States, but in every country of the world.

Over the post–World War II period, there has been a worldwide movement toward a policy of *free trade*—the unhindered movement of goods and services across national boundaries. An example of this movement was the creation—in 1995—of a new international body: the World Trade Organization (WTO). The WTO's goal is to help resolve trade disputes among its members and to reduce obstacles to free trade around the world.

And to some extent it has succeeded: Import taxes, import limitations, and all kinds of crafty regulations designed to keep out imports are gradually falling away. In recent years, almost one-third of the world's production has been exported to other countries. One hundred fifty-three countries have joined the WTO, and 30 others are eager to join.

But even though many barriers have come down, others remain—and new ones have come up. In 2009, the United States was still refusing to eliminate its long-standing quota on sugar imports or to allow Mexican trucks to carry products into the United States. The European Union continued to restrict imports of U.S. beef, and took steps to keep out imports of Chinese screws and bolts. China banned all of its government agencies from purchasing imported goods unless domestic versions were unavailable. U.S. Congress members complained about China's ban, even though just a few months earlier, they had inserted "Buy American" orders in the new economic stimulus program.

Looking at the contradictory mix of trade policies that exist in the world, we are left to wonder: Is free international trade a good thing that makes us better off, or is it bad for us and something that should be kept in check? In this chapter, you'll learn to apply the tools of economics to issues surrounding international trade. Most important, you'll see how we can extend economic analysis to a global context, in which markets extend across international borders.

The Logic of Free Trade

Many of us like the idea of being self-reliant. A very few even prefer to live by themselves in a remote region of Alaska or the backcountry of Montana. But consider the defects of self-sufficiency: If you lived all by yourself, you would be poor. You could not *export* or sell to others any part of your own production, nor could you *import* or buy from others anything they have produced. You would be limited to consuming the goods and services that you produced. Undoubtedly, the food, clothing, and housing you would manage to produce by yourself would be small in quantity and poor in quality—nothing like the items you currently enjoy. And there would be many things you could not get at all—electricity, television, cars, airplane trips, or the antibiotics that could save your life.

The defects of self-sufficiency explain why most people do not choose it. Rather, people prefer to specialize and trade with each other. In Chapter 2, you learned that specialization and exchange enable us to enjoy greater production and higher living standards than would otherwise be possible.

This principle applies not just to individuals, but also to *groups* of individuals, such as those living within the boundaries that define cities, counties, states, or nations. That is, just as we all benefit when *individuals* specialize and exchange with each other, so, too, we can benefit when *groups* of individuals specialize in producing different goods and services, and exchange them with other *groups*.

Imagine what would happen if the residents of your state switched from a policy of open trading with other states to one of self-sufficiency, refusing to import anything from "foreign states" or to export anything to them. Such an arrangement would be preferable to individual self-sufficiency; at least there would be specialization and trade *within* the state. But the elimination of trading between states would surely result in many sacrifices. Lacking the necessary inputs for their production, for instance, your state might have to do without bananas, cotton, or tires. And the goods that *were* made in your state would likely be produced inefficiently. For example, while residents of Vermont *could* drill for oil, and Texans *could* produce maple syrup, they could do so only at great cost of resources.

Thus, it would make no sense to insist on the economic self-sufficiency of each of the 50 states. And the founders of the United States knew this. They placed prohibitions against tariffs, quotas, and other barriers to interstate commerce right in the U.S. Constitution. The people of Vermont and Texas are vastly better off under free trade among the states than they would be if each state were self-sufficient.

What is true for states is also true for entire nations. The members of the WTO have carried the argument to its ultimate conclusion: National specialization and exchange can expand world living standards through free *international* trade. Such trade involves the movement of goods and services across national boundaries. Goods and services produced domestically, but sold abroad, are called **exports**; those produced abroad, but consumed domestically, are called **imports.** The long-term goal of the WTO is to remove all barriers to exports and imports in order to encourage among nations the specialization and trade that have been so successful within nations.

Exports Goods and services produced domestically, but sold abroad.

Imports Goods and services produced abroad, but consumed domestically.

International Comparative Advantage

In Chapter 2, you were introduced to the notions of absolute and comparative advantage. Let's focus on these two concepts as they apply to trade between nations.

A country has an *absolute advantage* in a good when it can produce it using *fewer resources* than another country. Economists who first considered the benefits of international trade focused on absolute advantage. As the early economists saw it, the citizens of every nation could improve their economic welfare by specializing in the production of goods in which they had an absolute advantage and exporting them to other countries. In turn, they would import goods from countries that had an absolute advantage in those goods.

In 1817, however, the British economist David Ricardo disagreed. Absolute advantage, he argued, was not a necessary ingredient for mutually beneficial international trade. The key was *comparative advantage:*

> *A nation has a comparative advantage in producing a good if it can produce it at a* lower opportunity cost *than some other country.*

As you learned in Chapter 2, there is a key difference between absolute advantage and comparative advantage. While absolute advantage in a good is defined by the resources used to produce it, comparative advantage is based on the *opportunity cost* of producing it. The opportunity cost of producing something is the *other goods* that these resources *could* have produced instead.

Ricardo argued that a potential trading partner could be absolutely inferior in the production of every single good—requiring more resources per unit of each good than any other country—and still have a comparative advantage in some good. The comparative advantage would arise because the country was *less* inferior at producing some goods than others. Likewise, a country that had an absolute advantage in producing everything could—contrary to common opinion—still benefit from trade. It would have a comparative advantage only in some, but not all, goods.

To illustrate Ricardo's insight, let's go back to our example from Chapter 2, involving trade between the U.S. and China. We'll first review what you learned there, and then carry our analysis and discussion further.

Determining a Nation's Comparative Advantage

In our example, the U.S. and China each produce two goods—soybeans and T-shirts—using just one resource: labor. Table 1 shows the number of labor hours required to produce one unit of each good, in each country. We assume that these labor requirements remain constant, no matter how much is produced.

Notice that the United States has an *absolute* advantage in producing both goods: It takes fewer hours to produce either a bushel of soybeans or a T-shirt in the United States than in China. But China has a *comparative advantage* in T-shirts.

TABLE 1

Labor Requirements for Soybeans and T-Shirts

	Labor Required for:	
	1 Bushel of Soybeans	**1 T-Shirt**
United States	½ hour	¼ hour
China	5 hours	1 hour

That is, the opportunity cost of producing T-shirts is lower in China than in the United States. How do we know? For China to produce one more T-shirt, they must shift one hour of labor out of soybean production. Because each bushel of soybeans requires five hours in China, taking away one hour causes soybean production to fall by one-fifth of a bushel. So in China, the opportunity cost of one T-shirt is one-fifth of a bushel of soybeans.

In the United States, producing another T-shirt means shifting one-fourth hour from soybeans to T-shirts. This shift decreases soybean production by one-half of a bushel. Thus, in the United States, the opportunity cost of one T-shirt is one-half bushel of soybeans. Because China has the lower opportunity cost (one-fifth of a bushel, rather than one-half of a bushel), China has the comparative advantage in T-shirts.

Similar reasoning (which you should do on your own) will show that the United States has a lower opportunity cost, and therefore a comparative advantage, in producing in soybeans.

So far, this should all be familiar: It's the same analysis, done in the same way, as in Chapter 2. But now, as we continue with the example, we'll delve deeper. For each country, we'll compare production and consumption of each good before trade and after trade. And we'll see precisely how each country can gain by exporting its comparative advantage good.

How Specialization Increases World Production

Figure 1 shows production possibilities frontiers for the United States and China. In the left panel, we assume that the United States has 100 million hours of labor per year, which it must allocate between soybeans (on the horizontal axis) and T-shirts (on the vertical axis). To obtain the PPF for the United States, we first suppose that all 100 million hours of labor were allocated to T-shirts. The United States could then produce 400 million of them per year (because each one requires $\frac{1}{4}$ hour of labor). Accordingly, the upper-most point on the PPF represents 400 million T-shirts and zero bushels of soybeans.

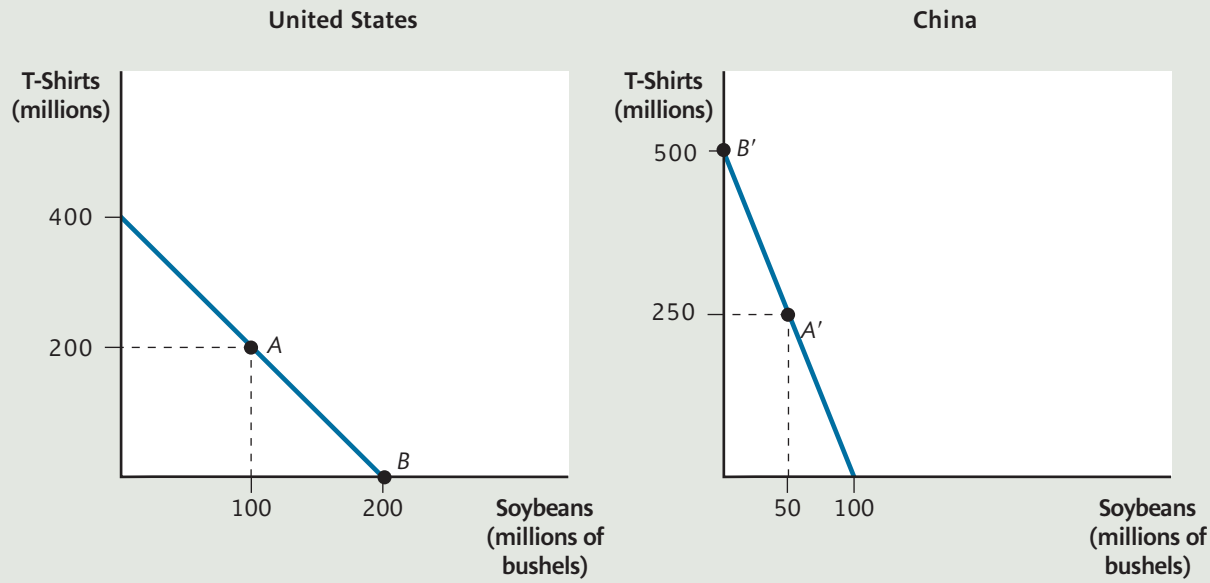
To get the rest of the points, remember that the opportunity cost of one more bushel of soybeans is 2 T-shirts. Therefore, each time we move rightward by one unit (one more bushel), we must move downward by 2 units (2 fewer T-shirts). Accordingly, the PPF for the United States will be a straight line, with a slope of -2. The PPF ends where the United States would be allocating all of its 100 million hours to soybeans, producing 200 million bushels.

Notice that this PPF is a straight line, unlike the curved PPFs in Chapter 2. A linear PPF follows from our assumption that hours per unit—and therefore opportunity costs—remain constant no matter how much of either good is produced. Essentially, to keep things simple, we are assuming *constant opportunity costs,* rather than increasing opportunity cost as in the PPFs drawn in Chapter 2. (We'll discuss the implications of this in a few pages.)

The right panel shows China's PPF, under the assumption that China has 500 million hours of labor per year. On your own, be sure you can see how the two endpoints of China's PPF are determined. Also, be sure you understand why the slope of China's PPF will be -5. (Hint: What is the opportunity cost of another bushel of soybeans in China?)

Before international trade occurs, we assume (arbitrarily) that both countries are operating in the middle of their respective PPFs. The United States is at point *A,*

FIGURE 1 How Trade Changes Production



	United States	China	World (United States + China)
Pre-trade Production			
Soybeans (million bushels)	100	50	150
T-shirts (millions)	200	250	450
Post-trade Production			
Soybeans (million bushels)	200	0	200
T-shirts (millions)	0	500	500

Before international trade opens up in soybeans and T-shirts, the United States is assumed to produce in the middle of its linear PPF at point A, *representing 100 million bushels of soybeans and 200 million T-shirts. Similarly, China is assumed to produce at point* A′, *representing 50 million bushels of soybeans and 250 million T-shirts.*

After trade, each country will completely specialize in its comparative advantage good. The United States will shift all resources into soybeans (200 million bushels) at point B, *and China will put all resources into T-shirts (500 million) at point* B′. *The result is greater world production of both goods. World soybean production rises from 150 million to 200 million bushels, and world T-shirt production rises from 450 million to 500 million.*

producing 100 million bushels of soybeans and 200 million T-shirts each year. This combination of goods is also U.S. *consumption* per year: Without trade, you can only consume what you produce. In the right panel, China is at point *A′*, producing and consuming 50 million bushels of soybeans and 250 million T-shirts.

Now look at the table that accompanies the figure. The first two rows tell us the production of each good in each country before trade opens up, and also world production of each good. As you can see, with the United States producing at point *A* along its PPF and China producing at point *A′*, the world (the United States and China combined) produces 150 million bushels of soybeans and 450 million T-shirts per year.

Let's now see what happens to world *production* when trade opens up. We'll have each country devote *all* of its resources to the good in which it has a comparative advantage. The United States, with a comparative advantage in soybeans, moves to point *B* on its PPF, producing 200 million bushels of soybeans and zero T-shirts. China moves to point *B*' on its PPF, producing 500 million T-shirts and zero soybeans. The new production levels for each country are entered in the last two rows of the table in Figure 1.

Finally, look at the last column of numbers in the table. For both goods, world production has increased. Soybean output is up from 150 million to 200 million bushels, and T-shirt production is up from 450 million to 500 million. This increase in world production has been accomplished without adding any resources to either country. The world's resources are simply being used more efficiently.

Although our example has just two countries and two goods, it illustrates a broader conclusion:

When countries specialize according to their comparative advantage, the world's resources are used more efficiently, enabling greater production of every good.

How Each Nation Gains from International Trade

Now let's show that both countries can gain from trade. As you've seen, when the two countries specialize in their comparative advantage good, they produce more of that good but none of the other. For example, China produces more T-shirts but no soybeans. However, by trading some of its comparative advantage good for the other good, each country can *consume* more of both goods.

Table 2 illustrates this conclusion, under the assumption that the U.S. will trade 80 million bushels of soybeans for 240 million T-shirts from China. This is an arbitrary assumption, and we could change it and still come to the same conclusion. But let's see how international trade can benefit both countries in this case.

Because there is so much going on in this table, we'll go through it one step at a time, starting with the column of numbers for soybeans in the United States. The first entry U.S. shows how specialization changes U.S. *production* of soybeans, based on the movement along the PPF in Figure 1. When the U.S. moves from point *A* to

TABLE 2

The Gains from Specialization and Trade

	United States		China	
	Soybeans (million bushels)	T-Shirts (millions)	Soybeans (million bushels)	T-Shirts (millions)
Change in Production	+100	−200	−50	+250
Exports (−) or Imports (+)	−80	+240	+80	−240
Net Gain in Consumption	+20	+40	+30	+10

point *B* in Figure 1, soybean production increases from 100 million to 200 million bushels. This is an increase of 100 million, hence the entry +100 in the first row. If there were no international trade, this would also be the change in U.S. consumption of soybeans.

But because of international trade, the U.S. will *not* consume all the soybeans it produces. Instead, it will export some of them. Our assumption is that the U.S. exports 80 million bushels. These bushels have to be deducted from its production in order to determine how many bushels are left for the U.S. to consume. Hence, the entry −80 for U.S. exports of soybeans in the second row.

Finally, the third row subtracts exports from the change in production to get the change in *consumption* of soybeans. Since production increased by 100 million bushels, and 80 million of them were exported, U.S. consumption of soybeans rises by 100 − 80 = 20 million bushels.

Next, we move to the next column, for U.S. T-shirts. When the U.S. moves from point *A* to point *B* in Figure 1, its T-shirt production decreases from 200 million to zero, represented by the entry −200. Without trade, this would mean that consumption of T-shirts decreases by 200 as well. However, remember our assumption: In exchange for its soybean exports, the U.S. gets 240 million T-shirts from China. These imports enter with a plus sign (+240). Combining the change in production (−200) with the T-shirts obtained as imports (+240), we are left with the change in U.S. *consumption* of T-shirts: +40. So U.S. consumption of T-shirts rises by 40 million.

Notice that the United States—by specializing in and exporting its comparative advantage good (soybeans)—ends up with more of *both* goods. One might think that if the U.S. ends up with more of both goods, then the Chinese must end up with less. But don't forget: world production of both goods has increased too. And in our example, China will end up with more of both goods as well.

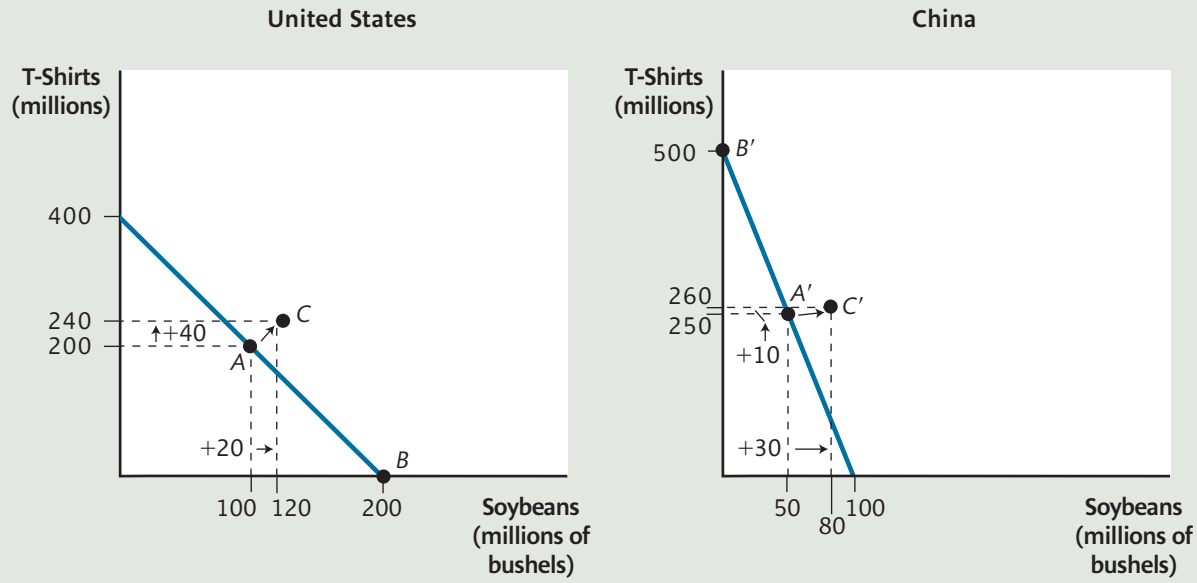
The first row shows the consequences of China's move from point *A′* to point *B′* along its PPF in Figure 1. China's soybean production decreases from 50 million to zero (−50), while T-shirt production rises from 250 million to 500 million (+250). In the second row, we remember our assumption about trade: The U.S. exports 80 million bushels of soybeans to China. That means China must be *importing* 80 million bushels of soybeans (+80). Similarly, we assume the U.S. imports 240 million T-shirts from China, so that must be how many T-shirts China *exports* (−240). Summing up the numbers for China, we see that China's soybean consumption rises by 30 million (+30), while its T-shirt consumption rises by 10 million (+10).

Illustrating Gains from Trade Graphically

Figure 2 illustrates the change in consumption on a PPF diagram similar to Figure 1. Once again, we start with the PPFs for the United States and China. But now we compare consumption in each country before trade with consumption *after* trade. The United States began with 100 million bushels of soybeans and 200 million T-shirts (point *A*). After specialization and trade, it consumes 120 million bushels of soybeans and 240 million T-shirts (point *C*). Similarly, China began by consuming 50 million bushels of soybeans and 250 million T-shirts (point *A′*). After specialization and trade, it consumes 80 million bushels of soybeans and 260 million T-shirts (point *C′*).

Notice that points *C* and *C′* lie *beyond* each country's PPF. While the PPF still shows possibilities for *production* of the two goods, *consumption* is no longer limited to what is produced. Instead, with trade, a country can consume more of both goods than it would be capable of producing and consuming on its own.

FIGURE 2 The Gains from Specialization and Trade



With international trade, the U.S. moves production from point A *to point* B*, but consumes at point* C*—beyond its* PPF*. Soybean production increases from 100 million (at point* A*) to 200 million (at point* B*). After exporting 80 million bushels to China, the United States is left with 120 million (at point* C*), for a net gain of 20 million bushels. With terms of trade assumed to be 3 T-shirts for 1 bushel of soybeans, the United States trades its soybeans for 240 million T-shirts. United States T-shirt consumption rises from 200 million (point* A*) to 240 million (point* C*).*

In China, production moves from point A′ *to point* B′*, but consumption is at point* C′*. T-shirt production rises from 250 million (at point* A′*) to 500 million (at point* B′*). After exporting 240 million T-shirts to the United States, China is left with 260 million (point* C′*), for a net gain of 10 million. China trades its T-shirts for 80 million bushels of soybeans from the United States, so China's soybean consumption rises from 50 million (at point* A′*) to 80 million (at point* C′*).*

Let's take a step back and consider what we've discovered. First, look back at Table 1. Based on the required labor hours, the United States has an *absolute advantage* in both goods: It can produce both soybeans and T-shirts using fewer hours of labor than can China. But we've determined that the United States has a *comparative advantage* in only *one* of these goods—soybeans—and China has a comparative advantage in the other—T-shirts. This is because the *opportunity costs* of each good differ in the two countries. Then, in Figure 1, we saw how world production of both goods increases when each country shifts its resources toward its comparative advantage good. Finally, in the last row of Table 2 and in Figure 2, we saw that *international trade* can enable *each* country to end up with more of *both* goods.

Although we've illustrated this result for just two countries and two goods, it also holds more generally:

> *Through international trade based on comparative advantage, all nations can achieve greater total consumption of goods and services, and therefore higher living standards, than is possible without trade.*

The Terms of Trade

In our ongoing example, China exports 240 million T-shirts in exchange for 80 million bushels of soybeans. This exchange ratio (240 million to 80 million, or 3 to 1) is known as the **terms of trade**—the quantity of one good that is exchanged for one unit of the other.

Terms of trade The ratio at which a country can trade domestically produced products for foreign-produced products.

The terms of trade determine how the gains from international trade are *distributed* among countries. Our particular choice of 3 to 1 apportioned the gains as shown in the last row of Table 2. But with different terms of trade, the gains would have been apportioned differently. In the problems at the end of this chapter, you will be asked to recalculate the gains for each country with different terms of trade. You'll see that with different terms of trade, both countries still gain, but the distribution of the gains between countries changes.

But notice that the terms of trade were not even *used* in our example until we arrived at Table 2. The gains from trade for the *world as a whole* were demonstrated in Figure 1, and were based entirely on the increase in world production when countries specialize according to comparative advantage.

> *For the world as a whole, the gains from international trade are due to increased production as nations specialize according to comparative advantage.* How *those world gains are distributed among specific countries depends on the terms of trade.*

We won't consider here precisely *how* the terms of trade are determined (it's a matter of supply and demand). But we *will* establish the limits within which the terms of trade must fall.

Look again at Table 1. China would never give up *more* than 5 T-shirts to import 1 bushel of soybeans. Why not? Because it could always get a bushel for 5 T-shirts *domestically,* simply by shifting resources into soybean production.

Similarly, the United States would never export a bushel of soybeans for *fewer* than 2 T-shirts because it could get 2 T-shirts for a bushel domestically (again, by shifting resources). Therefore, when these two nations trade, we know the terms of trade will lie *somewhere between* 5 T-shirts for 1 bushel and 2 T-shirts for 1 bushel. Outside of that range, one of the two countries would refuse to trade. Note that in our example, we assume terms of trade of 3 T-shirts for 1 bushel—well within the acceptable range.

Countries can gain when they shift production toward their comparative advantage goods (such as soybeans in the United States and), and trade them for other goods from other countries (such as textiles in China).

Some Provisos about Specialization

Our simple example seems to suggest that countries should specialize *completely,* producing *only* the goods in which they have a comparative advantage. That is, it seems that China should get out of soybean production *entirely,* and the United States should get out of T-shirt production *entirely*.

The real world, however, is more complicated than our simplified example might suggest. Despite divergent opportunity costs, sometimes it does *not* make sense for two countries to trade with each other, or it might make sense to trade, but *not* completely specialize. Following are some real-world considerations that can lead to reduced trade or incomplete specialization.

Costs of Trading

If there are high transportation costs or high costs of making deals across national boundaries, trade may be reduced and even become prohibitively expensive. High

transportation costs are especially important for perishable goods, such as ice cream, which must be shipped frozen, and most personal services, such as haircuts, eye exams, and restaurant meals. These goods are less subject to trade according to comparative advantage. (Imagine the travel cost for a U.S. resident to see an optometrist in China, where eye exams are less expensive.)

The costs of making deals are generally higher for international trade than for trade within domestic borders. For one thing, different laws must be dealt with and different business and marketing customs must be mastered. In addition, international trade involves the exchange of one country's currency for another. This can introduce additional costs and risks that don't exist for domestic trade, because exchange rates can change before a contract is settled with payment. High transportation costs and high costs of making deals help explain why nations continue to produce some goods in which they do not have a comparative advantage and why there is less than complete specialization in the world.

Sizes of Countries

Our earlier example featured two large economies capable of fully satisfying each other's demands. But sometimes a very large country, such as the United States, trades with a very small one, such as the Pacific island nation of Tonga. If the smaller country specialized completely, its output would be insufficient to fully meet the demand of the larger one. While the smaller country would specialize *completely,* the larger country would not. Instead, the larger country would continue to produce both goods. This helps to explain why the United States continues to produce bananas, even though we do so at a much higher opportunity cost than many small Latin American nations.

Increasing Opportunity Cost

In our example, we have assumed that the labor hours required to produce another unit of each good remains constant as production changes. Therefore, opportunity costs remain constant as well. For example, the opportunity cost of a bushel of soybeans remains at 2 T-shirts for the United States, regardless of how many bushels it produces.

But more typically, the opportunity cost of a good rises as more of it is produced. (Why? You may want to review the law of increasing opportunity cost in Chapter 2.) In that case, each step on the road to specialization would change the opportunity cost. A point might be reached—before complete specialization—in which opportunity costs became *equal* in the two countries, and there would be no further mutual gains from trading. (Remember: Opportunity costs must *differ* between the two countries in order for trade to be mutually beneficial.) In the end, while trading will occur, there will not be complete specialization. Instead, each country will produce both goods, just as China and the United States each produce T-shirts *and* soybeans in the real world.

Government Barriers to Trade

Governments can enact barriers to trading. In some cases, these barriers increase trading costs; in other cases, they make trade impossible. Since this is such an important topic, we'll consider government-imposed barriers to trade in a separate section, later in the chapter.

The Sources of Comparative Advantage

We've just seen how nations can benefit from specialization and trade when they have comparative advantages. But what determines comparative advantage in the first place?

Resource Abundance and Comparative Advantage

In many cases, comparative advantage arises from the *resources* a country has at its disposal.

> *A country that has relatively large amounts of a particular resource will tend to have a comparative advantage in goods that make heavy use of that resource.*

Remember that resources include not just land and natural resources, such as coal or oil, but also labor, human and physical capital, and entrepreneurship. A country that is relatively abundant in *any* of these resources will tend to have a comparative advantage in those goods that use relatively large quantities of that resource.

Comparative Advantage Based on Natural Resources and Climate

The most obvious cases of comparative advantage arise from gifts of nature, such as a specific natural resource like oil or coal, or a climate especially suited to a particular product.

The top part of Table 3 contains some examples. Saudi Arabia has a comparative advantage in the production of oil because it has oil fields with billions of barrels of oil that can be extracted at low cost. The United States' comparative advantage in crops such as wheat and soybeans is partly explained by its abundant farmland.

TABLE 3

Examples of National Specialties in International Trade

Country	Specialization Resulting from Natural Resources or Climate
Saudi Arabia	Oil
Canada	Timber
United States	Grain
Spain	Olive oil
Mexico	Tomatoes
Jamaica	Aluminum ore
Italy	Wine
Israel	Citrus fruit
Niger	Uranium

Country	Specialization *Not* Based on Natural Resources or Climate
Japan	Cars, consumer electronics
United States	Software, movies, music, aircraft
Switzerland	Watches
Korea	Cars, steel, ships
China	Textiles, toys, shoes
Great Britain	Financial services
Pakistan	Textiles

Canada is a major exporter of timber because its climate and geography make its land more suitable for growing trees than other crops. Canada is a good example of comparative advantage without absolute advantage: It grows a lot of timber, not because it can do so using fewer resources than other countries, but because its land is even more poorly suited to growing other things.

But now look at the bottom half of Table 3. It shows examples of international specialization that arise from some cause *other* than natural resources. Japan has a strong comparative advantage in making automobiles. Yet none of the *natural* resources needed to make cars are available in Japan; the iron ore, coal, and oil needed to produce cars are all imported.

What explains the cases of comparative advantage in the bottom half of Table 3?

Comparative Advantage from Other Resources

Some of the examples in the lower half of Table 3 arise from resources *other* than natural resources or climate. The United States is rich in both physical capital and human capital. As a result, the United States tends to have a comparative advantage in goods and services that make heavy use of computers, tractors, and satellite technology, as well as goods that require highly skilled labor. This, in part, explains the U.S. comparative advantage in the design and production of aircraft, a good that makes heavy use of physical capital (such as computer-based design systems) and human capital (highly trained engineers).

The United States is also relatively abundant in entrepreneurship. This helps explain why many of the Internet-related innovations in recent years (Google's search engine, Facebook, YouTube, Twitter) and first-generation high-tech products (Apple's iPhone, Tivo's Digital Video Recorder) were developed in the United States. Creating these inventions and bringing them to market not only requires the human capital of scientists and managers, but also requires entrepreneurs who perceive market opportunities and are willing to take risks to exploit them.

In less developed countries, by contrast, capital and skilled labor are relatively scarce, but less-skilled labor is plentiful. Accordingly, these countries tend to have a comparative advantage in products that make heavy use of less-skilled labor, such as textiles and light manufacturing. Note, however, that as a country develops—and acquires more physical and human capital—its pattern of comparative advantage can change. Japan, Korea, and Singapore, after a few decades of very rapid development, acquired a comparative advantage in several goods that, at one time, were specialties of the United States and Europe—including automobiles, steel, and sophisticated consumer electronics.

Beyond Resources

Another aspect of the bottom half of Table 3 is harder to explain: Why do specific countries develop a *particular* specialty? For example, you may take the worldwide dominance of American movies for granted. But if you try to explain it based on the availability of resources like physical capital or highly skilled labor, or cultural traditions that encouraged artists, writers, or actors, then why not Britain or France? At the time the film industry developed in the United States, these two countries had similar endowments of physical and human capital, and much older and stronger theatrical traditions than the United States. Yet their film industries—in spite of massive government subsidies—are a very distant second and third compared to that of the United States.

In even the most remote corner of the world, the cars, cameras, and VCRs will be Japanese, the movies and music American, the clothing from Hong Kong or China, and the bankers from Britain. These specialties are certainly *consistent* with the capital and other resources each nation has at its disposal, but explaining why each *specific* case of comparative advantage arose in the first place is not easy.

We can, however, explain why a country *retains* its comparative advantage once it gets started. Japan today enjoys a huge comparative advantage in cars and consumer electronics in large part because it has accumulated a capital stock—both physical capital and human capital—well suited to producing those goods. The physical capital includes the many manufacturing plants and design facilities that the Japanese have built over the years. The human capital includes the accumulated knowledge and skills of Japanese managers, scientists, product designers, and factory workers.

The stocks of physical and human capital in Japan sustain its comparative advantage just as stocks of natural resources sustain comparative advantages in other countries. More likely than not, Japan will continue to have a comparative advantage in cars and electronics, just as the United States will continue to have a comparative advantage in making movies.

Countries often develop strong comparative advantages in the goods they have produced in the past, regardless of why they began producing those goods in the first place.

Why Some People Object to Free Trade

Given the clear benefits that nations can derive by specializing and trading, why would anyone ever *object* to free international trade? Why do the same governments that join the WTO turn around and create roadblocks to unhindered trade? The answer is not too difficult to find: Despite the benefit to the nation as a whole, some groups within the country, especially in the short run, are likely to lose from free trade, even while others gain a great deal more. Unfortunately, instead of finding ways to compensate the losers—to make them better off as well—we often allow them to block free trade policies. The simple model of supply and demand helps illustrate this story.

Figure 3 shows the market for shrimp in the United States. Both the supply and demand curve in the figure represent the *domestic* market only. That is, the supply curve tells us the quantity supplied at each price by U.S. producers; the demand curve tells us quantity demanded at each price by U.S. consumers. With no international trade in shrimp, the U.S. market would achieve equilibrium at point *A*, at a price of $7 per pound. This relatively high price reflects the relatively high opportunity cost of producing shrimp in the United States. Both production and consumption would be 400 million pounds per year.

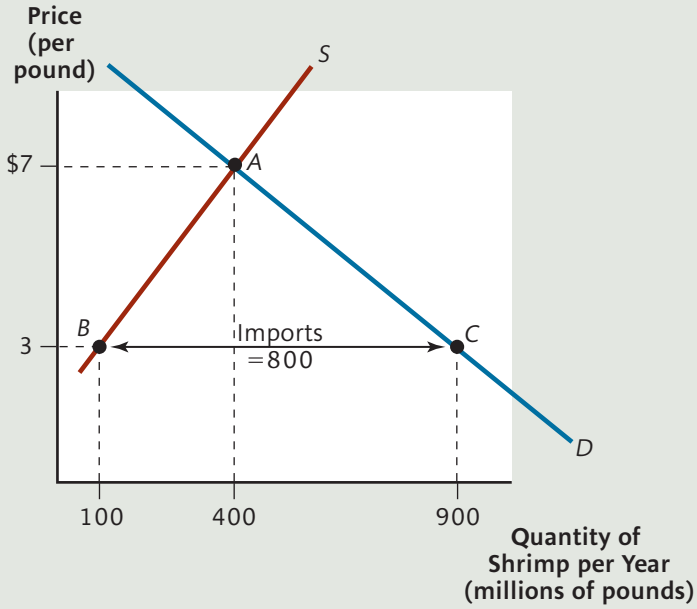
The United States does not have a comparative advantage in shrimp. Other countries that *do* have a comparative advantage (such as Vietnam, Thailand, and Brazil) would like to sell it to us. Moreover, because their opportunity cost of producing shrimp is less than in the United States, their price tends to be lower as well. Let's suppose that the *world price* of shrimp—the price at which other countries offer to sell it to Americans—is $3 per pound. To keep our example simple, we'll also assume this price remains constant, no matter how much shrimp Americans buy from the

FIGURE 3 The Impact of Trade

With no international trade, equilibrium in the U.S. market for shrimp is at point A, *where domestic quantity supplied equals domestic quantity demanded. Price is $7 per pound, and 400 million pounds are consumed each year.*

When U.S. consumers can import shrimp at the lower world price of $3.00 per pound, quantity supplied falls to 100 million pounds, while quantity demanded rises to 900 million. The difference—800 million pounds—is imports from the world market.

Consumers of shrimp gain from trade—they enjoy a greater quantity at a lower price. But producers lose—they sell less at a lower price.



rest of the world. (In effect, we're assuming that under international trade, the United States would be a relatively small buyer in a much larger world market.)

Now let's open up free trade in shrimp. Because Americans can buy unlimited quantities of imported shrimp at $3, domestic producers will have to lower their price to $3 as well in order to sell any. So the price of *all* shrimp in the U.S. market falls to $3 per pound—the same as the world price.

As the price drops, two things happen. On the one hand, we move along the demand curve from point *A* to point *C:* U.S. consumers buy more shrimp (900 million pounds) because it is cheaper. On the other hand, we move along the supply curve from point *A* to point *B:* U.S. producers decrease their quantity supplied (to 100 million pounds). The difference between domestic supply of 100 million and domestic demand of 900 million is the amount of shrimp the United States imports each year: 800 million pounds.

You've already learned (in the last section) that international trade according to comparative advantage makes each country as a whole better off: It increases total world production and enables consumers to enjoy greater quantities of goods and services. But not *everyone* is better off. It is easy to figure out who will be happy and who will be unhappy in the United States. American consumers are delighted: They are buying more shrimp at a lower price. American shrimp producers are miserable: They are selling less shrimp at a lower price.

International trade makes each country, as a whole, better off. But not everyone gains, because cheap imports from abroad—while beneficial to domestic consumers—are harmful to domestic producers.

The Anti-Trade Bias

Imagine that a bill comes before Congress to prohibit or restrict the sale of cheap shrimp from abroad, so that its U.S. price can rise above $3.00. Domestic producers would favor the bill. Domestic consumers would oppose it. But not with equally loud voices. After all, the harm to consumers from this restriction of trade would be spread widely among *all* U.S. consumers. The loss to any individual would be very small. For example, if the total loss to U.S. consumers were $200 million per year, the total harm to any single consumer would be less than a dollar. As a result, no individual consumer of shrimp has a strong incentive to lobby Congress, or to join a dues-paying organization that would act on behalf of shrimp consumers to oppose this antitrade bill.

By contrast, the benefits from this restriction of trade would be highly concentrated on a much smaller group of people: those who work in or own firms in the domestic shrimp industry. They have a powerful incentive to lobby against free trade in shrimp. Not surprisingly, when it comes to trade policy, the voices raised *against imports* are loud and clear, while those *for imports* are often nonexistent. Since a country has the power to restrict imports from other countries, the lobbying can—and often does—lead to a restriction on free trade. The United States, for example, continues to restrict imports of shrimp from low-cost producers, largely due to powerful lobbying by the U.S. shrimp industry.

For any particular good or service, the costs from expanded trade are highly concentrated among relatively few parties, while the benefits are widely dispersed among many. As a result, those harmed by international trade generally have more incentive to mobilize and lobby than those who benefit.

A similar process works against U.S. *exports* to other countries. In this case, the *foreign* producers who would have to compete with U.S. goods will complain loudest, while foreign consumers who stand to gain will be mostly silent. The U.S. exporters—who are not constituents of these foreign governments—will have little influence in the debate. Thus, just as there is a policy bias against U.S. imports in the United States, there is a policy bias against U.S. exports in other countries.

Some Antidotes to the Anti-Trade Bias

As we've seen, those harmed by international trade in a good often have a stronger incentive to mobilize than those who benefit. As a result, new agreements to reduce trade barriers are often blocked, and new barriers are often created. There are, however, some forces that work *against* the anti-trade bias and in favor of expanded international trade.

The World Trade Organization

One important antidote to the anti-trade bias is the World Trade Organization. By setting standards for acceptable and unacceptable trade restrictions and making rulings in specific cases, the WTO has some power to influence nations' trade policies. But its influence is limited, because the WTO has no enforcement power. For example, after the WTO, in 2003, ruled against a U.S. policy that encouraged companies to seek trade barriers from the U.S. government, the U.S. continued to violate the ruling for several years. And the WTO has ruled several times against European trade barriers against U.S. beef, with little effect. Still, a negative WTO ruling puts

public relations pressure on a country and allows a nation harmed by restrictions on its exports to retaliate, in good conscience, with its own trade barriers.

All or Nothing Trade Agreements

In a bilateral or multilateral trade agreement, two or more countries agree to trade freely in many goods—or even all goods—simultaneously. These agreements are typically negotiated by government officials and then presented to legislatures as "all-or-nothing" deals: The agreement must be approved or rejected as a whole, without any amendments that make exceptions for specific industries.

Such agreements can bring in another constituent to lobby for free trade: *exporters* in both countries. Ordinarily, exporters have no ability to influence the debate because their ability to export is decided in the *importing* country, where they have little influence. But in an all-or-nothing free trade deal, they can lobby their *own* country to allow imports as a way of enabling them to sell their exports. In this way, a balance of forces is created. Domestic producers threatened by imports will lobby against trade agreements in each country. But potential exporters will lobby just as strongly *for* the agreement.

An example was the North American Free Trade Agreement (NAFTA) between the United States, Canada, and Mexico, which went into effect in 1994, and has eliminated barriers on most products produced by the three nations. NAFTA was hotly opposed in all three countries by many producers and some labor unions who stood to lose from imports, but was just as hotly favored by producers and workers who stood to gain from exports. (The biggest gainers—consumers in the three countries—were hardly involved in the debate, for reasons we've discussed.)

More recently, a similar conflict arose over the Dominican Republic–Central American Free Trade Agreement (DR-CAFTA), between the United States and six other countries. In each country (including the U.S.) producers who would have to compete with imports lobbied against the bill, while exporters in each country lobbied for it. In mid-2005, the U.S. Congress narrowly approved the agreement; by 2007, the other countries involved had given their approval as well.

Industries as Importers

When we hear the word *imports,* we might think only about consumer goods: French wine, Japanese cars, Columbian coffee. But many imported goods and services are *inputs* used by domestic industries. If imported inputs are an important part of firms' costs, they have an incentive to lobby for free trade in the good. For example, in late 2004, the textile industry lobbied the Bush administration to slow the rise in clothing imports from China. But U.S. clothing retailers and importers—for whom clothing is an input—lobbied strongly *against* any trade barriers. While the retailers and importers ultimately lost the battle in May 2005, their opposition delayed the restrictions for months and influenced the final policy adopted.

How Free Trade Is Restricted

So far in this chapter, you've learned that specialization and trade according to comparative advantage can dramatically improve the well-being of entire nations. This is why governments generally favor free trade. Yet international trade can, in the short run, hurt particular groups of people. These groups often lobby their government to restrict free trade.

When governments decide to accommodate the opponents of free trade, they are apt to use one of two devices to restrict trade: tariffs or quotas.

Tariffs

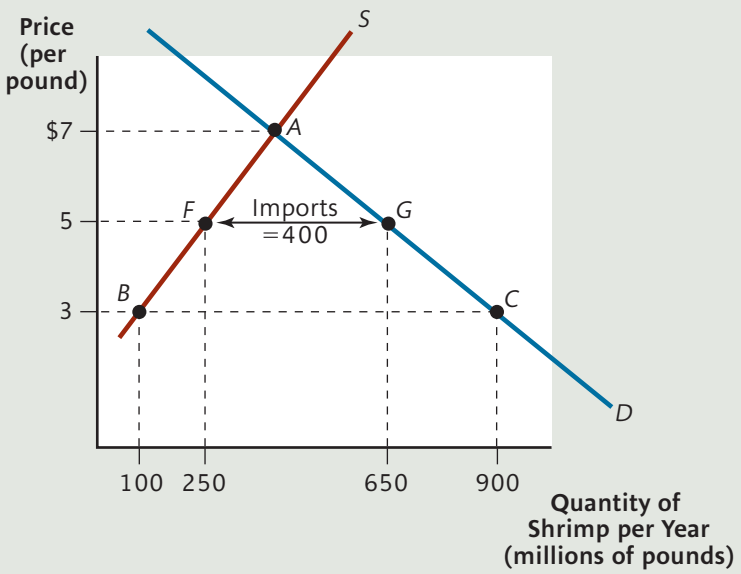
A **tariff** is a tax on imported goods. It can be a fixed dollar amount per physical unit, or it can be a percentage of the good's value. In either case, the effect in the tariff-imposing country is similar.

Tariff A tax on imports.

Figure 4 illustrates the effect of a U.S. tariff of $2 per pound on imported shrimp. Before the tariff is imposed, the price of shrimp under free trade is the world price: $3 per pound. The U.S. imports 800 million pounds per year (the distance *BC*). When the tariff is imposed, U.S. importers must still pay the same $3 per pound to their foreign suppliers. But now they must also pay $2 per pound to the U.S. government. Thus, the price of imported shrimp will rise from $3 to $5 per pound to cover the additional cost of the tariff.[1]

The higher price for imported shrimp allows U.S. producers to charge $5 for their domestic shrimp as well. As the price of shrimp rises, domestic quantity supplied increases (a movement along the supply curve from point *B* to point *F*). At the same time, domestic quantity demanded *decreases* (a movement along the demand curve from point *C* to point *G*). The final result is a reduction in imports, from 800 million before the tariff to 400 million after the tariff.

FIGURE 4 The Effects of a Tariff



With free trade in shrimp, the price in the United States is the same as the world price: $3.00 per pound. U.S. imports are equal to the distance from B *to* C*, 800 million pounds per year.*

A tariff of $2.00 per pound raises the price of imported shrimp to $5.00 per pound, and the price of domestically produced shrimp rises to the same level. Domestic quantity supplied increases from 100 million to 250 million (the move from A *to* B*), while domestic quantity demanded falls from 900 million to 650 million (the move from* C *to* G*). The result is lower imports of 400 million pounds.*

Domestic suppliers gain from the tariff: They sell more shrimp at a higher price. But domestic consumers lose: They pay a higher price and consume less.

[1] If the United States is a large buyer in the world market for shrimp, the reduction in imports caused by the tariff would cause the world price to fall. This would change the quantitative results in our example. However, the price in the United States would still rise above $3.00, and all of our conclusions about the impact of tariffs would still hold.

As you can see, American consumers are worse off: They pay more for shrimp and enjoy less of it. But U.S. producers are better off: They sell more shrimp at a higher price.

But we also know this: Since the volume of trade has decreased, the gains from trade according to comparative advantage have been reduced as well. The United States, as a whole, is worse off as a result of the tariff.

A tariff reduces the volume of trade and raises the domestic price of an imported good. In the country that imposes the tariff, producers of the good gain, but consumers lose. The country as a whole loses, because tariffs decrease the volume of trade and therefore decrease the gains from trade.

Quotas

Quota A limit on the physical quantity of imports.

A **quota** is a government decree that limits the imports of a good to a specified maximum physical quantity, such as 400 million pounds of shrimp per year. Because the goal is to restrict imports, a quota is set below the level of imports that would occur under free trade. Its general effects are very similar to the effects of a tariff.

Figure 4, which we used to illustrate tariffs, can also be used to analyze the impact of a quota. In the figure, we start with our free trade price of $3. Consumers are buying 900 million pounds, and domestic producers are selling 100 million pounds per year. The difference of 800 million pounds is satisfied by imports.

Now suppose the United States imposes a quota of 400 million pounds (equal to the distance *FG* in the figure). At $3 per pound, the gap between the domestic supply curve and the domestic demand curve would still be 900 million pounds, which is more than the 400 million pounds of foreign shrimp allowed into the country. There is an excess demand for shrimp, which drives up the price. The price will keep rising until the gap between the supply and demand curves shrinks to the quantity allowed under the quota. As you can see in the figure, only when the price rises to $5 would the gap shrink to 400 million pounds. Thus, a quota of 400 million gives us exactly the same result as did a tariff of $2: In both cases, the price rises to $5, and yearly imports shrink to 400 million pounds.

A quota has effects similar to a tariff: It reduces imports, raises the domestic price, thereby helping domestic producers of the good but reducing the gains from trade to the country as a whole.

Quotas versus Tariffs

The previous discussion seems to suggest that tariffs and quotas are pretty much the same. But even though the price and level of imports may end up being the same, there is one important difference between these two trade-restricting policies. A tariff, after all, is a *tax* on imported goods. Therefore, when a government imposes a tariff, it collects some revenue every time a good is imported. Even though the country loses from a tariff, it loses a bit less (compared to a quota) because at least it collects some revenue from the tariff. This revenue can be used to fund government programs or reduce other taxes, to the benefit of the country as a whole. When a government imposes a quota, however, it typically gains no revenue at all.

Both quotas and tariffs reduce the gains from trade. But a tariff has one saving grace that a quota lacks: increased tax revenue.

Economists, who generally oppose measures such as quotas and tariffs to restrict trade, argue that, if one of these devices must be used, tariffs are the better choice. While both policies reduce the gains that countries can enjoy from specializing and trading with each other, the tariff provides some compensation in the form of additional government revenue.

Protectionism

This chapter has outlined the *gains* that arise from international trade, but it has also outlined some of the *pain* trade can cause to different groups within a country. While the country as a whole benefits, those who own or work in firms that have to compete with cheap imports will be harmed. The groups who suffer from trade with other nations have developed a number of arguments against free trade. Together, these arguments form a position known as **protectionism**—the belief that a nation's industries should be *protected* from free trade with other nations.

Protectionism The belief that a nation's industries should be protected from foreign competition.

Protectionist Myths

Some protectionist arguments are rather sophisticated and require careful consideration. We'll consider some of these a bit later. But antitrade groups have also promulgated a number of myths to support their protectionist beliefs. Let's consider some of these myths.

Myth #1. "A high-wage country cannot afford free trade with a low-wage country. The high-wage country will either be undersold in everything and lose all of its industries, or else its workers will have to accept equally low wages and equally low living standards."

It's true that some countries have much higher wages than others. Here are 2007 figures for average hourly wages of manufacturing production workers, including benefits such as holiday pay and health insurance: Germany, $37.66; United States, $24.59; Japan, $19.75; Italy, $28.23; Korea, $16.02; Singapore, $8.35; Brazil, $5.96; Mexico, $2.92; and less than a dollar in China and Bangladesh. This leads to the fear that the poorer countries will be able to charge lower prices for their goods, putting American workers out of jobs unless they, too, agree to work for low wages.

But this argument is incorrect, for two reasons. First, it is true that American workers are paid more than Chinese workers, but this is because the average American worker is more *productive* than his or her Chinese counterpart. After all, the American workforce is more highly educated, and American firms provide their workers with more sophisticated machinery than do Chinese firms. If an American can produce more output than a Chinese worker in an hour, then even though wage rates in the United States may be greater, cost *per unit* produced can still be lower in the United States.

But suppose the cost per unit *were* lower in China. Then there is still another, more basic argument against the fear of a general job loss or falling wages in the United States: comparative advantage. Let's take an extreme case. Suppose that

labor productivity were the same in the United States and China, so that China—with lower wages—could produce *everything* more cheaply than the United States could. Both countries would still gain if China specialized in products in which its cost advantage was relatively large and the United States specialized in goods in which China's cost advantage was relatively small. That is, the United States would still have a comparative advantage in some things and there would be mutual gains from trade.

Myth #2. "A low-productivity country cannot afford free trade with a high-productivity country. The former will be clobbered by the latter and lose all of its industries."

This argument is the flip side of the first myth. Here, it is the poorer, less-developed country that is supposedly harmed by trade with a richer country. But this myth confuses absolute advantage with comparative advantage. Suppose the high-productivity country (say, the United States) could produce *every* good with fewer resources than the low-productivity country (say, China). Once again, the low-productivity country would *still* have a comparative advantage in *some* goods. It would then gain by producing those goods and trading with the high-productivity country. This is the case in our hypothetical example that began with Table 1. In that example, the United States has an absolute advantage in both goods, yet—as we've seen—trade still benefits both countries.

To make the point even clearer, let's bring it closer to home. Suppose there is a small, poor town in the United States where workers are relatively uneducated and work with little capital equipment, so their productivity is very low. Would the residents of this town be better off sealing their borders and not trading with the rest of the United States, which has higher productivity? Before you answer, think what this would mean: The residents of the poor town would have to produce everything on their own: grow their own food, make their own cars and television sets, and even provide their own entertainment. Clearly, they would be worse off in isolation. And what is true *within* a country is also true *between* different countries: Closing off trade will make a nation, as a whole, worse off, regardless of its level of wages or productivity. Even a low-productivity country is made better off by trading with other nations.

Myth #3. International trade decreases the total number of jobs in a country.

It is true that a sudden opening up of trade temporarily disrupts markets. There can even be a temporary drop in employment as jobs are lost in some sectors before they are created in other sectors. But neither logic nor observation supports the view that international trade causes any long-lasting drop in total employment. In the United States, for example, as international trade has expanded rapidly in recent decades, total employment has risen steadily. And the U.S. unemployment rate has trended downward, not upward.

This myth about losses in total employment comes from looking at only one side of the international trade coin: imports. It is true that international trade destroys jobs in those industries that now have to compete with cheaper imports from abroad. But trade also creates *new* jobs in the export sector. When trade is balanced—exports and imports are equal—there is no reason to expect the jobs lost in the import-competing sector to exceed the jobs gained in the export sector.

What about when trade is unbalanced, as when a country runs a *trade deficit* (the value of imports exceeds the value of exports)? Even in this case, there is no

reason for total employment to decrease. The United States has run a trade deficit every year for decades: We spend more dollars buying imports from other countries than they return to us by buying our products. As a result, producers in these other countries start to pile up dollar balances. But they don't just hold onto these dollar balances, which pay no interest or other return. Instead, they invest these dollars in U.S. financial markets, purchasing stocks and bonds and making bank deposits. The funds are then lent out to U.S. firms and households who, in turn, spend them—on new capital equipment, new housing construction, or other things. In this way, while some jobs are lost when Americans spend their dollars on imports rather than U.S. goods, other jobs are created when the dollars flow back into the U.S. through the financial markets. A trade deficit can cause other problems for a country (as you will learn when you study macroeconomics), but it does not reduce total employment.

Myth #4. "In recent times, the declining wages of America's unskilled workers are due to ever-expanding trade between the United States and other countries."

True enough, unskilled workers have lost ground over the past 25 years. College graduates have enjoyed growing purchasing power from their earnings, while those with only a high school education or less have lost purchasing power. Rising trade with low-wage countries has been blamed for this adverse trend.

But before we jump to conclusions, let's take a closer look. Recall (from Chapter 12) that the college wage premium can be explained by supply and demand. For college graduates, demand has been increasing especially rapidly due to skill-biased technological change, while supply has not kept pace. For high school graduates, the same technological change has *slowed* the growth of demand. These forces suggest we would observe a widening gap between college and high school graduates even if there were *no* international trade.

Still, it is true that international trade contributes to this gap. All else equal, U.S. imports of goods that would otherwise be produced by low-wage labor shrink the demand for low-wage labor in the United States. And U.S. exports of goods produced by high-wage labor raise the demand for such labor in the United States.

Economists who have looked at the impact of trade on U.S. labor markets have consistently concluded that international trade is a small contributor to the depressed earnings of low-wage workers in the United States. Technological change, and the greater skills needed to work with new technologies, appears to be the main driver.

But suppose that international trade were just as important as technological change in depressing low-wage workers' earnings. Would trade barriers be justified? Only if barriers to technological progress are also justified. After all, both technological progress and international trade have remarkably similar effects. They both raise living standards and benefit the country as a whole. And they both disproportionately benefit high-wage workers while causing harm to many low-wage workers. If concern for low-wage workers is grounds for opposing international trade, then logical consistency suggests it is grounds for opposing technological progress too.

Most economists believe that neither technological change nor international trade should be opposed. Rather, we should take advantage of the higher average living standards they enable, and use other policies to address the problems of low-wage workers more directly. The other policies include the earned income tax

credit (discussed in Chapter 12), the removal of artificial barriers to attending college, government assistance for college tuition, job retraining programs, and more.

Sophisticated Arguments for Protection

While most of the protectionist arguments we read in the media are based on a misunderstanding of comparative advantage, some more recent arguments for protecting domestic industries are based on a more sophisticated understanding of how markets work. These arguments have become collectively known as **strategic trade policy.** According to its proponents, a nation can gain in some circumstances by assisting certain *strategic industries* that benefit society as a whole, but that may not thrive in an environment of free trade.

Strategic trade policy Protectionist policies designed to capture social benefits, such as greater tax revenue, from having an industry in the domestic country.

Strategic trade policy is most effective in situations where a market is dominated by a few large firms. With few firms, the forces of competition—which ordinarily reduce profits in an industry to very low levels—will not operate. Therefore, each firm in the industry may earn high profits. These profits benefit not only the owners of the firm but also the nation more generally, since the government will be able to capture some of the profit with the corporate profits tax. When a government helps an industry compete internationally, it increases the likelihood that high profits—and the resulting general benefits—will be shifted from a foreign country to its own country. Thus, interfering with free trade—through quotas, tariffs, or even a direct subsidy to domestic firms—might actually benefit the country.

Infant industry argument The argument that a new industry in which a country has a comparative advantage might need protection from foreign competition in order to flourish.

An argument related to strategic trade policy is the **infant industry argument.** This argument begins with a simple observation: In order to enjoy the full benefits of trade, markets must allocate resources toward those goods in which a nation has a comparative advantage. This requires well-organized markets for resources such as labor and land, and also well-organized *financial markets,* where firms obtain funds for new investment. But in some countries—especially developing countries—financial markets do not work very well. Poor legal systems or incomplete information about firms and products may prevent a new industry from obtaining financing, even though the country would have a comparative advantage in that industry once it was formed. In this case, protecting the infant industry from foreign competition may be warranted until the industry can stand on its own feet.

Strategic trade policy and support for infant industries are controversial. Opponents of these ideas stress three problems:

1. Once the principle of government assistance to an industry is accepted, special-interest groups of all kinds will lobby to get the assistance, whether it benefits the general public or not.
2. When one country provides assistance to an industry by keeping out foreign goods, other nations may respond in kind. If they respond with tariffs and quotas of their own, the result is a shrinking volume of world trade and falling living standards. If subsidies are used to support a strategic industry, and another country responds with its own subsidies, then both governments lose revenue, and neither gains the sought-after profits.
3. Strategic trade policy assumes that the government has the information to determine which industries, infant or otherwise, are truly strategic and which are not.

Still, the arguments related to strategic trade policy suggest that government protection or assistance *may* be warranted in some circumstances, even if putting

this support into practice proves difficult. Moreover, the arguments help to remind us of the conditions under which free trade is most beneficial to a nation:

> *Production is most likely to reflect the principle of comparative advantage when firms can obtain funds for investment projects and when they can freely enter industries that are profitable. Thus, free trade, without government intervention, works best when markets are working well.*

This may explain, in part, why the United States, where markets function relatively well, has for decades been among the strongest supporters of the free trade ideal.

Protectionism in the United States

Americans can enjoy the benefits of importing many of the products listed in Table 3: olive oil from Spain, watches from Switzerland, tomatoes from Mexico, cars and VCRs from Japan. But on the other side of the ledger, U.S. consumers have suffered and U.S. producers have gained from some persistent barriers to trade. Table 4 lists some examples of American protectionism—through tariffs, quotas, or similar policies—that have continued for years.

As you can see, protectionism is costly. Quotas and tariffs on apparel and textiles, the most costly U.S. trade barrier, force American consumers to pay $33.6 billion more for clothes each year. And while protection saves an estimated 168,786 workers in this industry from having to make the painful adjustment of finding other work, it does so at an annual cost of $199,241 per worker. Both workers and consumers could be made better off if textile workers were paid any amount up to $199,241 *not* to work and consumers were allowed to buy inexpensive textiles from abroad.

In some cases, the cost per job saved is staggering. The table shows that trade barriers preventing Americans from buying inexpensive luggage save just a couple of hundred jobs, at a yearly cost of more than $1 million each. Trade barriers on sugar are almost as bad: While 2,261 jobs are saved, the annual cost per job is $826,104.

In addition to the dozens of industries in the United States permanently protected from foreign competition, dozens more each year are granted temporary

TABLE 4

Some Examples of U.S. Protectionism

Protected Industry	Annual Cost to Consumers	Number of Jobs Saved	Annual Cost per Job Saved
Apparel and Textiles	$33,629 million	168,786	$ 199,241
Maritime Services	$ 2,522 million	4,411	$ 571,668
Sugar	$ 1,868 million	2,261	$ 826,104
Dairy Products	$ 1,630 million	2,378	$ 685,323
Softwood Lumber	$ 632 million	605	$1,044,271
Women's Nonathletic Footwear	$ 518 million	3,702	$ 139,800
Glassware	$ 366 million	1,477	$ 247,889
Luggage	$ 290 million	226	$1,285,078
Peanuts	$ 74 million	397	$ 187,223

Source: *The Fruits of Free Trade*, Federal Reserve Bank of Dallas, Annual Report, 2002, Exhibit 11.

protection when the U.S. government finds a foreign producer or industry guilty of *dumping*—selling their products in the United States at "unfairly" low prices that harm a U.S. industry. Most economists believe that these low prices are most often the result of comparative advantage, and that the United States as a whole would gain from importing the good. Vietnam, for example, has a clear comparative advantage in producing shrimp. But in 2005, based on a complaint by the Southern Shrimp Alliance, the U.S. government imposed tariffs of up to 26 percent on Vietnamese shrimp.

In the Using the Theory section that follows, we take a closer look at one of the longest-running examples of protectionism in the United States.

Using the Theory

THE U.S. SUGAR QUOTA[2]


© RICHARD LORD/PHOTOEDIT, INC.

The United States has protected U.S. sugar producers from foreign competition since the 1930s. Since the 1980s, the protection has been provided in the form of a price guarantee. Essentially, the government has promised U.S. sugar beet and sugar cane producers and processors that they can sell their sugar at a predetermined price—22 cents a pound—regardless of the world price of sugar.

This may not sound like a high price for sugar. But in the rest of the world, people and businesses can buy sugar for a lot less. From 2000 to 2009, the world price of sugar has averaged about 10 cents a pound, while Americans have continued to pay about 22 cents. Even in the late 1990s, when the world price of sugar plunged to just 5 cents a pound—a bonanza for sugar buyers around the world—American buyers were not invited to the party: The United States price remained at 22 cents.

Because the world price of sugar is so consistently below the U.S. price, the government cannot keep its promise to support sugar prices while simultaneously allowing free trade in sugar. With free trade, the price of sugar in the United States would plummet. The government's solution is a sugar quota. More accurately, the government decides how much foreign sugar it will allow into the United States each year, free of any tariff; all sugar beyond the allowed amount is hit with a heavy tariff of about 16 cents a pound. Since the tariff is so high, no one in the United States imports sugar beyond the allowed amount. So, in effect, the United States has a sugar quota.

The *primary* effects of the sugar quota are on sugar producers and sugar consumers. As you've learned, an import quota raises the domestic price of sugar (the quota's purpose). Sugar producers benefit. But sugar consumers are hurt even more.

And the harm is substantial. Table 4 shows that American consumers pay almost $2 billion more each year for sugar and products containing sugar due to the sugar quota. But spread widely over the

[2] Information in this section is based on: Mark A. Groombridge, "America's Bittersweet Sugar Policy," *Trade Briefing Paper No. 13,* Cato Institute, December 4, 2001; Lance Gay, "Soured on Sugar Prices, Candy Makers Leave the U.S.," *Scripps Howard News Service,* June 18, 2003; "Closing the 'Stuffed Molasses' Loophole," *White Paper,* United States Sugar Corporation (*http://www.ussugar.com/pressroom/white_papers/stuffed_molasses.html);* Remy Jurenas, "Sugar Policy Issues," *CRS Issue Brief for Congress,* Congressional Research Service, February 16, 2006, and U.S. Department of Agriculture—Briefing Room, "Sugar and Sweeteners: Policy," updated January 7, 2009.

U.S. population, this amounts to less than $15 per person per year. This probably explains why you haven't bothered to lobby for free trade in sugar.

But the costs of the sugar quota go beyond ordinary consumers. Industrial sugar users—such as the ice cream industry—are affected by the higher price too, not all of which can be passed on to consumers. So they try to avoid the quota's harm in other ways. One way is to waste resources buying sugar abroad disguised as other products. In the late 1990s and early 2000s, U.S. firms bought about 125,000 tons of sugar each year mixed with molasses, which was not restricted by the sugar quota. The sugar was then separated from the molasses. Even with these additional (and wasteful) processing costs, it was still a better deal to buy the disguised sugar abroad than to buy it through regular channels in the United States.

And sometimes a firm decides it's just not worth it anymore. In the past decade, several candy and baked-goods manufacturers have simply given up trying to pay the high cost of sugar in the United States, and moved their production facilities to other countries that don't have quotas, and where sugar can be purchased at the lower, world price.

Taxpayers, too, pay a cost for the sugar quota because as part of its price support program, the U.S. government must occasionally buy excess sugar from producers. In 2005, the U.S. government was storing about 759,000 tons of sugar at a cost of more than $1 million per month. The government must also hire special agents to detect and prevent sugar from entering the country illegally.

A final cost of the sugar quota is one that we have not yet considered in our discussion of international trade. In Figure 4, when a U.S. tariff or quota caused imports to shrink, we assumed that the world price of the good remained unchanged. But when a country is a very large buyer in the world market, a reduction in its purchases can cause the world price to drop. Essentially, the quota—by keeping sugar out—causes greater quantities of sugar to be dumped onto the world market, depressing its price. This hurts the poorest countries in the world that rely on sugar as an important source of export revenue. The sugar quota's harm to these countries has been estimated at about $1.5 billion per year.

Why do we bear all of these costs? Because of lobbying by groups who enjoy highly concentrated benefits. There are about 13,000 sugar farms in the United States. When the $2 billion in additional spending by U.S. consumers is spread among this small number of farms, the additional revenue averages out to more than $150,000 per farm per year. Those benefits are sizable enough to mobilize sugar producers each time their protection is threatened.

And mobilize they do. In 2004, the United States negotiated a free trade agreement with Australia that eliminated barriers on almost every good or service . . . except sugar. In 2005, the United States approved DR-CAFTA—a free trade agreement with five Central American countries and the Dominican Republic. Once again, sugar was an exception: Additional sugar imports from all six countries combined were restricted to less than 2 percent of the U.S. market.

Even the NAFTA agreement between the U.S., Mexico, and Canada put sugar in a special category: Restrictions on sugar from Mexico remained in place until 2008. When that year arrived, so did Mexican sugar. But the U.S. Congress came to the rescue with a new farm bill that promised even higher prices for U.S. sugar producers, from increased government purchases and continued restrictions on imports from other countries.

There is another group that receives concentrated benefits from the sugar quota: producers of high-fructose corn syrup, the closest substitute for sugar. Because of the sugar quota, high-fructose corn syrup can be sold at a substantially higher price.

Not surprisingly, the largest producer of high-fructose corn syrup in the U.S. market—the Archer Daniels Midland (ADM) company—has funded organizations that lobby Congress and try to sway public opinion in the United States. Occasionally, you may see a full-page newspaper advertisement paid for by one of these groups, arguing that sugar in the United States is cheap. And it is . . . until you find out what other countries are paying.

SUMMARY

A country has a *comparative advantage* in a good when it can produce it at a lower opportunity cost than another country. When countries specialize in the production of their comparative advantage goods, world production rises. Both countries benefit as consumption rises in each country. The distribution of the benefits between countries depends on the *terms of trade*—the rate at which the imported goods are traded for the exported goods.

Despite the benefits to each nation as a whole, those who supply goods that must compete with cheaper imports are harmed. Because the gains from trade are spread widely while the harm is concentrated among a smaller number of people, the latter have an incentive to lobby against free trade. Those harmed often encourage government to block or reduce trade through the use of *tariffs* (taxes on imported goods) and *quotas* (limits on the volume of imports). Tariffs and quotas both decrease the gains from trade through similar effects on prices and production. But tariffs at least provide additional government revenue, while quotas generally do not.

A variety of arguments have been proposed in support of protectionism. Some are clearly invalid and fail to recognize the principle that both sides gain when countries trade according to their comparative advantage. More sophisticated arguments for restricting trade may have merit in certain circumstances. These include strategic trade policy—the notion that governments should assist certain strategic industries—and the idea of protecting infant industries when financial markets are imperfect.

PROBLEM SET

Answers to even-numbered Questions and Problems can be found on the text Web site at www.cengage.com/economics/hall.

1. Suppose that the costs of production of winter hats and wheat in two countries are as follows:

	United States	Russia
Per winter hat	$10	5,000 rubles
Per bushel of wheat	$ 1	2,500 rubles

   a. What is the opportunity cost of producing one more winter hat in the United States? In Russia?
   b. What is the opportunity cost of producing one more bushel of wheat in the United States? In Russia?
   c. Which country has a comparative advantage in winter hats? In wheat?

2. Suppose that the Marshall Islands does not trade with the outside world. It has a competitive domestic market for VCRs. The market supply and demand curves are reflected in this table:

Price ($/VCR)	Quantity Demanded	Quantity Supplied
500	0	500
400	100	400
300	200	300
200	300	200
100	400	100
0	500	0

   a. Plot the supply and demand curves and determine the domestic equilibrium price and quantity.
   b. Suddenly, the islanders discover the virtues of free exchange and begin trading with the outside world. The Marshall Islands is a very small country, and so its trading has no effect on the price established in the world market. It can import as many VCRs as it wishes at the world price of $100 per VCR. In this situation, how many VCRs will be purchased in the Marshall Islands? How

many will be produced there? How many will be imported?
c. After protests from domestic producers, the government decides to impose a tariff of $100 per imported VCR. Now how many VCRs will be purchased in the Marshall Islands? How many will be produced there? How many will be imported?
d. What is the government's revenue from the tariff described in part (c)?
e. Compare the effect of the tariff described in part (c) with a quota that limits imports to 100 VCRs per year.

3. The following table provides hypothetical data about the supply and demand for beef in the European Union. The prices are in euros, and quantities are millions of pounds of beef per month. (You may wish to draw the supply and demand curves to help you visualize what is happening.)

Price	Quantity Supplied	Quantity Demanded
0	0	160
2	20	140
4	40	120
6	60	100
8	80	80
10	100	60
12	120	40

a. In the absence of international trade, what is the equilibrium price and quantity of beef?
b. If trade opens up, and the world price of beef is (and remains) 2 euros per pound of beef, how much beef will EU producers supply? How much beef will EU consumers demand? How much beef will be imported?
c. Within the EU, who gains and who loses when trade opens up?

4. Using the data on supply and demand in problem 3, suppose the EU imposed a tariff of 2 euros on each pound of beef.
a. How much beef would EU producers supply?
b. How much beef would EU consumers demand?
c. How much beef would the EU import?
d. How much total revenue would EU government authorities collect from the tariff?

5. Using the data on supply and demand in problem 3, suppose the EU imposed a quota on imports of beef equal to 40 million pounds of beef per month.
a. What would be the price of beef in the EU?
b. How much beef would EU producers supply?
c. How much beef would EU consumers demand?

6. Refer to Table 2 in the chapter. Suppose the terms of trade are *two and a half* T-shirts for each bushel of soybeans (instead of three for one as in the chapter). As in the chapter, assume the United States increases soybean production by 100 million bushels and exports 80 million of them to China, and that China decreases its own soybean production by 50 million bushels. Some of the remaining numbers in the table will have to change to be consistent with these new specifications. Then, answer each of the following questions.
a. Does China still gain from trade? Explain briefly.
b. Does the United States still gain from trade? Explain briefly.
c. Compare the effects of trade for China under the new and old terms of trade. In which case does China fare better? Explain briefly.
d. Compare the effects of trade for the United States under the new and old terms of trade. In which case does the United States fare better? Explain briefly.

7. Refer to Table 2 in the chapter. Suppose the terms of trade are *four* T-shirts for each bushel of soybeans (instead of three for one as in the chapter). Assume the United States increases soybean production by 100 million bushels and exports 60 million to China. China increases its T-shirt production by 250 million. Some of the remaining numbers in the table will have to change, to be consistent with these new specifications. Then, answer each of the following questions.
a. Does China still gain from trade? Explain briefly.
b. Does the United States still gain from trade? Explain briefly.
c. Compare the effects of trade for China under the new and old terms of trade. In which case does China fare better? Explain briefly.
d. Compare the effects of trade for the United States under the new and old terms of trade. In which case does the United States fare better? Explain briefly.

8. Redraw the PPFs for the United States and China from Figure 2 in the chapter. Assume that the initial production and consumption points (A and A') and the new production points (B and B') are the same as in that figure. Plot the new consumption points (C and C') that correspond to your results from the previous problem (7).

9. The following table shows the hypothetical labor requirements per ton of wool and per hand-knotted rug, for New Zealand and for India.

Labor Requirements per Unit

	New Zealand	India
Per ton of wool	10 hours	40 hours
Per hand-knotted rug	60 hours	80 hours

a. Which country has an absolute advantage in each product?
b. Calculate the opportunity cost in each country for each of the two products. Which country has a comparative advantage in each product?
c. If India produces one more rug and exports it to New Zealand, what is the lowest price (measured in tons of wool) that it would accept? What is the highest price that New Zealand would pay? Within what range will the equilibrium terms of trade lie?

10. Using the data from problem 9, suppose that New Zealand has 300 million hours of labor per period, while India has 800 million hours.
a. Draw PPFs for both countries for the two goods (put quantity of wool on the vertical axis).
b. Suppose that, before trade, each country uses half of its labor to produce wool and half to produce rugs. Locate each country's production point on its PPF (label it *A* for New Zealand, and *A′* for India).
c. After trade opens up and each country completely specializes in its comparative advantage good, locate each country's production point on its PPF (label it *B* for New Zealand, and *B′* for India).

More Challenging

11. This problem uses the data from problem 9, and the graphs you drew in problem 10. Suppose that the terms of trade end up at 4 tons of wool for 1 hand-knotted rug. Suppose, too, that New Zealand decides to export 12 million tons of wool to India.
a. How many rugs will New Zealand import from India?
b. What will be New Zealand's consumption of each good after trade?
c. What will be India's consumption of each good after trade?
d. On the PPFs you drew for problem 10, plot each country's consumption point after trade. Label it *C* for New Zealand, and *C′* for India.

12. In Figures 3 and 4, we assumed that the world price of a good was fixed, and not affected by the quantity of imports a country chooses. But if a country is large relative to the world market, its imports can influence the world price.

Suppose the market for good X involves only two large countries (A and B), with supply and demand schedules as shown below:

Country A			Country B		
Price per Unit of Good X (measured in dollars)	Quantity Demanded of Good X	Quantity Supplied of Good X	Price per Unit of Good X (measured in dollars)	Quantity Demanded of Good X	Quantity Supplied of Good X
$10	1	25	$10	5	11
9	2	22	9	6	10
8	3	19	8	7	9
7	4	16	7	8	8
6	5	13	6	9	7
5	6	10	5	10	6
4	7	7	4	11	5
3	8	4	3	12	4

a. Plot the supply and demand curves for each country.
b. Before international trade, what is the equilibrium price and quantity in each country? For the remaining questions, assume that the two countries can trade in good X.
c. Which country will export good X?
d. What will be the equilibrium world price? (Hint: This will be the price at which the quantity of exports from one country equals the quantity of imports to the other.)
e. What will happen to production and consumption in Country A?
f. What will happen to production and consumption in Country B?
g. What quantity will be exported (and also imported) in equilibrium?
h. On your graph, label the new levels of production and consumption in each country, as well as distances representing exports and imports.

A

Absolute advantage The ability to produce a good or service, using fewer resources than other producers use.
Accounting profit Total revenue minus accounting costs.
Adverse selection A situation in which asymmetric information about quality eliminates high-quality goods from a market.
Aggregation The process of combining distinct things into a single whole.
Alternate goods Other goods that firms in a market could produce instead of the good in question.
Alternate market A market other than the one being analyzed in which the same good could be sold.
Asymmetric information A situation in which one party to a transaction has relevant information not known by the other party.
Average cost pricing Setting a monopoly's regulated price equal to long-run average cost where the *LRATC* curve crosses the market demand curve.
Average fixed cost Total fixed cost divided by the quantity of output produced.
Average total cost Total cost divided by the quantity of output produced.
Average variable cost Total variable cost divided by the quantity of output produced.

B

Behavioral economics A subfield of economics focusing on decision-making patterns that deviate from those predicted by traditional consumer theory.
Black market A market in which goods are sold illegally at a price above the legal ceiling.
Bond A promise to pay back borrowed funds, issued by a corporation or government agency.
Budget constraint The different combinations of goods a consumer can afford with a limited budget, at given prices.
Budget line The graphical representation of a budget constraint, showing the maximum affordable quantity of one good for given amounts of another good.
Business firm An organization, owned and operated by private individuals, that specializes in production.

C

Capital A long-lasting tool that is used to produce other goods.
Capital gain The return someone gets by selling a financial asset at a price higher than they paid for it.
Capital loss The loss to the owner of an asset when it is sold for a price lower than its price when originally purchased.
Capital stock The total amount of capital in a nation that is productively useful at a particular point in time.
Capitalism A type of economic system in which most resources are owned privately.
Cartel A group of firms that selects a common price that maximizes total industry profits.
Ceteris paribus Latin for "all else remaining the same."
Change in demand A shift of a demand curve in response to a change in some variable other than price.
Change in quantity demanded A movement along a demand curve in response to a change in price.
Change in quantity supplied A movement along a supply curve in response to a change in price.
Change in supply A shift of a supply curve in response to a change in some variable other than price.
Circular flow A simple model that shows how goods, resources, and dollar payments flow between households and firms.
Coase theorem When a side payment can be arranged without cost, the market will solve an externality problem—and create the efficient outcome—on its own.
Command or centrally planned economy An economic system in which resources are allocated according to explicit instructions from a central authority.
Common resource A nonexcludable and rival good. Generally available free of charge, though efficiency would require a positive price.
Comparative advantage The ability to produce a good or service at a lower opportunity cost than other producers.
Compensating wage differential A difference in wages that makes two jobs equally attractive to a worker.
Complement A good that is used together with some other good.

Complementary input An input that is used *by* a particular type of labor, making it more productive.
Constant cost industry An industry in which the long-run supply curve is horizontal because each firm's cost curves are unaffected by changes in industry output.
Constant returns to scale Long-run average total cost is unchanged as output increases.
Consumer surplus The difference between the value of a unit of a good to the buyer and what the buyer actually pays for it.
Copyright A grant of exclusive rights to sell a literary, musical, or artistic work.
Coupon payments A series of periodic payments that a bond promises before maturity.
Critical assumption Any assumption that affects the conclusions of a model in an important way.
Cross-price elasticity of demand The percentage change in the quantity demanded of one good caused by a 1 percent change in the price of another good.

D

Deadweight loss The dollar value of potential benefits not achieved due to inefficiency in a particular market.
Decreasing cost industry An industry in which the long-run supply curve slopes downward because each firm's *LRATC* curve shifts downward as industry output increases.
Demand curve A graph of a demand schedule; a curve showing the quantity of a good or service demanded at various prices, with all other variables held constant.
Demand curve facing the firm A curve that indicates, for different prices, the quantity of output that customers will purchase from a particular firm.
Demand schedule A list showing the quantities of a good that consumers would choose to purchase at different prices, with all other variables held constant.
Derived demand The demand for a resource that arises from, and varies with, the demand for the product it helps to produce.
Diminishing marginal returns to labor The marginal product of labor decreases as more labor is hired.
Discount rate The interest rate used in computing present values.
Discounting The act of converting a future value into its present-day equivalent.
Discrimination When a group of people have different opportunities because of personal characteristics that have nothing to do with their abilities.
Diseconomies of scale Long-run average total cost increases as output increases.
Dividends Part of a firm's current profit that is distributed to shareholders.
Dominant strategy A strategy that is best for a player no matter what strategy the other player chooses.
Duopoly An oligopoly market with only two sellers.

E

Economic efficiency A situation in which every possible Pareto improvement is being exploited.
Economic profit Total revenue minus all costs of production, explicit and implicit.
Economics The study of choice under conditions of scarcity.
Economies of scale Long-run average total cost decreases as output increases.
Efficient market A market that instantaneously incorporates all available information relevant to a stock's price.
Elastic demand A price elasticity of demand greater than 1.
Entrepreneurship The ability and willingness to combine the other resources—labor, capital, and land—into a productive enterprise.
Equilibrium price The market price that, once achieved, remains constant until either the demand curve or supply curve shifts.
Equilibrium quantity The market quantity bought and sold per period that, once achieved, remains constant until either the demand curve or supply curve shifts.
Excess demand At a given price, the amount by which quantity demanded exceeds quantity supplied.
Excess supply At a given price, the amount by which quantity supplied exceeds quantity demanded.
Exchange The act of trading with others to obtain what we desire.
Excise tax A tax on a specific good or service.
Excludability The ability to exclude those who do not pay for a good from consuming it.
Exit A permanent cessation of production when a firm leaves an industry.
Explicit collusion Cooperation involving direct communication between competing firms about setting prices.
Explicit cost The dollars sacrificed—and actually paid out—for a choice.
Exports Goods and services produced domestically, but sold abroad.
Externality A by-product of consuming or producing a good that affects someone other than the buyer or seller.

F

Factor markets Markets in which resources—labor, capital, land and natural resources, and entrepreneurship—are sold to firms.
Financial asset A promise to pay future income in some form, such as future profits or future interest payments.
Firm's supply curve A curve that shows the quantity of output a competitive firm will produce at different prices.
Fixed costs Costs of fixed inputs, which remain constant as output changes.
Fixed input An input whose quantity must remain constant, regardless of how much output is produced.

Flow variable A variable representing a process that takes place over some time period.
Free rider problem When the efficient outcome requires a side payment but some individuals will not contribute.

G

Game theory An approach to modeling the strategic interaction of oligopolists in terms of moves and countermoves.
Government franchise A government-granted right to be the sole seller of a product or service.

H

Human capital The skills and training of the labor force.

I

Imperfect competition A market structure in which there is more than one firm but one or more of the requirements of perfect competition is violated.
Imperfectly competitive market A market in which a single buyer or seller has the power to influence the price of the product.
Implicit cost The value of something sacrificed when no direct payment is made.
Imports Goods and services produced abroad, but consumed domestically.
Income The amount that a person or firm earns over a particular period.
Income effect As the price of a good decreases, the consumer's purchasing power increases, causing a change in quantity demanded for the good.
Income elasticity of demand The percentage change in quantity demanded caused by a 1 percent change in income.
Increasing cost industry An industry in which the long-run supply curve slopes upward because each firm's *LRATC* curve shifts upward as industry output increases.
Increasing marginal returns to labor The marginal product of labor increases as more labor is hired.
Inelastic demand A price elasticity of demand between 0 and 1.
Infant industry argument The argument that a new industry in which a country has a comparative advantage might need protection from foreign competition in order to flourish.
Inferior good A good that people demand less of as their income rises.
Input Anything (including a resource) used to produce a good or service.
Input-substitution effect A change in the wage rate alters the price of labor relative to the costs of other inputs, and therefore changes the quantity of labor demanded.
Investment Firms' purchases of new capital over some period of time.

L

Labor The time human beings spend producing goods and services.
Labor demand curve Curve indicating the total number of workers all firms in a labor market want to employ at each wage rate.
Labor supply curve A curve indicating the number of people who want jobs in a labor market at each wage rate.
Land The physical space on which production takes place, as well as the natural resources that come with it.
Law of demand As the price of a good increases, the quantity demanded decreases.
Law of diminishing (marginal) returns As more and more of any input is added to a fixed amount of other inputs, its marginal product will eventually decline.
Law of diminishing marginal utility As consumption of a good or service increases, marginal utility decreases.
Law of supply As the price of a good increases, the quantity supplied increases.
Long run A time horizon long enough for a firm to vary all of its inputs.
Long-run average total cost The cost per unit of producing each quantity of output in the long run, when all inputs are variable.
Long-run elasticity An elasticity measured a year or more after a price change.
Long-run supply curve A curve indicating price and quantity combinations in an industry after all long-run adjustments have taken place.
Long-run total cost The cost of producing each quantity of output when all inputs are variable and the least-cost input mix is chosen.
Loss The difference between total cost (*TC*) and total revenue (*TR*), when $TC > TR$.
Lumpy input An input whose quantity cannot be increased gradually as output increases, but must instead be adjusted in large jumps.

M

Macroeconomics The study of the behavior of the overall economy.
Marginal approach to profit A firm maximizes its profit by taking any action that adds more to its revenue than to its cost.
Marginal cost The increase in total cost from producing one more unit of output.
Marginal cost pricing Setting a monopoly's regulated price equal to marginal cost where the marginal cost curve crosses the market demand curve.
Marginal product of labor The additional output produced when one more worker is hired.
Marginal revenue The change in total revenue from producing one more unit of output.

Marginal social benefit (MSB) The full benefit provided by another unit of a good, including the benefit to the consumer *and* any benefits enjoyed by third parties.
Marginal social cost (MSC) The full cost of producing another unit of a good, including the marginal cost to the producer *and* any harm caused to third parties.
Marginal utility The change in total utility an individual obtains from consuming an additional unit of a good or service.
Market A group of buyers and sellers with the potential to trade with each other.
Market consumer surplus The total consumer surplus enjoyed by all consumers in a market.
Market economy An economic system in which resources are allocated through individual decision making.
Market failure A market that operates inefficiently without government intervention.
Market power The ability of a seller to raise price without losing all demand for the product being sold.
Market producer surplus The total producer surplus gained by all sellers in a market.
Market signals Price changes that cause changes in production to match changes in consumer demand.
Market structure The characteristics of a market that influence how trading takes place.
Market supply curve A curve indicating the quantity of output that all sellers in a market will produce at different prices in the short run.
Marketable public good An excludable and nonrival good. Generally provided by the market for a price, though efficiency would require a price of zero.
Maturity date The date at which a bond's principal amount will be paid to the bond's owner.
Microeconomics The study of the behavior of individual households, firms, and governments; the choices they make; and their interaction in specific markets.
Minimum efficient scale The lowest output level at which the firm's *LRATC* curve hits bottom.
Model An abstract representation of reality.
Monopolistic competition A market structure in which there are many firms selling products that are differentiated, and in which there is easy entry and exit.
Monopoly The only seller in a market, or a market with just one seller.
Moral hazard When someone is protected from paying the full costs of their harmful actions and acts irresponsibly, making the harmful consequences more likely.
Mortgage A loan given to a home-buyer for part of the purchase price of the home.

N

Nash equilibrium A situation in which every player of a game is taking the best action for themselves, given the actions taken by all other players.
Natural monopoly A monopoly that arises when, due to economies of scale, a single firm can produce for the entire market at lower cost per unit than could two or more firms.
Natural oligopoly A market that tends naturally toward oligopoly because the minimum efficient scale of the typical firm is a large fraction of the market.
Network externalities Additional benefits enjoyed by all users of a good or service because others use it as well.
Nonprice competition Any action a firm takes to shift its demand curve rightward.
Nonwage job characteristic Any aspect of a job—other than the wage—that matters to a potential or current employee.
Normal good A good that people demand more of as their income rises.
Normal profit Another name for zero economic profit.
Normative economics The practice of recommending policies to solve economic problems.

O

Oligopoly A market structure with a small number of strategically interacting firms.
Opportunity cost What is given up when taking an action or making a choice.
Output effect A change in the wage rate alters the profit-maximizing output level, and therefore changes the quantity of labor demanded.

P

Pareto improvement An action that makes at least one person better off, and harms no one.
Patent A temporary grant of monopoly rights over a new product or scientific discovery.
Payoff matrix A table showing the payoffs to each of two players for each pair of strategies they choose.
Perfect competition A market structure in which there are many buyers and sellers, the product is standardized, sellers can easily enter or exit the market, and buyers and sellers are well-informed.
Perfect price discrimination Charging each customer the most he or she would be willing to pay for each unit purchased.
Perfectly competitive labor market A market with many well-informed buyers and sellers of standardized labor, with no barriers to entry and exit.
Perfectly competitive market (informal definition) A market in which no buyer or seller has the power to influence the price.
Perfectly inelastic demand A price elasticity of demand equal to 0.
Perfectly (infinitely) elastic demand A price elasticity of demand approaching infinity.

Physical capital The part of the capital stock consisting of physical goods, such as machinery, equipment, and factories.
Plant The collection of fixed inputs at a firm's disposal.
Positive economics The study of how the economy works.
Present value The value, in today's dollars, of a sum of money to be received or paid at a specific date in the future with certainty.
Price The amount of money that must be paid to a seller to obtain a good or service.
Price ceiling A government-imposed maximum price in a market.
Price discrimination Charging different prices to different customers for reasons other than differences in cost.
Price elasticity of demand The sensitivity of quantity demanded to price; the percentage change in quantity demanded caused by a 1 percent change in price.
Price elasticity of supply The percentage change in quantity supplied of a good or service caused by a 1 percent change in its price.
Price floor A government-imposed minimum price in a market.
Price leadership A form of tacit collusion in which one firm sets a price that other firms copy.
Price setter A firm (with market power) that selects its price, rather than accepting the market price as a given.
Price taker A firm that treats the price of its product as given and beyond its control.
Primary market The market in which newly issued financial assets are sold for the first time.
Principal–agent problem When one party (the principal) hires another (the agent), who in turn can pursue goals that conflict with the principal's because of asymmetric information.
Principal (face value) The amount of money a bond promises to pay when it matures.
Principle of asset valuation The idea that the value of an asset is equal to the total present value of all the future benefits it generates.
Producer surplus The difference between what the seller actually gets for a unit of a good and the cost of providing it.
Product markets Markets in which firms sell goods and services to households.
Production possibilities frontier (PPF) A curve showing all combinations of two goods that can be produced with the resources and technology currently available.
Productively inefficient A situation in which more of at least one good can be produced without sacrificing the production of any other good.
Profit Total revenue minus total cost.
Protectionism The belief that a nation's industries should be protected from foreign competition.
Pure discount bond A bond that promises no payments except for the principal it pays at maturity.
Pure private good A good that is both rivalrous and excludable.
Pure public good A good that is both nonrival and nonexcludable.

Q

Quantity demanded The quantity of a good that all buyers in a market would choose to buy during a period of time, given their constraints.
Quantity supplied The specific amount of a good that all sellers in a market would choose to sell over some time period, given their constraints.
Quota A limit on the physical quantity of imports.

R

Rational preferences Preferences that satisfy two conditions: (1) Any two alternatives can be compared, and one is preferred or else the two are valued equally, and (2) the comparisons are logically consistent or transitive.
Relative price The price of one good relative to the price of another.
Rent controls Government-imposed maximum rents on apartments and homes.
Rent-seeking activity Any costly action a firm undertakes to establish or maintain its monopoly status.
Repeated play A situation in which strategically interdependent sellers compete over many time periods.
Resource markets Markets in which households that own resources sell them to firms.
Resources The labor, capital, land (including natural resources), and entrepreneurship that are used to produce goods and services.
Rivalry A situation in which one person's consumption of a unit of a good or service means that no one else can consume that unit.

S

Scarcity A situation in which the amount of something available is insufficient to satisfy the desire for it.
Secondary market The market in which previously issued financial assets are sold.
Share of stock A share of ownership in a corporation.
Short run A time horizon during which at least one of the firm's inputs cannot be varied.
Shortage An excess demand not eliminated by a rise in price, so that quantity demanded continues to exceed quantity supplied.
Short-run elasticity An elasticity measured just a short time after a price change.
Short side of the market The smaller of quantity supplied and quantity demanded at a particular price.

Shutdown price The price at which a firm is indifferent between producing and shutting down.
Shutdown rule In the short run, the firm should continue to produce if total revenue exceeds total variable costs; otherwise, it should shut down.
Simplifying assumption Any assumption that makes a model simpler without affecting any of its important conclusions.
Single-price monopoly A monopoly firm that is limited to charging the same price for each unit of output sold.
Socialism A type of economic system in which most resources are owned by the state.
Specialization A method of production in which each person concentrates on a limited number of activities.
Stock variable A variable representing a quantity at a moment in time.
Strategic trade policy Protectionist policies designed to capture social benefits, such as greater tax revenue, from having an industry in the domestic country.
Subsidy A government payment to buyers or sellers on each unit purchased or sold.
Substitutable input An input that can be used *instead of* a particular type of labor.
Substitute A good that can be used in place of some other good and that fulfills more or less the same purpose.
Substitution effect As the price of a good falls, the consumer substitutes that good in place of other goods whose prices have not changed.
Sunk cost A cost that has been paid or must be paid, regardless of any future action being considered.
Supply curve A graph of a supply schedule, showing the quantity of a good or service supplied at various prices, with all other variables held constant.
Supply schedule A list showing the quantities of a good or service that firms would choose to produce and sell at different prices, with all other variables held constant.
Surplus An excess supply not eliminated by a fall in price, so that quantity supplied continues to exceed quantity demanded.

T

Tacit collusion Any form of oligopolistic cooperation that does not involve an explicit agreement.
Tariff A tax on imports.
Tax incidence The division of a tax payment between buyers and sellers, determined by comparing the new (after tax) and old (pretax) market equilibriums.
Technology The methods available for combining inputs to produce a good or service.
Terms of trade The ratio at which a country can trade domestically produced products for foreign-produced products.
Tit-for-tat A game-theoretic strategy of doing to another player this period what he has done to you in the previous period.
Total benefits The sum of consumer and producer surplus in a particular market.
Total cost The costs of all inputs—fixed and variable.
Total fixed cost The cost of all inputs that are fixed in the short run.
Total product The maximum quantity of output that can be produced from a given combination of inputs.
Total revenue The total inflow of receipts from selling a given amount of output.
Total variable cost The cost of all variable inputs used in producing a particular level of output.
Tradable permit A license that allows a company to release a unit of pollution into the environment over some period of time.
Traditional economy An economy in which resources are allocated according to long-lived practices from the past.
Tragedy of the commons The problem of overuse when a good is rivalrous but nonexcludable.

U

Unit elastic demand A price elasticity of demand equal to 1.
Utility A quantitative measure of pleasure or satisfaction obtained from consuming goods and services.

V

Variable costs Costs of variable inputs, which change with output.
Variable input An input whose usage can change as the level of output changes.

W

Wealth The total value of everything a person or firm owns, at a point in time, minus the total amount owed.

Y

Yield The annual rate of return a bond earns for its owner.

INDEX

E

F

G

H

I

L

M

N

O

P

Q

R

S

T

U

V

W

Y