GLOBAL EDITION

Modern Industrial Organization

FOURTH EDITION

Dennis W. Carlton • Jeffrey M. Perloff

ALWAYS LEARNING

PEARSON

Modern Industrial Organization

Fourth Edition

Global Edition

Modern Industrial Organization

Fourth Edition

Global Edition

Dennis W. Carlton

University of Chicago

Jeffrey M. Perloff

University of California, Berkeley

PEARSON

Boston San Francisco New York Hoboken
London Toronto Sydney Tokyo Singapore Madrid
Mexico City Munich Paris Cape Town Hong Kong Montreal

To Janie and Jackie

Editor in Chief: Denise Clinton
Acquisitions Editor: Adrienne D'Ambrosio
Director of Development: Sylvia Mallory
Managing Editor: Jim Rigney
Senior Production Supervisor: Nancy Fenton
Senior Media Producer: Melissa Honig
Senior Acquisitions Editor, Global Edition: Steven Jackson
Assistant Project Editor, Global Edition: Amrita Kar
Manager, Media Production, Global Edition: Vikram Kumar
Senior Manufacturing Controller, Production,
Global Edition: Trudy Kimber
Marketing Manager: Deborah Meredith
Design Supervisor: Regina Kolenda
Interior Designer: Leslie Haimes
Cover Designer: Shree Mohanambal Inbakumar, Lumina Datamatics
Illustrator: Jim McLaughlin
Senior Prepress Supervisor: Caroline Fell
Senior Manufacturing Buyer: Hugh Crawford
Compositor: Cenveo Publisher Services
Cover Image: (c)somchai rakin/Shutterstock

Pearson Education Limited
Edinburgh Gate
Harlow
Essex CM20 2JE
England

and Associated Companies throughout the world

Visit us on the World Wide Web at:
www.pearsonglobaleditions.com

Authorized adaptation from the United States edition, entitled Modern Industrial Organization, 4th edition, *ISBN 978-0-321-18023-0, by Dennis W. Carlton and Jeffrey M. Perloff, published by Pearson Education © 2016.*

ISBN-10:1-292-08785-4
ISBN-13: 978-1-292-08785-6

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

10 9 8 7 6 5 4 3 2 1

Typeset by Courier Westford
Printed and bound by Courier Westford

Brief Contents

Contents

PART 1 *Introduction and Theory* *25*

PART 2 *Market Structures* *79*

CHAPTER 16 Patents and Technological Change

CHAPTER 17 How Markets Clear: Theory and Facts

PART 6 *Government Policies and Their Effects* 619

Preface

There's no IO without U.—Sesame Street

Modern Industrial Organization, Fourth Edition, combines the latest theories with empirical evidence about the organization of firms and industries. It goes beyond the descriptive traditional structure-conduct-performance approach by using the latest advances in microeconomic theory, including transaction-cost analysis, game theory, contestability, and information theory. Practical examples illustrate the role of each theory in current policy debates, such as whether mergers promote economic efficiency (Chapter 2), whether predatory pricing is likely to be a serious problem (Chapter 11), whether preventing manufacturers from restricting distributors' prices benefits consumers (Chapter 12), whether providing consumers with more information about prices or products increases welfare (Chapter 13), whether advertising is harmful (Chapter 14), whether joint ventures are the best means of encouraging research (Chapter 16), whether current antitrust laws promote competition and increase welfare (Chapter 19), and whether government regulation does more harm than good (Chapter 20).

Modern Industrial Organization is designed for use by both undergraduate and graduate students. The theories presented in the chapters require only a microeconomics course as a prerequisite and do not involve calculus. Technical appendixes supplement selected chapters and provide a rigorous foundation for graduate students. Starred sections are relatively difficult and may be skipped.

We have used this book in both undergraduate and graduate courses. In our undergraduate courses, we rely on the chapters and skip the technical appendixes. In graduate courses, we use the chapters and technical appendixes along with supplementary readings based on selected articles that are discussed within chapters or recommended at the end of chapters.

Structure of the Book

The first half of the book covers the basics of competition, monopoly, oligopoly, and monopolistic competition. Chapter 1 discusses the basic approach used in the book. Chapter 2 discusses the reasons why firms exist, merger activity, and costs. Chapters 3 and 4 develop the basics of microeconomic theory—costs, competition, monopoly, barriers to entry, and externalities—that we use throughout the rest of the book. Variations on the standard models (such as a dominant firm facing a competitive fringe) are also presented.

Chapters 5 through 7 explain the recent developments in the theory of oligopoly and monopolistic competition. Chapter 5 covers cooperative oligopoly behavior (cartels), and Chapter 6 examines both cooperative and noncooperative behavior based on game theory. Chapter 7 focuses on monopolistic competition and product differentiation. Chapter 8 concludes the first part of the book with a thorough review and assessment of empirical work on market structure.

The remainder of the book covers the "new industrial organization"—material that is often missing from traditional texts. These topics, essential for applying the theories of industrial organization to everyday problems, are at the heart of many public policy debates and are the focus of considerable recent research. Chapters 9 and 10 cover common pricing strategies such as price discrimination through quantity discounts and tie-in sales. Chapter 11 examines strategic behavior where firms determine the best ways to do battle with their rivals. Chapter 12 discusses common business practices between manufacturers and distributors (vertical integration and vertical restrictions) and the dramatic changes in public policy toward these practices in recent years. The next two chapters, Chapters 13 and 14, address the problems that arise when consumers are not perfectly informed and when firms must advertise their products. The role of time is introduced in Chapters 15 and 16, which analyze how the durability of a product affects the market and how innovation can be encouraged. Chapter 17 considers evidence on the ways markets operate, and explores how modern microeconomic models of industrial organization may affect the macroeconomic economy. Chapter 18 examines the industrial organization issues that arise in international trade. The two concluding chapters, Chapters 19 and 20, analyze antitrust policy and government regulation.[1]

Although we believe that *Modern Industrial Organization* contains innovative ideas, we recognize that any textbook must borrow from existing research. We have tried to indicate when we have relied on the insights of others. However, we may have occasionally omitted a reference to an author whose ideas predated ours. We apologize for any such oversights.

Changes in the Fourth Edition

There are three major changes in the Fourth Edition. First, we have added many new applications, as well as discussions of important recent policies and new theories. Much of this new material is based on significant findings from more than 250 relevant articles and books published since our last edition. We have substantially updated material on cartels, particularly international cartels, and antitrust activities (Chapter 5); we have included a new section on estimation issues concerning differentiated

[1] Sometimes commonly used words have special meanings in the law that differ from the standard usage by economists and the general public. We try to use clear language to express economic rather than legal principles. For example, we might say that the "price of wheat in the market in Chicago affects the price of wheat in the market in Kansas City." Although such a statement uses the word *market* loosely, the point of the statement—that the prices of wheat in Chicago and Kansas City are related—is clear. In an antitrust trial, however, a specific legal definition of a market (see Chapter 19) is used and whether there are two separate markets or a single combined market is often of central interest. Our statement should not be interpreted to mean that there are necessarily two distinct wheat markets in Chicago and Kansas City for legal purposes.

goods oligopolies (Chapter 7); we have added a major new section on Sutton's modern approach to structure-conduct-performance analysis (Chapter 8); and we have substantially updated our discussion of patents and copyrights (Chapter 16) and regulation (Chapter 20).

Second, we have updated 18 examples and added 51 new examples. For instance, in one updated application, we conducted a new study of how the prices of Coke and Tropicana orange juice vary across grocery stores within a city. Our new examples spotlight a range of current events, among them the Enron scandal, the importation of low-price drugs from Canada, genetically modified organisms, the effect of 9/11 on flag sales, Blockbuster's innovative pricing polices, mergers in Europe, a monopsony in hiring priests, the change of China's tobacco monopoly to dominant firm status, the international vitamins cartel, the value of minivans, the certification of thoroughbreds, counterfeit Halal meat, Napster and piracy issues, and many others.

Third, we have significantly augmented our Web site, **www.aw-bc.com/carlton_perloff,** with extensive supporting material. Still-timely material that we removed from the Third Edition is available on the Web site. Further, we have written many new applications for the site.

Alternative Course Outlines

To cover the entire book takes two quarters or semesters. The book is designed, however, so that shorter courses can be constructed easily by choosing selected chapters, as shown in the following proposed reading lists.

Chapter 2 through 4 review and extend the basic material that is often covered in an intermediate microeconomics course: the theory of the firm, costs, the theory of competition, the theory of monopoly, and externalities. These chapters can be reviewed quickly for students with extensive preparation in microeconomics. Chapters 2 through 8 comprise the basic material for any course. Depending on the interests of the students and the instructor, a one-quarter or semester course could then sample a few of the chapters in the remainder of the book to obtain a flavor of the ways industrial organization can be used to study real-world problems.

All courses:
Carefully cover the core material in Chapters 2 and 5–8.

For courses that do not assume a strong background in microeconomic theory:
Cover Chapters 3 and 4.

Courses that assume a strong background in microeconomic theory:
Quickly review Chapters 3 and 4.

Courses that require calculus:
Include the technical appendixes and material on the Web.

Policy-oriented courses:
Cover international trade, antitrust, and regulation (Chapters 18 through 20). As time allows, include strategic behavior (Chapter 11), price discrimination (Chapters 9

and 10), vertical relationships (Chapter 12), limited information, advertising, and disclosure (Chapters 13 and 14), government policies toward innovation (Chapter 16), and macroeconomics (Chapter 17).

Regulation courses:
Regulations are dealt with throughout the book. Cover, in particular, externalities (Chapters 3 and 4), vertical relations (Chapter 12), limited information (Chapter 13), advertising and disclosure (Chapter 14), government policies toward innovation (Chapter 16), international trade (Chapter 18), and other government regulation (Chapter 20).

Business courses:
Include strategic behavior (Chapter 11), price discrimination (Chapter 9 and, optionally, nonlinear pricing, Chapter 10), vertical relations (Chapter 12), information and advertising (Chapters 13 and 14), and international trade (Chapter 18).

Courses that stress the latest theories:
Include strategic behavior (Chapter 11), vertical relations (Chapter 12), information and advertising (Chapters 13 and 14), government policies toward innovation (Chapter 16), market operation (Chapter 17), and international trade (Chapter 18).

Advanced courses:
Add chapters on nonlinear pricing (Chapter 10) and durability (Chapter 15).

Acknowledgments

We are especially grateful to George Stigler, our departed colleague, for his encouragement and insightful comments. His contributions to industrial organization influenced this and every other text in the field. We also thank the many people who helped us in the planning, writing, revising, and editing of this book:

Donald L. Alexander, *Western Michigan University*
Mark Bagnoli, *Michigan State University*
Kyle Bagwell, *Northwestern University*
Gustavo Bamberger, *Lexecon, Inc.*
Francis Bloch, *Brown University*
Giacomo Bonanno, *University of California, Davis*
Ralph Bradburd, *Williams College*
Reuven Brenner, *McGill University*
Timothy Bresnahan, *Stanford University*
Jeremy Bulow, *Stanford University*
David Butz, *University of California, Los Angeles*
Catherine Carey, *Western Kentucky University*
Kathleen Carroll, *University of Maryland, Baltimore County*
Phillip P. Caruso, *Western Michigan University*

Richard Clarke, *AT&T Bell Laboratories*
Charles Cole, *California State University, Long Beach*
John Connor, *Purdue University*
Ron Cotterill, *University of Connecticut*
Keith Crocker, *Pennsylvania State University*
Anna P. Della Valle, *New York University*
Craig A. Depken II, *University of Texas, Arlington*
Frank Easterbook, *University of Chicago,* and *Judge, Federal Court of Appeals*
Nicholas Economides, *New York University*
Gregory Ellis, *University of Washington*
Robert Feinberg, *American University*
Daniel Fischel, *University of Chicago*
Trey Fleisher, *Metropolitan State College of Denver*
Alan Frankel, *LECG*
Drew Fudenberg, *Harvard University*
Anita Garten, *A. Garten Consulting*
Robert Gertner, *University of Chicago*
Richard Gilbert, *University of California, Berkeley*
J. Mark Gidley, *White and Case*
Luis Guash, *University of California, San Diego*
Timothy Guimond, *Lexecon, Inc.*
Jonathan Hamilton, *University of Florida*
Mehdi Haririan, *Bloomsburg University*
Gloria Helfand, *University of Michigan*
James Holcolm, *University of Texas, El Paso*
Charles Holt, *University of Virginia*
Jorge Ibarra-Salazar, *ITESM*
Adam Jaffe, *Brandeis University*
Harvey James, *University of Missouri*
Larry Karp, *University of California, Berkeley*
Theodore Keeler, *University of California, Berkeley*
Alvin Klevorick, *Yale University*
William Kolasky, *Wilmer, Cutler and Pickering*
Dan Kovenock, *Purdue University*
John Kwoka, *George Washington University*
William Landes, *University of Chicago*
Richard Langlois, *University of Connecticut*
Jim Lee, *Fort Hays State University*
Bart Lipman, *Carnegie-Mellon University*

Nancy Lutz, *Yale University*

William Lynk, *Lexecon, Inc.*

Frank Mathewson, *University of Toronto*

Rachel McCulloch, *Brandeis University*

James Meehan, *Colby College*

John Menge, *Dartmouth College*

Robert Michaels, *California State University, Fullerton*

Richard A. Miller, *Wesleyan University, Connecticut*

David E. Mills, *University of Virginia*

Herbert Mohring, *University of Minnesota*

Janet Netz, *Purdue University*

Gregory Pelnar, *Lexecon, Inc.*

Marty Perry, *Rutgers University*

Nicola Persico, *University of Pennsylvania*

Russell Pittman, *Justice Department*

Richard Posner, *University of Chicago,* and *Judge, Federal Court of Appeals*

Stanley Reynolds, *University of Arizona*

Richard Rogers, *University of Massachusetts*

Andrew Rosenfield, *Lexecon, Inc.*

Thomas Ross, *University of British Columbia*

Charles K. Rowley, *George Mason University*

Stephen Salant, *University of Michigan*

Garth Saloner, *Stanford University*

Steven Salop, *Georgetown University*

Richard Schmalensee, *Massachusetts Institute of Technology*

Suzanne Scotchmer, *University of California, Berkeley*

Robert Sherwin, *Analysis Group*

Steven Sklivas, *Columbia University*

Edward Snyder, *University of Michigan*

Pablo Spiller, *University of California, Berkeley*

Mark Stegman, *University of North Carolina*

George Stigler, *University of Chicago*

Stephen Stigler, *University of Chicago*

Joseph Stiglitz, *Columbia University*

Dmitry Stolyarov, *University of Michigan*

Valerie Suslow, *University of Michigan*

Ming-Je Tang, *University of Illinois*

Mihkel M. Tombak, *Helsinki School of Economics*

Lien Tran, *Federal Trade Commission*

W. van Hulst, *Tilburg University*
Frank van Tongeren, *Erasmus University of Rotterdam*
Klaas van't Veld, *University of Michigan*
John Vernon, *Duke University*
Rickard Wall, *Linköping University*
Roger Ware, *Queen's University*
Avi Weiss, *Bar-Ilan University*
Leonard Weiss, *University of Wisconsin*
Gregory Werden, *Department of Justice*
Douglas West, *University of Alberta*
Lawrence White, *New York University*
Oliver Williamson, *University of California, Berkeley*
Robert Willig, *Princeton University*
Asher Wolinsky, *Northwestern University*
Brian Wright, *University of California, Berkeley*
Edwin Zimmerman, *Covington & Burling*

We are grateful to Keith Crocker, Stan Reynolds, and particularly Gregory Pelnar, who have made invaluable contributions to several editions. We thank David Buschena, Gary Casterline, Hayley Chouinard, Laona Fleischer, George Frisvold, Carolyn Harper, Colleen Loughlin, David Mitchell, Margaret Sheridan, and Deborah Zimmermann for excellent research assistance. We gratefully acknowledge the typing and other assistance of Julie Rodriguez and Hazel Young.

We thank our four excellent acquisitions editors: George Lobell (who provided extensive help on the First Edition), John Greenman (who arranged for much of the supplemental material for the Second Edition and the early stages of the Third Edition), Denise Clinton (who was very supportive on our Third Edition), and Adrienne D'Ambrosio (who helped to create the Fourth Edition). We are particularly grateful to Director of Development Sylvia Mallory, who has been extremely helpful as a development editor and in arranging for the supplements (with Diana Theriault) and creating a Web site (with Melissa Honig). Nancy Fenton and Julie DeSilva ably handled production and art coordination. Cynthia Benn and Robin MacFarlane provided careful copy editing and proofreading. Jim McLaughlin carefully drew all the two-color figures. Regina Kolenda directed the design of both the text and the cover, and Nesbitt Graphics executed the paging, resulting in the most attractive edition yet.

This book benefited extensively from the comments of our many students who cheerfully served as "guinea pigs," reading and using earlier editions and drafts of this version. Most importantly, we thank our families for their support.

Each author blames the other for any mistakes. Each takes credit for any good jokes.

Dennis W. Carlton
Jeffrey M. Perloff

Pearson would like to thank and acknowledge **Srishti Chakrabarti** (**Ansrsource.com**) for her contribution to the Global Edition, and **Gordon Lenjosek** for reviewing the Global Edition.

PART ONE

Introduction and Theory

CHAPTER 1

Overview

Leave all hope, ye that enter. —*Dante Alighieri*

This text presents both traditional and new theories of **industrial organization**: the study of the structure of firms and markets and of their interactions. Introductory microeconomics analyzes idealized models of firms and markets; this text takes a closer, more realistic look at them, warts and all.[1] In introductory physics, one first disregards gravity and friction in studying the movement of bodies, and then adds these complications to the analysis. The study of industrial organization adds to the perfectly competitive model real-world frictions such as limited information, transaction costs, costs of adjusting prices, government actions, and barriers to entry by new firms into a market. It then considers how firms are organized and how they compete in such a world. This chapter describes some of the approaches that help to organize the study of industrial organization and gives an overview of the material in later chapters. Finally, it describes some of the analytic tools that are used.

Models

There are at least two major approaches to the study of industrial organization, and, because they are compatible as organizing principles, this text uses both of them. The first approach, *structure-conduct-performance,* is primarily descriptive and provides an overview of industrial organization. The second,

[1] We use the terms *market* and *industry* loosely and interchangeably. In antitrust cases, important distinctions are made between these terms, as is discussed in later chapters.

price theory, uses microeconomic models to explain firm behavior and market structure.

According to the structure-conduct-performance approach, an industry's **performance** (the success of an industry in producing benefits for consumers) depends on the **conduct** (behavior) of its firms, which, in turn, depends on the **structure** (factors that determine the competitiveness of the market).[2] The structure of an industry depends on basic conditions, such as technology and demand for a product. For example, in an industry with a technology such that the average cost of production falls as output increases, the industry tends to have only one firm, or possibly a small number of firms. If only one firm (a monopoly) sells output in an industry, it may be able to set a price that is well above its marginal costs of production. If the basic conditions make the demand for the monopoly's product relatively inelastic (people are relatively insensitive to price), then the price in that market is higher than if the demand is relatively elastic (people are price sensitive).

Figure 1.1 illustrates the relationships among structure, conduct, and performance and shows how basic conditions and government policy interact. The relationships among the five boxes are complex. For example, government regulations affect the number of sellers in an industry, and firms may influence government policy to achieve higher profits. Similarly, if entry barriers lead to monopoly and monopoly profits, new industries may develop new, substitute products that affect the demand for the original product. Empirical researchers who rely on this paradigm typically use data at the industry level. They ask, for example, if industries with certain structural features (for example, few firms) have high prices.

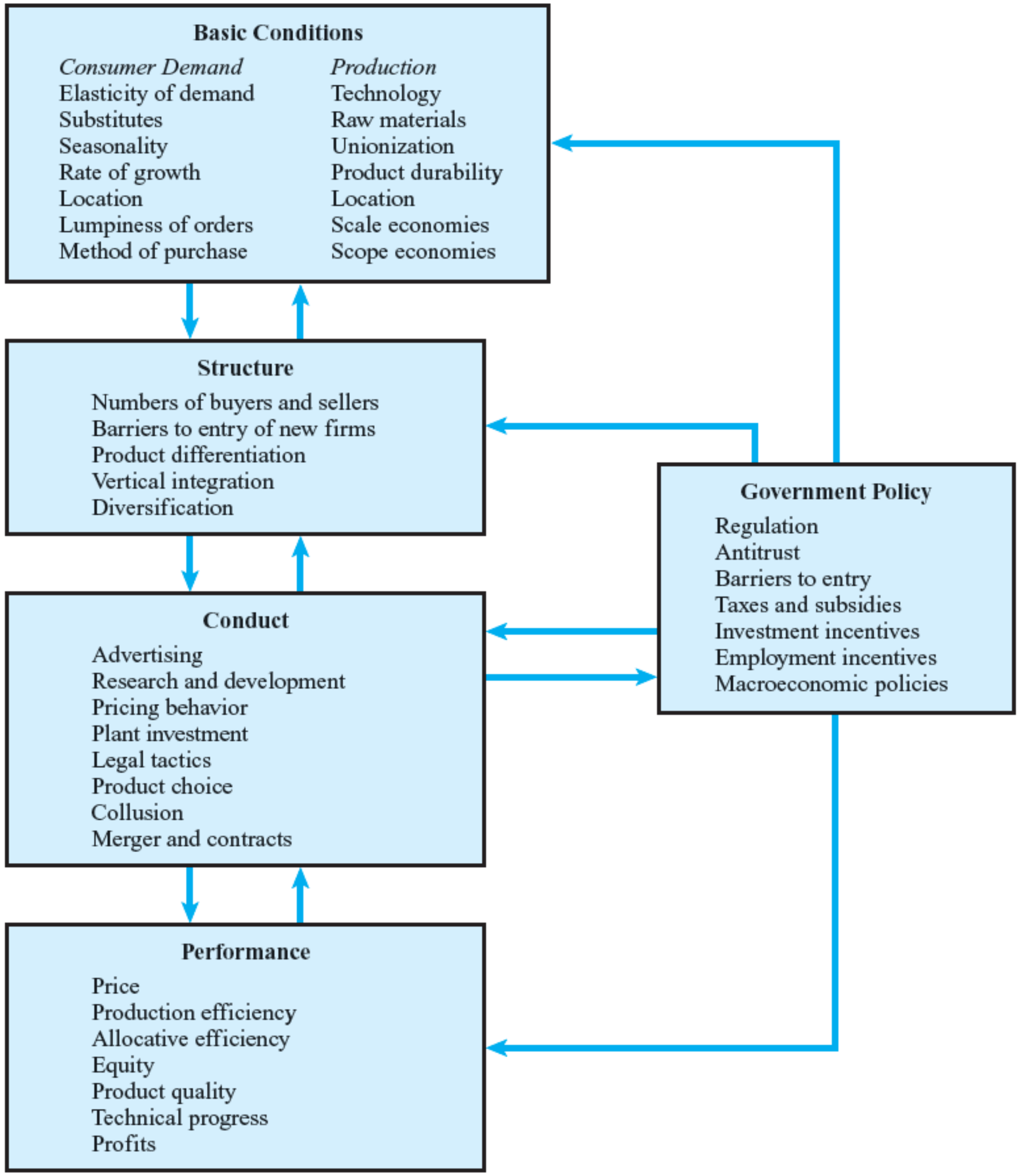
The structure-conduct-performance approach is a very general way to organize the study of industrial organization, and can be used to organize the material in the rest of this book. The second major approach, the price theory paradigm, can also be used to organize and interpret this material.

Price Theory

Price theory models analyze the economic incentives facing individuals and firms to explain market phenomena. George J. Stigler (1968), an early proponent of this analytical approach, believed that industrial organization researchers should use microeconomic theory to design empirical studies of markets and of the effects of public policy. Today, most industrial organization research and courses are well grounded in microeconomic theory. Two reasons for the shift to this approach are the recent availability of data at a more micro level and advances in price theory. In recent years, three specific theoretical applications of price theory have won substantial support—transaction cost analysis, game theory, and contestable market analysis—and help to explain structure, conduct, and performance.

[2] The structure-conduct-performance approach was developed at Harvard by Edward S. Mason (1939, 1949) and his colleagues and students, such as Joe S. Bain (1959).

FIGURE 1.1 Structure, Conduct, and Performance



Transaction Costs

Transaction costs are the expenses of trading with others above and beyond the price, such as the cost of writing and enforcing contracts. Using formal price theory analysis, the transaction cost approach uses differences in transaction costs to explain why structure, conduct, and performance vary across industries.

Over 60 years ago, Ronald H. Coase (1937) explained that a firm and a market are alternative means of organizing economic activity. Coase emphasized that the use of the marketplace involves costs. These costs help to determine market structure. For example, where the cost of buying from other firms is relatively low, a firm is more likely to buy supplies from others than produce the supplies itself.

Oliver Williamson (1975, 8–10), one of the major proponents of the transaction cost approach, says that four basic concepts underlie this analysis:

1. Markets and firms are alternative means for completing related sets of transactions. For example, a firm can either buy a product or a service or produce it.
2. The relative cost of using markets or a firm's own resources should determine the choice.
3. The transaction costs of writing and executing complex contracts across a market "vary with the characteristics of the human decision makers who are involved with the transaction on the one hand, and the objective properties of the market on the other" (p. 32).
4. These human and environmental factors affect the transaction costs across markets and within firms.

This approach aims to identify a set of environmental and human factors that explain both the internal organization of firms and organization of industries. The key environmental factors are *uncertainty* and the *number of firms*; the key human factors are *bounded rationality* and *opportunism*. **Bounded rationality** is the limited human capacity to anticipate or solve complex problems. Problems arise when uncertainty is combined with bounded rationality, or where the managers of the few firms in an industry behave opportunistically (take advantage of a situation).

Thus, in a world of great uncertainty, it may be too difficult or costly to negotiate contracts that deal with all possible contingencies. As a result, firms may produce internally even though, otherwise, it would be cost-effective to rely on markets.

When the number of firms is small and individuals are opportunistic, firms may not want long-term contracts for fear of being victimized in the future. For example, a firm that relies on another to supply a factor that is essential to its production process may be exploited because it cannot operate if its supply is stopped. This problem is likely to be important if there are few alternative suppliers.

Thus, reliance on markets is more likely when (1) there is little uncertainty and (2) there are many firms (competition) and limited opportunities for opportunistic behavior. When these conditions are reversed, firms are more likely to produce for themselves than to rely on markets. The transaction cost approach has been very successful because of its broad explanatory power.

Game Theory

Another approach that is increasingly important to economic theorists is game theory (von Neumann and Morgenstern 1944), which uses formal models to analyze conflict and cooperation between firms and individuals. Competition among firms is viewed as a game of strategies, or battle plans of the actions of a firm, that describe the behavior of each firm. A firm's strategy determines, for example, its output, price, and advertising level. In the game, firms compete for profits. Game theory describes how firms form their strategies and how these strategies determine the profits.

Game theory provides insights in games in which there are relatively few firms. Much of this text concerns such markets, and many of the models it presents are examples of game theory.

Contestable Markets

The importance of entry to the competitive process has been recognized for a long time. Demsetz (1968) and Baumol, Panzar, and Willig (1982) emphasize that industries with only a few firms (or just one) can be very competitive if there is a threat of entry by other firms. Markets in which many firms can enter rapidly if prices exceed costs and can exit rapidly if prices drop below costs are called contestable. As Baumol, Panzar, and Willig explain, firms are reluctant to enter an industry if it is very costly to exit.

With few firms but easy entry and exit, the market is contestable and can have the properties of a competitive market: Price equals marginal cost and strategic behavior is irrelevant. There are few known examples of such markets. If there are few firms in an industry and entry or exit is difficult, the market is not contestable and the strategic behavior studied by game theorists is relevant.

Organization

Where I am not understood, it shall be concluded that something very useful and profound is couched underneath. —*Jonathan Swift*

The main objective of this text is to provide a systematic presentation of the basic theories—both traditional and new—of how firms and markets are organized and how they behave. Rather than treating structure and conduct as given, the text explains them as the outcome of individuals' maximizing behaviors. That is, it shows how the price theory models provide the underpinnings for the structure-conduct-performance paradigm. The paradigms complement each other, and both are useful for developing an understanding of industrial organization.

Basic Theory

Chapter 2 reviews basic microeconomic theories about costs and introduces the theory of the firm. The chapter starts by covering the internal organization and ownership of firms, pointing out that the division between firms and markets is not always clear and that the structure of an industry may change rapidly as costs shift. It examines the role of mergers and acquisitions in achieving efficiency in production.

The chapter then turns to costs. Special attention is paid to costs because they are crucial in explaining market structure. For example, costs (especially transaction costs) are crucial in determining whether it pays for a firm to produce an input or buy it in a market. The chapter reviews several cost concepts and discusses empirical evidence on costs.

Market Structures

The theory and empirical evidence on basic market structures are covered in Chapters 3 through 8. Table 1.1 shows the basic taxonomy used in this text to describe several structures. The number of firms in a market and the ease of entry and exit by new firms determine the type of structure.

When a market has many potential buyers and sellers and has no entry or exit barriers, the market structure is that of **competition**. When one firm sells to many buyers, and no new sellers can enter, the firm is a **monopoly**. Conversely, the only firm that buys from many sellers is a **monopsony**. If sellers can influence price even though they face competition from other firms, the market structure is either oligopolistic or monopolistic competition. An **oligopoly** is a small group of firms in a market with substantial barriers that prevent new sellers from entering the market. If there are no substantial barriers to entry and exit and each firm has some control over the price of its product, then the market is one of **monopolistic competition**: Firms can set price above the competitive level but earn zero profit.

The market structure often depends on the presence or absence of barriers to entry and exit, which are discussed in Chapter 3. For example, a new airline company cannot offer service between New York and Tokyo without permission from both the

TABLE 1.1 Some Basic Market Structures

	Sellers		Buyers	
Market Structure	Entry Barriers	Number	Entry Barriers	Number
Competition	no	many	no	many
Monopoly	yes	one	no	many
Monopsony	no	many	yes	one
Oligopoly	yes	few	no	many
Oligopsony	no	many	yes	few
Monopolistic competition	no	many	no	many

United States and Japan. Such permission is usually not granted unless a company currently flying ceases operation; thus, a government-created entry barrier exists in this market.

Chapters 3 and 4 review and extend the theory of competition and monopoly. Chapter 3 discusses the basic theory of competition. Competitive firms are too small to affect the market price, so they take that price as given (the firms are said to be **price takers**) and choose how many units of output to produce. The chapter shows that such behavior has desirable consequences for social welfare. It is the market structure to which all other structures are compared. Because there are no barriers to entry, firms enter competitive markets whenever positive profits can be made. This influx of sellers drives profits to zero for all firms in the market in the long run.

In contrast, as the only firm in the market, a monopoly (Chapter 4) is a **price setter**: It determines the price of its good, and typically sets it above the competitive level. The ability to price profitably above the competitive level is referred to as **market power**, and such conduct leads to welfare losses by society. Because of entry barriers, the monopoly can earn positive economic profits in the long run. Analogously, monopsony results in a lower price than a competitive market would set, which also has undesirable welfare implications.

Chapter 4 introduces another structure, which is not described in Table 1.1. It is a hybrid of the competitive and monopolistic structures, in which there is a *dominant firm* and a *competitive fringe*. The dominant firm has some market power so that it can set prices, and the other (fringe) firms are price takers. For example, such a structure is observed where a monopoly in one country competes in world markets with a higher-cost competitive industry located in another country.

Chapter 5 shows how monopoly-like conduct may occur in a market with more than one firm. The firms may form a **cartel**: an association of firms that explicitly agree to coordinate their activities, typically to maximize joint profits. That is, the separate firms imitate the behavior of a monopoly. If they all restrict output and raise the industry price above the competitive level, they can increase their profits. Government antitrust laws may be used to prevent explicit cartels from forming. The chapter considers why cartels form in only some industries and why they fall apart. Members of cartels are shown to have an incentive to cheat on one another. This chapter shows how cartel theory provides an explanation of oligopoly behavior in the absence of explicit agreements.

Chapter 6 continues our study of oligopolies. Unlike competitive and monopolistic firms, oligopolistic firms expect their rivals to react to their behavior or strategies. Using game theory, we consider when oligopolies compete vigorously and when they do not. The chapter presents some experimental evidence on oligopolistic behavior.

Chapter 7 studies monopolistic competition by modifying the oligopoly model of Chapter 6 in two ways. First it allows entry. In monopolistic competition, unlike in an oligopoly, entry by new firms drives economic profits to zero. Thus, other things being equal, removing entry barriers typically increases output.

Second, Chapter 7 considers the implications of product differentiation on social welfare and the effect of government interventions in these markets. For example, consumers presumably prefer low prices and many choices of differentiated products. Thus, government intervention that results in fewer firms and products but lower average prices may be a mixed blessing. Whether consumers prefer slightly higher prices with more variety becomes an empirical question for each market.

Chapter 8 surveys the available empirical evidence on performance and market structure in the United States and other economies. Tests of the market structure theories discussed in Chapters 3 through 7 are examined. Both traditional and modern empirical approaches to assessing performance are presented.

Business Practices: Strategies and Conduct

Chapters 9 through 12 cover general business practices using some of the latest research in game theory and transaction cost theory. In the basic market structures, covered in the earlier chapters, firms concentrate on only a few strategies: Firms vary only price, output levels, or the degree of differentiation of their products, usually on a once-only basis.

Chapters 9 and 10 concentrate on complex pricing behavior. Chapter 9 covers **price discrimination**: A firm charges different categories of customers different unit prices for the identical good. Firms with market power can increase their profits by charging some consumers who are less price sensitive a higher price than others for identical products. Chapter 10 deals with other pricing schemes that are related to price discrimination. For example, an electrical utility may charge one price to be connected to the system and another for each kilowatt consumed. Similarly, a firm may sell you one of its products only if you agree to buy another.

Chapter 11 considers sophisticated competitive strategies in dynamic game theory models. For example, a firm may set such a low price that it drives its competitors out of business and then raises its price. Similarly, a firm may engage in behavior designed to raise its rivals' costs, so they cannot compete as effectively. Other more complex strategies involve exchanging (or not exchanging) information with competitors.

Chapter 12 examines the reasons for vertical integration. When a firm produces an input itself, the firm is said to be *vertically integrated*. Costs help to determine whether the firm vertically integrates or not. The chapter discusses why some industries buy inputs and others produce them. It also examines the welfare implications of vertical integration.

Chapter 12 then discusses why some firms, instead of vertically integrating, use *vertical restraints*. For example, an automobile manufacturer may require that its dealers, which are independent firms, agree in contracts about the way they will conduct their business. Thus, the manufacturer uses contractual restrictions to approximate vertical integration. The recent change in public policy toward vertical restraints is discussed.

Information, Advertising, and Disclosure

Chapters 13 and 14 examine the effects of limited information on markets and how strategic behavior by firms can alter information. Chapter 13 discusses the effect of information on quality and prices in a market and shows that many typical properties of a competitive market disappear if information is limited. Limits on consumer information often give firms market power; thus, better information may reduce market power and increase competition.

Chapter 14 examines advertising and how it may either increase or decrease welfare. The chapter also explains how laws designed to limit lying or to require disclosure of important facts to consumers may have paradoxical effects.

Dynamic Models and Market Clearing

Except for Chapter 11's discussion of multiperiod strategies, the models discussed prior to Chapter 15 use a **static analysis**: models of markets that last for only one period. Like snapshots, static models tell us what happens at a point in time. Typically, static models are used for long-term analysis. In contrast, multiperiod or dynamic models describe the evolution of markets and firm behavior over time. Although such models are more difficult to use than static ones, they provide additional insights.

Chapters 15, 16, and 17 use models in which current actions affect future profits. Chapter 15 examines firms' decision making in markets for durable goods. For example, would a car that lasts 15 years produce higher or lower profits for the manufacturer than one that lasts 10 years? One surprising result of this investigation is that a durable goods monopoly may have more market power if it rents its product than if it sells it.

Chapter 16 considers how government behavior affects technological change. New discoveries that reduce production costs or create new products are obviously highly desirable. Unfortunately, a competitive industry produces too few inventions because inventors do not capture the full value of their discoveries. To encourage greater inventive activity, governments provide many incentives. For example, governments grant patents that allow inventors to be monopoly sellers of new products.

Chapter 17 is the only chapter to deal explicitly with macroeconomic issues. However, as in the other chapters, the focus is on price theory. This chapter examines how a market adjusts over time as a function of its structure. Other means of *clearing a market* (forcing quantity demanded to equal quantity supplied) besides price adjustments are also discussed.

Government Policies and Their Effects

Chapters 18, 19, and 20 analyze the effects of government actions that increase or decrease welfare. Chapter 18 examines how market structure and government actions affect international trade markets. Particular attention is paid to the effects of tariffs, subsidies, and quotas on the performance of markets.

Chapter 19 considers *antitrust* laws, which are intended to prevent conduct that adversely affects welfare, such as the formation of cartels or mergers that might lead to substantial market power. The chapter points out, however, that antitrust laws sometimes have been used to prevent rather than encourage competitive behavior.

Finally, Chapter 20 discusses how governments regulate business conduct and market structure. The chapter examines the effects of the recent trend toward deregulating markets. Unfortunately, regulation does not always benefit consumers or society. Government intervention in some markets leads to inefficiency, and many laws proposed with the noblest objectives benefit special interest groups at the expense of the general population.

CHAPTER 2

The Firm and Costs

Few have heard of Fra Luca Parioli, the inventor of double-entry bookkeeping; but he has probably had much more influence on human life than has Dante or Michelangelo. —Herbert J. Muller

A firm is an organization that transforms *inputs* (resources it purchases) into *outputs* (valued products that it sells). It earns the difference between what it receives as revenue and what it spends on inputs, which are used in manufacturing and selling. For example, a steel firm builds a plant, hires workers, purchases raw materials, and then produces and sells steel. The firm decides the quantity of resources to buy, how to combine the resources to make steel, and how and where to sell it. The firm makes a profit if it sells its steel for more than the cost of producing and selling the steel.

We start by discussing the objective, organization, and ownership of firms. Most firms try to maximize their profits. To maximize profit, a firm must produce its output at the least possible cost, given technology and the price of inputs.

Next we examine costs. Knowledge of costs is necessary to understand industrial organization for three reasons. First, many of the predictions of economic theory, such as those involving price and firms' size, revolve around concepts like marginal costs and profits. Without a knowledge of cost concepts, one cannot understand or empirically test these predictions. Second, theoretical work (Baumol, Panzar, and Willig 1982) emphasizes that oligopoly behavior depends crucially on certain types of fixed cost. Third, governments often regulate industries in which competitive entry leads to unusually high costs. Knowing how to regulate these industries requires an intimate familiarity with cost concepts (see Chapter 20).

This chapter introduces the concepts of marginal, average, and variable costs and then discusses some subtleties associated with economic costs. It analyzes the theory and evidence concerning economies of scale and concludes with a discussion of costs for a multiproduct firm.

The key issues we study in this chapter are:

1. Most firms maximize profits.
2. Acquisitions and other mergers may (but do not always) force firms to operate efficiently and profitably.
3. Economists use the concept of an opportunity cost that includes a normal profit.
4. The costs of a single-product firm depend on the prices of factors of production and the output level.
5. A multiproduct firm's cost of producing a single product depends on factor prices, the output level of that product, and the output level of its other products.
6. Production processes may have various properties such as economies of scale and economies of scope.

The Firm

Most goods and services produced in Western countries are produced by firms. In the United States, firms produce 84 percent of national production, the government produces 11 percent, nonprofit institutions (such as some universities and hospitals) produce 5 percent, and private households produce less than 0.1 percent.[1] In contrast, the government's share of total national production can be much higher in developing countries, reaching 43 percent in Ethiopia, 44 percent in Kyrgyzstan, 46 percent in Yemen, and 59 percent in Lesotho, though it is less than 5 percent in Guinea, Ireland, and Luxembourg (Heston et al., 2002). We now examine the objective, organization, and ownership of firms.

The Objective of a Firm

Most firms are *for-profit* firms: They exist to make money. Unless we state otherwise, when we refer to a firm we mean a for-profit firm and not a firm that exists for charitable or other nonprofit reasons.

The standard assumption in most economic models is that the primary objective of a manager of a firm is to maximize the firm's profits. The manager must sell the optimal amount of output, and the firm engages in efficient production: No more output could be produced with existing technology, given the quantity of inputs used.

[1] These are shares of the U.S. gross domestic product for 2002 from the U.S. Department of Commerce, Bureau of Economic Analysis, National Income and Products Accounts Table 1.7, Gross Domestic Product by Sector (**www.bea.doc.gov/bea/dn/nipaweb**). These figures exclude nonmarketed output of private households such as meal production.

Managers may have objectives other than profit maximization, however. For example, if managers want to control a large firm, they may maximize sales rather than profits. Similarly, managers may spend the firm's money on luxurious offices, company planes, and other amenities that reduce the profitability of the firm but benefit managers directly.

Various forces keep managers from deviating from profit-maximizing behavior. If a firm is run inefficiently and unprofitably, it may be driven out of business by rival firms that do maximize profits. Managers who lose their jobs when their firm is driven out of business or who are fired for inefficiency or laziness find it difficult to obtain new jobs. Incentives, such as stock ownership and other bonuses, also motivate managers to maximize profit. Thus, throughout most of this book, we assume that profit maximization is a reasonable approximation of a firm's objectives.

In Chapter 12, we examine how firms are organized to make them as efficient and profitable as possible, why failing to monitor causes problems, and what incentives firms provide employees to minimize these problems.[2]

Ownership and Control

Firms are owned and controlled in a variety of ways. A firm must raise money to finance itself, decide how its business is to be managed, and distribute its revenues to those who have contributed to its activity.

Forms of Ownership. The three basic business forms in the United States are sole proprietorships (single owner), partnerships (multiple owners), and corporations. Before the twentieth century, most firms were sole proprietorships or partnerships. Sole proprietors and partners are personally liable for the debts of their business. *All* the owners' assets, not just those invested in the business, are at risk. For example, a partner bears full personal liability for the debts of a failed business if the other partners have no assets, even if the business fails through no fault of the partner with the assets. Partnerships have a second problem as well. If one member of a partnership leaves, the entire partnership is automatically dissolved. To continue, the business must form a new partnership.

In the United States, 87 percent of business sales are made by corporations, even though only 20 percent of all firms are corporations. Nearly 72 percent of all firms are sole proprietorships. Sole proprietorships tend to be small, however, so they are responsible for only 5 percent of all sales. Partnerships are 8 percent of all firms and make 9 percent of sales.[3]

Corporations are companies whose capital is divided into shares that are held by individuals who have only limited responsibility for the debts of the company. That is, a shareholder has **limited liability**: If the corporation fails (is unable to pay its bills), the

[2] For classic works on these issues see March and Simon (1958), Cyert and March (1963), Marris (1964), and Williamson (1964). See **www.aw-bc.com/carlton_perloff** "How Firms Are Organized" for a discussion of these issues.

[3] Data for 1999 from the *Statistical Abstract of the United States*, Table 699, 2002:471.

stockholders need not pay for the debt using their personal assets. A shareholder's losses are limited to the price paid for the stock. With limited liability, individuals are more willing to buy shares than they would be if they could lose more than they paid to acquire the shares.

Today, most sales in the United States are made by corporations. Large corporations whose stock is publicly traded account for the bulk of economic activity and own a large percentage of all assets. According to the *1997 Census of Manufactures* (2001), out of 316,952 manufacturing firms, 246,189 (78 percent) are corporations. In manufacturing, corporations produce 95 percent of all the value added, account for 94 percent of all new capital expenditures, and hire 94 percent of all workers and 93 percent of all production workers. Individual proprietorships are 16 percent of all manufacturing firms but produce only 0.7 percent of the value added. Partnerships are about 4 percent of all manufacturing firms and produce 1.6 percent of the value added.

The importance of corporations has risen over time. In 1947, they comprised only 49 percent of all manufacturing firms (compared to 78 percent in 1997) and produced 92 percent of the value added (compared to 95 percent in 1997).

The rise of the corporation coincided with the need to increase the size of firms (see Example 2.1). The money needed to finance large enterprises could be efficiently raised only through the corporate form of organization. Otherwise, investors were not willing to accept the potential liabilities arising from the actions of managers whom they neither knew nor had the ability to monitor. The increase in the importance of the corporation and the coincident rise in stock trading is a relatively recent phenomenon of the last 100 years. In 1900, only 113 companies were listed on the New York Stock Exchange; in 1920, there were 391; today, over 1900 companies are listed.[4]

A corporation may raise money by selling shares of stock. Its shareholders elect a board of directors to run the corporation. In practice, the board of directors of a large corporation rarely becomes involved in day-to-day affairs; it delegates that responsibility to officers of the company. In large corporations, after the stock is issued, the stock is typically traded publicly (for example, IBM stock is traded on the New York Stock Exchange) and is not necessarily concentrated in the hands of a few key employees. Once stock is issued, the corporation receives nothing when individuals buy or sell the shares on a stock market.

Shareholders (also called *equity owners* because they own rights to the capital or equity of the firm) are entitled to receive dividend payments, which come out of the corporation's profits. Dividends are one way stockholders earn returns on their investments, but even if a corporation pays no dividends, shareholders can earn returns. If the price of the stock rises above what the shareholder paid, the shareholder can sell it for a profit.

Corporations also raise money by issuing debt. They promise to pay those who lend them money (*debt holders*) a stipulated amount of interest plus repayment of the loan. For example, General Electric might sell a *note* for $1 million in which it promises to pay 10

[4] *Wall Street Journal,* December 24, 1900, 1920, and telephone communications with the New York Stock Exchange. The number of companies is calculated as the number listed in the table entitled "New York Stock Exchange Composite Transactions," which appears daily in the *Wall Street Journal.* Some companies may not appear if their stock was not traded.

EXAMPLE 2.1 *Value of Limited Liability*

The rise of limited liability coincided with an increase in the size of firms. If it is efficient for firms to become large, and limited liability is the best structure for large firms to have, then a group of firms could limit competition if they could get a law passed that grants limited-liability protection to them alone. In Scotland until 1879, limited liability was granted to only three Edinburgh banks; all other competing banks had to accept unlimited liability.

We would expect that the limited-liability banks were larger and more successful than the others. In fact, even though over 50 banks with unlimited liability failed between 1845 and 1879, none of the three limited-liability banks failed. Furthermore, data from 1825 indicate that the three limited-liability banks averaged about 10 times the assets of the average bank with unlimited liability. After 1879, laws were changed, and all banks effectively became protected by limited liability.

Source: Carr and Mathewson (1988). See also Rasmusen (1988) for a discussion of ownership form and banks.

percent, or $100,000, per year for three years and repay the $1 million at the end of three years. Debt holders are paid first; stockholders are paid from what remains.

Table 2.1 illustrates this distinction between the claims of debt holders and shareholders. Suppose a corporation raises $1 million by borrowing $500,000 at 20 percent interest and selling 500,000 shares at $1 each. The corporation invests in Project 1, which has an equal chance of succeeding or failing. If the project succeeds, the corporation earns $2 million, of which $100,000 goes to interest payments and $500,000 to repay the loan. The remaining $1,400,000 goes to shareholders as dividends. If the project fails, the corporation goes out of business and sells its machines for $500,000, which goes to the debt holders.

On average, debt holders expect to receive a payoff of $550,000 (= 1/2 × $600,000 + 1/2 × $500,000), and equity owners expect to receive a payoff of $700,000 (= 1/2 × $1,400,000 + 1/2 × 0). The expected return on an investment is the expected payoff in excess of the initial investment divided by the initial investment, which the table shows is 10 percent for debt holders and 40 percent for equity owners.

The example shows that the expected return is higher for equity owners than for debt holders and that the payoff is not as variable for debt holders as for equity owners. In general, because debt holders get paid before equity owners, it is safer to hold debt. But, because debt is less risky, the expected return tends to be lower than for equity owners—if it were not, no one would hold equity. As the firm becomes more *highly leveraged*—increases its ratio of debt to equity—the expected returns to the equity holders rise. That is why stock prices fluctuate more for companies with high debt/equity ratios than for companies with low debt/equity ratios, all other things equal.

The amount of taxes firms pay depends on whether they are corporations, single proprietorships, or partnerships. For example, corporate income is taxed before it is

TABLE 2.1 Returns to Debt Holders and Equity Owners

Project 1

Outcome of Project	Probability	Payoff Received by Debt Holders	Payoff Received by Equity Owners
Success	.5	$600,000	$1,400,000
Failure	.5	500,000	0
Expected payoff		$550,000	$700,000
Initial investment		500,000	500,000
Expected payoff minus initial investment		50,000	200,000
Expected return		10%	40%

Project 2

Outcome of Project	Probability	Payoff Received by Debt Holders	Payoff Received by Equity Owners
Success	.5	$600,000	$1,300,000
Failure	.5	600,000	0
Expected payoff		$600,000	$650,000
Initial investment		500,000	500,000
Expected payoff minus initial investment		100,000	150,000
Expected return		20%	30%

distributed to a shareholder, who then pays personal income taxes. In contrast, income from proprietorships and partnerships is not directly taxed; it flows untaxed to the owners, who then pay personal income taxes.

Separation of Ownership and Control. The dramatic rise in the importance of the corporation caused a clamor in the 1930s about whether this organizational form was efficient. The debate was precipitated in part by *The Modern Corporation and Private Property*, by Berle and Means (1932), who argued that the corporate form separates ownership from control.[5] With separation of ownership and control, the owners of a

[5] Their book was and still is very influential. See Leibenstein (1966) and "The Symposium on Berle and Means" in the *Journal of Law and Economics,* June 26, 1983.

corporation, the shareholders, are typically not the managers, who are employees of the corporation. In contrast, single proprietorships and partnerships are run by the owners.

When control is separated from ownership, managers may not attempt to maximize profits and may pursue other objectives, like maximizing their own incomes, not working hard, and having plush offices (see Example 2.2).

In many corporations, there is often no single shareholder with the incentive to monitor managers' actions. Shareholders elect a board of directors to minimize the conflicts that arise because of the separation of ownership and control. The board's primary function is to act as an agent for the shareholders and oversee the efficient management of the company. But who monitors the board of directors? If they do a bad job, how will they be punished? One potential punishment is that they may not be reelected and may acquire bad reputations that make it difficult for them to get other good jobs. For example, in 1992, when facing massive debts, the large retailer R. H. Macy & Co. brought in outside directors to take control and to ensure that, in the event of a filing for bankruptcy, a majority of the board members would not be company employees.[6]

This control over the board of directors and over the managers may be inadequate to ensure profit-maximizing behavior. Therefore, according to Berle and Means, the actions of corporations cannot be predicted by a traditional economic analysis based on profit maximization. They implied that the severity of the Great Depression was at least in part attributable to the rise of this new and inefficient form of business.

Aside from the conflict that Berle and Means pointed out between equity owners and managers, conflict can also arise between debt holders and equity owners. For example, suppose the firm in Table 2.1 has already raised its $500,000 from debt holders and $500,000 from equity holders and is deciding between Project 1, which we've already examined, and Project 2, which pays $600,000 if it fails and $1,900,000 if it succeeds. The total expected payoff to the latter project is $1,250,000, as before. Yet the division of the payoff between equity owners and debt holders is different: Debt holders now receive $600,000 for sure, whereas the equity owners can expect to receive $650,000. The payoffs of the new project are summarized in the table.

The debt holders prefer this new project, but the equity owners prefer the original one. Because debt holders recognize that their interests may diverge from those of equity owners, debt holders often insist on **bond covenants**, which are restrictions on the corporation's choices of investment projects or further financing.

One interpretation of Berle and Means is that they were focusing attention on the monitoring problems and conflicts that arise as a firm grows. There is nothing inefficient about incurring costs as long as they are offset by benefits. Large corporations are not inefficient just because they entail monitoring costs. These costs can be offset by the benefits of larger size and the ability to raise money cheaply.

[6] Laura Evenson, "Macy's Board Facing Major Shakeup," *San Francisco Chronicle,* April 25, 1992: B1–B2.

EXAMPLE 2.2 *Conflicts of Interest Between Managers and Shareholders*

The stock market decline beginning March 2000 was followed by a series of revelations that managers had perpetrated outrageous fraud. The managers had engaged in actions that directly benefited themselves at the expense of unsuspecting shareholders. Often the frauds involved misreporting earnings to mislead investors and to raise the price of the stock at least temporarily.

One of the most spectacular frauds involved Enron. Enron had been primarily a natural gas pipeline company until it transformed itself into an energy-trading company. With the deregulation of energy markets, there was great incentive for buyers and sellers of energy to trade with each other through a variety of sometimes complicated contracts. Enron became a hugely successful company whose stock price by 2000 was four times what it had been four years earlier, and whose annual revenue was $200 billion. Then in fall 2001, everything began to come apart. Enron executives had apparently set up partnerships for themselves that did business with Enron. The partners allegedly stood to earn significant profits. The investments performed poorly for Enron, and huge losses piled up—losses that were unknown to the public because they were not explicitly revealed in the accounting information released by Enron. In November 2001, Enron made the startling announcement that it was correcting its past accounting numbers and was writing off several hundred million dollars of earnings; the firm disavowed its past four years of audited financial statements. Enron declared bankruptcy, and litigation followed.

Several Enron executives pleaded guilty to felonies, and others, proclaiming innocence, were indicted. Enron's auditor, Arthur Andersen, one of the leading accounting firms in the world, was convicted of obstruction of justice in connection with the shredding of documents and subsequently fell apart.

This scandal was only one of many. Massive frauds were uncovered in other industries. One of the largest telecommunication firms, WorldCom (now called MCI), was thrown into bankruptcy as a result of poor investments that were allegedly not handled

Size of Firms. A firm may expand because it wants to produce more of its basic output or because it chooses to produce inputs as well or distribute its output. The market and the firm are alternative means of providing goods and services. The higher the costs of doing business with other firms, the more tasks a firm performs itself. For example, as the relative costs of dealing with others changed, General Motors went from purchasing car bodies from others, as it did before 1926, to producing the car bodies itself.[7]

Although a firm may want to grow so as to avoid the cost of doing business with other firms, a larger firm faces higher costs and greater difficulty monitoring its own

[7] See the April 2000 issue of the *Journal of Law and Economics* for four articles that examine and dispute this claim.

properly in its accounts nor revealed. In addition, WorldCom made a loan of about $400 million to its CEO. The Securities and Exchange Commission (SEC) began to look into tightening accounting standards to prevent firms from reporting income in ways that, while technically accurate, distort a company's overall financial picture. Congress passed the Sarbanes-Oxley Act, which increases firms' reporting requirements and attempts to limit possible conflicts of interest between managers and shareholders.

Finally, the use of stock options as a device to pay and motivate employees came under scrutiny. An option provides the owner with the right to buy a stock at a fixed price. So, for example, an option on Microsoft at a strike price of $10 allows the owner to purchase one share of Microsoft stock for $10 even if the stock price is $50. Firms, especially high-tech companies, give options to employees to motivate them. If an employee receives an option for $10 when the stock is trading at $5, the option is not worth much. But the option becomes valuable if the stock price rises substantially. The option creates a strong incentive for employees to act to raise the stock price. Some observers believe that options so cloud executives' thinking that they are encouraged to misreport earnings and to engage in actions that cause stock prices to rise for long enough that they can exercise their options. Maybe, but an executive should expect that a fraudulent short-run strategy will be discovered when the company's stock eventually plummets.

There is not yet a simple answer as to why fraud was so prevalent in the late 1990s (nor even whether the amount of fraud was unusually large historically, given the opportunities), but two facts are clear. First, the conflict between managers and shareholders is a real one. Second, the use of accounting gimmicks and even the use of options (which may be a reasonable method to pay employees in many industries) are likely to diminish. In 2003, Microsoft, one of the largest granters of options as a payment device, announced that it would use stock ownership rather than options to motivate employees.

Source: David Nicklaus, "WorldCom Scandal Shows Dark Side of Stock Option Plans," *St. Louis Post Dispatch*, July 3, 2002:C1; "Running Out of Options," *Newsweek*, July 21, 2003:40; "'24 Days' Behind Enron's Demise," *Wall Street Journal*, August 8, 2003:C1.

managers and employees so as to ensure that they operate efficiently and profitably. The optimal size of a firm depends on this trade-off between the advantages and disadvantages of expanding. For example, Bill Gates, head of the software firm Microsoft, contends that[8]

> [A]ll large organizations are in a certain sense less efficient than small organizations. But they can do things that are super important and need to be done. Believe me, when I wrote and reviewed every line that went into the code, the overall quality of the code—according to me—was higher. And I've had to compromise.

[8] "Gates: Our Only Advantage Is We Bet on Windows." *InfoWorld*. August 3, 1992:102.

In Chapter 12, we contend that a firm is more profitable if it performs only those actions that it is good at and relies on others (the market) for other essential actions.

Most U.S. firms are small, though larger firms account for most employment and sales. In 1999, 56 percent of all manufacturing firms had nine or fewer employees and accounted for 4 percent of all manufacturing employees (*Statistical Abstract of the United States,* 2002, Table 715). The U.S. economy had about 6 million companies, roughly 89 percent of which employed fewer than 20 people. Only 0.3 percent of firms had 500 or more employees, but they accounted for 50 percent of all employees. The top 200 U.S. manufacturing firms in terms of value added accounted for about 22 percent of manufacturing employment and 40 percent of value of shipments of manufacturers in 1997 (*Concentration Ratios in Manufacturing,* 1997, Table 1).

The share of employment and assets of the largest U.S. firms has fallen since 1970. Because machines have become more productive, manufacturing output is now produced with fewer employees. For example, the fraction of the nonagricultural labor force in manufacturing declined from about 34 percent in 1950 to about 13 percent by the end of 2001 (*Economic Report of the President,* 2003, Table B–46). As a result, employment has shifted to industries such as services in which firms tend to be relatively small.

Mergers and Acquisitions

A firm may increase its size by expanding through investment, such as by building new factories, or by means of a merger: a transaction in which the assets of one or more firms are combined in a new firm. We use the term *merger* to include acquisitions. There are three types of mergers:

- Vertical merger: A firm combines with its supplier.
- Horizontal merger: Firms that compete within the same market combine.
- Conglomerate merger: Firms in unrelated lines of businesses combine.

Reasons for Mergers and Acquisitions

There are many explanations for mergers. The main motive is usually to increase profitability. Unfortunately for firms, not all mergers result in greater profitability. Moreover, some mergers may be profitable for the firm yet harm society by reducing efficiency. We now contrast the motives for mergers that increase efficiency with those that do not.

Mergers That Increase Efficiency. Acquisitions and other mergers that increase efficiency are desirable for society. There are a number of reasons why takeovers of existing firms may promote efficiency, including increasing scale to an optimal level, creating synergies, and improving management.[9]

[9] Although takeovers are common in the United States and United Kingdom, they are virtually nonexistent in Japan. Takeovers were rare in Germany before unification. Privatized East German state companies are being acquired: "Bidding for Europe's Takeover Business." *The Economist,* September 12, 1992:81.

Combining firms may reduce duplication or produce other benefits from increased size. For example, the firms may be able to save management costs by using a single set of managers to run both firms.

As the costs of factors of production change, the optimal size of a firm (that is, the output at which average cost is minimized) may increase. In the late 1800s the cost of transportation fell because of the development of railroads, and the cost of communication fell because of the advent of the telegraph and telephone. Further, the development of financial markets (for example, bond and stock markets) lowered the cost of raising large sums of money. These developments probably caused the optimal size of a firm to increase and led to the importance of the large corporation as the major organizational form in the U.S. economy.

Reduced transaction costs could explain why two firms that engage in different activities might prefer to merge. Bittlingmayer (1985) contends that the Sherman Act of 1890 created uncertainty about the legality of contracts between direct competitors and thereby created an incentive for firms that had been cooperating with each other through contracts to merge.

Firms that engage in different but complementary activities may benefit from mergers because of synergies or **economies of scope**: It is less costly for one firm to perform two activities than for two specialized firms to perform them separately. If one firm excels at designing fast cars and another firm excels at designing attractive cars, the two firms may gain by merging.

Acquiring a badly run firm and installing better management produces gains. Suppose the current managers of a firm are doing a poor job. The firm generates a large amount of cash, but the managers keep investing the money in unprofitable projects and raising their salaries, so that stockholders see little, if any, of the cash as a dividend. Stockholders could urge the board of directors to control management, but that may be difficult, especially if some members of the board are managers.[10]

An alternative way to discipline managers is to allow shrewd investors to discover inefficiently run firms. Such investors could then "take over" (acquire or gain control of) the inefficient firm at a low price, improve it, and either resell it or pass along the increased dividends to shareholders.

Imagine that the stock of the firm is worth $100 per share, based on the low dividends that current management is paying shareholders. You discover that this firm is badly run, acquire it, fire the current management, improve the firm's operations, and double dividends. As a result, the value of your stock in the company doubles to $200 per share. The threat that someone like you could come along and buy enough shares to gain control of the company might so scare the managers of the firm that they perform efficiently to avoid losing their jobs.

To gain control of the firm, you could offer to buy a controlling number of shares of stock from the current shareholders. Shareholders, however, stand to gain if (a) they

[10] Groups of investors may pressure firms to perform well. A study by Lilli Gordon for the California Public Employees' Retirement System finds that "relationship investors" (those who gain a seat on the board and induce the firm to behave well) have higher returns than the market as a whole. "A Fund in Wolf's Clothing?" *The Economist,* January 30, 1993:68.

keep their stock while you take over the firm, improve its performance, and raise dividends; (b) they sell to you at a price above $100; or (c) they hold on to their stock and you fail to gain control of the firm, but your attempt motivates current managers to improve their performance. Of course, the firm's managers may not care at all about the shareholders, and they may fight the attempted takeover in order to protect their comfortable jobs. If the managers are unsuccessful in preventing the firm from changing hands, a **hostile takeover** occurs. Battles to prevent hostile takeovers are intense, and managers often use clever tactics.[11]

It is also possible that the firm's managers believe that they could significantly improve profits if only the board of directors would allow them to fire employees, sell off parts of the business, and embark on new projects. Such radical changes in operation might not appeal to either shareholders or the board, so the managers themselves might decide to buy out the firm. A firm that is being taken over by its managers is said to be **going private**, because there are no longer any outside stockholders to whom management must answer. But how could a group of managers afford to buy out a corporation? One way is to use a **leveraged buyout (LBO)**, in which bonds based on the corporation's assets are sold in order to raise a tremendous amount of money. These bonds are sometimes called **junk bonds**, which are high-yield bonds backed by a corporation's assets and are considered riskier than typical corporate bonds. Junk bonds became popular in the 1980s as a way for investors to raise money to acquire control of a firm. It is safer to own a junk bond than a share of stock in the same firm because bondholders are paid before stockholders.

Mergers That Reduce Efficiency. Some mergers are disastrous: They reduce both efficiency and profitability.[12] Here, we focus on mergers where the new owners of a firm profit from the merger, yet production efficiency is reduced or other efficiency losses occur. Although the owners of the new firm may benefit, society loses. Such mergers may occur to take advantage of tax codes, for reasons of short-run exploitation, or to extend market or political power.

Because of the complexities of the U.S. tax code, firms may have a financial incentive to merge even if there is no gain from increased economic efficiency. Suppose Firm 1 has $100 in profits and Firm 2 has $100 in losses. If the corporate tax rate is 50 percent, Firm 1 must pay $50 in taxes, and Firm 2 pays nothing. If Firms 1 and 2 combine, their profit is zero. The profits of Firm 1 are offset by the losses of Firm 2, so the combined firm owes no taxes. The government gets $50 less, but the profit of the new firm is $50 more than the combined profits of the two firms had they not merged. Thus, although no economic efficiencies are created (the same amount of inputs is used to produce the same amount of outputs), the merger is privately profitable. Tax

[11] See **www.aw-bc.com/carlton_perloff** "Hostile Takeovers" for a discussion of how managers avoid hostile takeovers.

[12] For example, managers may desire to control large firms because they enjoy power, and they may pursue a policy of acquisition not because it is profitable but because it appeals to their ego, which may bias their judgment about value (Roll 1986).

reasons alone, however, do not account for much merger activity (Auerbach and Reishus 1988).

People might acquire a firm to take advantage of short-run gains, even if there are long-run losses. Suppose a firm has implicitly agreed to employ loyal workers even during slack times. As a result of this arrangement, workers receive lower wages in return for steadier employment. If management reneged on its arrangement and fired workers during slack times, workers would never again trust management. If you buy an inefficient firm and get rid of surplus labor in slack times, you can make a short-run gain. Workers will soon demand higher wages to compensate them for less steady employment, but in the meantime, you can run the firm more profitably than the previous management. Your action may harm the firm in the long run as the wage payments rise. Still, the short-run gain to the acquiring firm could offset the long-run loss (Shleifer and Summers 1988).[13]

If a sufficient number of firms in one industry merge, the resulting firm would face less competition and acquire additional *market power*: the ability of a firm to set price profitably above competitive levels. As we explain in Chapter 3, if price is greater than the competitive level, too little output is produced (production is inefficient). Therefore, the elimination of competitors through merging could lead to higher prices for consumers. Antitrust laws in the United States and in most other industrialized countries forbid mergers that are likely to reduce competition and lead to higher prices.

Some observers point to the relaxation of antitrust scrutiny as one of the reasons for the U.S. merger wave of the 1980s and 1990s. However, there is little evidence of significant increases in market power overall or in market concentration (Pautler 2001, White 2002). Even if firms are in different industries, so that there are no concerns about a reduction in competition, their amalgamation may create a potent political force that could influence legislation to their benefit at the expense of the rest of society.

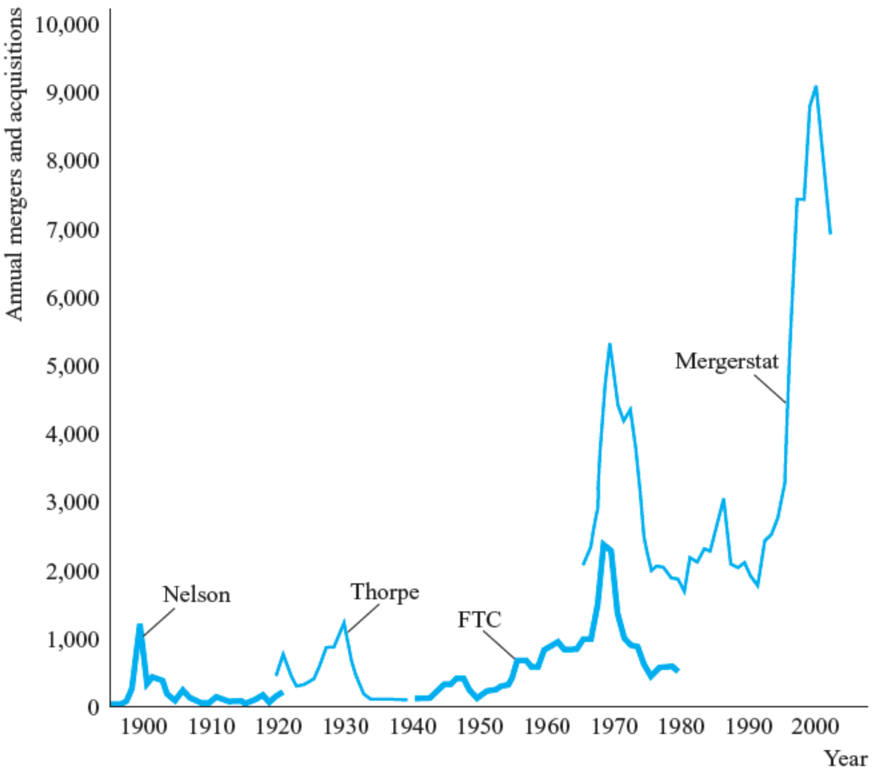
Merger Activity in the United States

Although newspaper articles often claim that the current period—starting with the Reagan era—is the period of greatest merger activity in history, even greater merger activity occurred in earlier times, when one adjusts for the size of the economy. Surprisingly, it is difficult to obtain consistent data on merger activity over time.[14] In early periods, data were kept on manufacturing and mining primarily. Over time these industries have declined in relative importance in the U.S. economy. The early data sources report only "large" transactions, ignoring mergers between small firms. As a result, measures of merger activity are biased downward, especially in earlier periods when firms tended to be smaller.

[13] Even if the firm's actions are inefficient, the firm's long-run losses from the higher wages could be offset by the firm's short-term gains.

[14] This section is based on Golbe and White (1988), and Andrade and Stafford (2001).

FIGURE 2.1 Annual Number of Mergers and Acquisitions



Notes: Nelson: Data derived by Nelson (1959) for manufacturing and mining. Thorpe: Data derived by W. Thorpe, reported in Nelson (1959, 166) for manufacturing and mining. FTC: Continuation of Thorpe series by Federal Trade Commission. Mergerstat: *Mergerstat Review* (2003).

Source: Adapted from Golbe and White (1988). Figure 9.6, in Alan J. Auerbach, ed., *Corporate Takeovers.* Copyright 1988 by the National Bureau of Economic Research. All rights reserved.

Bursts of merger activity appear to coincide with booms in the stock market for reasons that are not fully understood. Figure 2.1 shows the number of mergers since 1900 based on data from various sources. In the figure, there are five periods of extensive merger activity: one near the turn of the century, a second in the late 1920s, a third in the late 1960s, a fourth in the 1980s, and a fifth in the 1990s.

George Stigler (1950) called the first wave near the turn of the century the *merger to monopoly* movement. During this period, the U.S. economy was undergoing widespread changes in response to the development of railroads and communications. The stock market became a more important source of capital, and this period witnessed the creation of firms that, to this day, remain large and successful—among them, General Electric and U.S. Steel. The end of the first merger wave in the early 1900s coincided with a downturn in economic activity and with the Supreme Court's 1904 decision in the Northern Securities case, in which the Court found that certain (horizontal) mergers violated the antitrust law of the Sherman Act, which was passed in 1890.[15]

Stigler (1950) called the second wave in the 1920s the *merger to oligopoly* movement. The third wave in the 1960s is called the *conglomerate merger* movement because many of these mergers produced conglomerate firms or holding companies that own many firms that produce in different markets. There is no common name for the fourth wave. It was in this merger wave that hostile takeovers became more common, although they still remained a small share (less than 25 percent) of overall merger activity. The fifth wave could be labeled the *deregulation merger* wave because nearly half of the mergers took place in industries that had recently been deregulated, such as airlines, telecommunications, media, and banking.

News reports often proclaim that the 1980s and 1990s had unparalleled merger activity. Based on the pure number of mergers (Figure 2.1) or the nominal (not adjusted for inflation) value of the mergers, these statements are true. However, the economy is much larger today than near the turn of the century. If we compare mergers to the size of the economy, there was greater activity near the turn of the century. Figure 2.2 shows the ratio of the number of transactions per billion dollars of inflation-adjusted or "real" gross national product (GNP). Thus, the merger activity since the 1980s, though substantial, is not unprecedented.

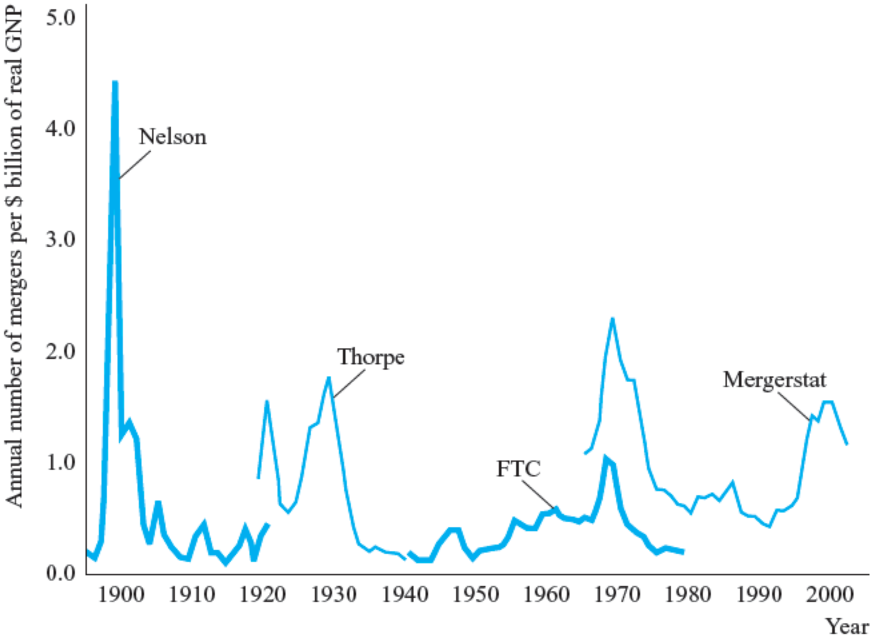
Merger Activities in Other Countries

Traditionally, mergers were much less common in Europe than in the United States. Now, however, mergers, hostile takeovers, and "going public" transactions are becoming more common in Europe, though there are still fewer mergers than in the United States. For example, the number of European Community mergers rose from 575 in 1984 to 1,159 in 1988 (Schmittmann and Vonnemann 1992). The number of mergers involving at least one of the top 1,000 EC firms rose from 185 in 1984/85 to 492 in 1988/89 (Jacquemin 1990). From 1980 to 1992, 95 mergers or acquisitions occurred in the U.S. defense industry and 40 in the comparable European industry (Reppy 1994).

One of the most debated issues in transitional central and eastern European economies has been the restructuring of state-owned enterprises. Some countries—such as Czechoslovakia and Russia—have privatized existing enterprises, while others—including Hungary and Poland—have tried to transform their enterprises before selling them. Regardless of the strategy these transitional economies used, they experienced massive and spontaneous breakups of state-owned enterprises at the beginning of the reforms.

[15] *Northern Securities Co. vs. U.S.,* 193 U.S. 197 (1904).

FIGURE 2.2 Annual Number of Mergers and Acquisitions per Billion Dollars of Real GNP



Note: Annual number of mergers and acquisitions per billion dollars of real GNP (in 1982 dollars); Nelson Series, FTC "Broad" series and Mergerstat.

Source: Adapted from Golbe and White (1988). Figure 9.7, in Alan J. Auerbach, ed., *Corporate Takeovers*. Copyright 1988 by the National Bureau of Economic Research. All rights reserved.

For instance, the number of industrial enterprises employing more than 25 workers went from about 700 in 1990 to 2,000 by mid-1992 in Czechoslovakia. Most industries have become less concentrated: The largest firms' share of the industry has declined in terms of both value of output and employment. This shift in industrial activities from large to small or medium-sized enterprises in the first years of transition was brought about by a combination of breakups of large companies and rapid entry of new firms.

Empirical Evidence on the Efficiency and Profitability of Mergers

Much debate has arisen about whether the recent wave of mergers and acquisitions benefits the economy. Obviously acquisitions and mergers that lead to production efficiency are desirable. Some worry, however, that many mergers and acquisitions merely involve a reshuffling of ownership that produces short-run stock gains for financial manipulators who are not interested in the long-term health of firms. Others are concerned that mergers create greater market power that hurts consumers by raising prices. Andrade and Stafford (2001) and Mueller (1997) survey the recent studies of the evidence on mergers and acquisitions. Earlier studies include Bradley, Desai, and Kim (1988), Jarrell, Brickley, and Netter (1988), Jarrell and Poulsen (1987), Jensen (1988), Jensen and Ruback (1983), Romano (1985), Scherer (1988), and Shleifer and Vishny (1988). The following paragraphs summarize the findings of these studies.

Returns to Acquired Firm. Shareholders of an acquired firm receive a premium of about 16 to 25 percent above the stock price prevailing prior to the acquisition activity. Much of the increase in the share price of an acquired firm occurs just *before* public announcement of the transaction. The premium received by shareholders rose significantly as a result of the Williams Act, which required firms to reveal their takeover plans, and the gains to shareholders of acquired firms have increased over time.

Effects of Preventing Mergers. Management tactics to thwart takeovers reduce the probability of a takeover but raise the acquisition price, if the takeover is successful. When someone fails to gain control of a firm, the increase in its stock price caused by that person's bidding is completely eliminated, and price returns to its previous level.

The evidence is mixed on the effect of defensive provisions such as supermajority amendments, greenmail, and poison pills on stock prices. Managers (who often own some stock) may try to enact a shareholder agreement under which anyone who seeks to gain control of the firm must obtain the approval of a supermajority (more than 50 percent) of the firm's shareholders. Such a rule makes it easier for a group of current shareholders to prevent a takeover. Adoption of supermajority amendments lowers a firm's stock price, presumably because of the reduced likelihood of a takeover.

A firm may try to dissuade a particular individual from taking over the firm by using **greenmail**, in which the firm buys back the shares of the person who is trying to take over the firm (and only that person's shares) at a premium. Greenmail has a negative effect on a firm's stock price. A firm that changes its state of incorporation to take advantage of the new state's strong antitakeover laws enjoys a slight increase (though not a statistically significant one) in its share price.

In **poison-pill** arrangements, the corporation must make stock available at bargain prices to original shareholders—but not to someone who takes over the firm—if the firm is taken over, thereby diluting the value of the new owner's stock. These arrangements significantly lower the stock price of the firm. Poison pills decrease the value of taking over the firm, raising the costs of acquisition and thereby reducing a potential buyer's incentives to try to acquire the firm.

Returns to Acquiring Firm. The shareholders of an acquiring firm do not earn substantial, above-average rates of return as a result of the acquisition. They do slightly better in hostile takeovers than in friendly mergers. The return to stockholders of acquiring firms has declined over time from about 4 percent in the 1960s to -3 percent in the 1980s and 1990s. The return to acquirers depends on whether the target is purchased with stock or cash, with acquirers doing better when cash is used. The use of stock as a means of payment increased by about 50 percent from the 1980s, with about 60 percent of transactions in the 1990s financed entirely by stock.

When faced with a hostile takeover attempt, a firm's managers may seek a friendly firm or individual, known as a **white knight**, to come to their rescue, obtain control of the firm, and leave current management in place. White knights, on average, overpay for the firms they acquire.

Returns to Society. Overall, total shareholder value of the combined companies rises about 2 to 7.5 percent after the consolidation. The increased value of a consolidated firm is not typically due to the creation of market power.

If the new firm acquires market power, the price consumers face will rise. This increased market power, however, also benefits the rivals of the combined firm, and hence their stock prices should rise. If the transaction is motivated by greater efficiency in production, the combined firm will be a more efficient competitor, and the stock price of its rivals should decline in anticipation of the increased competition. Stillman (1983), Eckbo (1983), and Banerjee and Eckard (1998), investigating the merger wave at the turn of the nineteenth century, conclude that the second explanation is more consistent with the evidence.

Instead of using stock-price data, some researchers look directly at accounting data from the consolidated firm to see if the new firm is more efficient. Data problems are more severe with this approach than with looking at stock prices because accounting data are often difficult to interpret. Moreover, the estimated efficiency gains for the firm are likely to be smaller than those estimated from data on stock prices because those figures apply to the increase in the *equity* value (not total value, which includes debt) that results from acquisition. In Mueller's (1997) survey of 20 studies covering 10 countries, only a few studies find increased profitability from merger. Two particularly ambitious U.S. studies, Scherer (1988) and Ravenscraft and Scherer (1987), do not find increases in profits after acquisition based on their examination of profit data by line of business from the 1960s and 1970s. In contrast, Lichtenberg and Siegel (1987) examine the productivity of individual plants using more recent data and detect significant improvements in efficiency in plants whose ownership had changed. Moreover, they find that the plants most likely to undergo an ownership change were those that were performing poorly. Andrade and Stafford (2001) criticized Scherer's studies for failing to control for industry benchmarks. Controlling for such benchmarks reveals that mergers generally improve the efficiency of the firm and lead to increased profits. Finally, that mergers are often concentrated in the same industries in any one time period lends further support to an efficiency rationale for mergers (Jovanovic and Rousseau 2001).

Moreover, contrary to what some commentators allege, there is no evidence that consolidated firms are "myopic" and cut back on research and development (R&D). Hall (1988) finds that R&D spending is not influenced by the change in control.

In summary, stock market evidence supports the view that merger activity improves efficiency and creates value. Shareholders of target firms are the primary beneficiaries of this increased value. As legislation and new management tactics have made it more difficult to gain control of firms, the returns to shareholders of target firms have increased and those to shareholders of acquiring firms have decreased. Moreover, there appears to be no increase in market concentration and market power, so consumers do not lose. Nor is there a reduction in R&D.

Additional research on profits subsequent to consolidation, not on stock prices, is needed to confirm these efficiency gains. Without such research, some may argue that mergers and takeovers create illusory stock market value that represents either the unjustified transfer of wealth from those dependent on the acquired firm (for example, employees) to its shareholders, or valuation errors by the stock market. The work of Andrade and Stafford (2001) appears to confirm the stock market evidence regarding the efficiencies of mergers.

Cost Concepts

By running efficiently, a firm can produce output at the lowest possible cost. Every firm needs to know what it costs to produce its products if it is to make sensible business decisions. There are a variety of ways to measure costs, and some cost concepts are more appropriate for certain problems than others. This section explores these different cost concepts and some subtleties in understanding them.

Types of Costs

Firms typically incur costs that do not vary with output and costs that do. A fixed cost (F) is an expense that does not vary with the level of output. A fee a government charges for a firm to incorporate and conduct business is a fixed cost. Whether the firm produces a lot or a little, it must pay the fee. Another example is the monthly rent that a lawyer must pay for an office after signing a one-year lease. The monthly rent must be paid regardless of how much business the lawyer does.

If the firm and the lawyer decide to go out of business, they would not renew their incorporation or rental agreement for the next year. But what if they decide to go out of business just one month after they began? Must they still pay their fee or monthly rental? If they have paid in advance, can they get a refund? The answer depends on the law or contract. The firm probably prepaid the entire incorporation fee, which is not refundable. The lawyer, although obligated to pay a monthly rent, may be able to rent to someone else and recoup some, if not all, of the cost. The portion of fixed costs that is not recoverable is a **sunk cost**. A sunk cost is like spilled milk: Once it is sunk, there is no use worrying about it, and it should not affect any subsequent decisions. In contrast, a fixed cost that is not sunk *should* influence decisions. For example, whether or not the lawyer should go out of business depends in part on how costly it is to get out of the lease (the financial penalty for breaking the lease). Costs, including fixed costs, that are not incurred if operations cease are called **avoidable costs**.

Variable costs (VC) are costs that change with the level of output, q. Because variable costs vary with output, we normally write them as a function of output: $VC(q)$. Typically, as output increases, so does the need for labor, electricity, and materials, so variable costs depend on the wages and prices that a firm must pay for inputs.

Total costs (C) are the sum of all fixed and variable costs: $C = F + VC$. Associated with the concepts of total cost and variable cost is **marginal cost** (MC), which is the *increment,* or addition, to cost that results from producing one more unit of output.[16] Because fixed cost does not change as output increases, the increase in total cost when output increases is identical to the corresponding increase in variable cost.

It is important to distinguish between the concept of marginal cost and the various concepts of average cost. There are three common types of average cost: *average total cost* (sometimes simply called *average cost*), *average variable cost,* and *average fixed cost*:

- **Average cost** (AC) (sometimes called *average total cost* or ATC) is total cost divided by output: $AC = C(q)/q$.
- **Average variable cost** (AVC) is variable cost divided by output: $AVC = VC(q)/q$.
- **Average fixed cost** (AFC) is fixed cost divided by output: $AFC = F/q$.

Because AC is the sum of AVC and AFC, AVC and AFC cannot exceed AC:

$$AC(q) = \frac{C(q)}{q} = \frac{VC(q) + F}{q} = \frac{VC(q)}{q} + \frac{F}{q} = AVC(q) + AFC(q).$$

Even though marginal cost is independent of fixed costs and average cost is not, it is *not* necessarily true that, at any given output level, marginal cost is less than average cost. The reason that marginal cost may exceed average cost is that marginal cost refers to *changes* in cost, not to levels.

Imagine going into a supermarket to buy fruit. You carry a bag and put in some apples, which naturally differ in weight. The total weight of the apples in the bag and the associated average weight per apple are easily determined. Suppose you *add* a very small apple to your bag. Its weight is the increment to the weight of the apples in the bag (the marginal weight). But the weight of the small apple is less than the average weight of the apples already in the bag, so the average weight falls. If, instead, you add a very large apple, its marginal weight exceeds the average weight of the apples already in the bag, so the average weight rises. The marginal weight is totally determined by the *one* additional apple. The average weight (after the additional apple) is determined in large part by the apples that were already there. Analogously, marginal cost can be either above or below average cost.

To illustrate further the relationship of marginal cost, average cost, and average variable cost, Table 2.2 shows how the various cost measures vary as output increases. In this example, the fixed cost is \$100 regardless of whether production occurs or not (output $= 0$). This fixed cost is an obligation that cannot be avoided by going out of business, so the fixed cost is sunk or nonrecoverable.[17]

[16] If $C(q)$ is the total cost of producing q units, then the marginal cost is $MC = dC(q)/dq$.

[17] If part of the fixed cost were recoverable, as when a license fee is refundable, then the relevant cost for output of zero would be only the sunk cost. For example, a firm that goes out of business but obtains a \$60 refund on its \$100 state license fee has costs of \$40 for producing nothing.

TABLE 2.2 An Example of Cost Concepts

Output	Fixed Cost	Average Fixed Cost	Total Variable Cost	Average Variable Cost	Total Cost	Average Total Cost	Marginal Cost
0	100		0		100		
1	100	100	10	10	110	110	10
2	100	50	19	9.5	119	59.5	9
3	100	33.3	25	8.3	125	41.7	6
4	100	25	32	8.0	132	33	7
5	100	20	40	8.0	140	28	8
6	100	16.7	49	8.2	149	24.8	9
7	100	14.2	60	8.6	160	22.9	11
8	100	12.5	73	9.1	173	21.6	13
9	100	11.1	88	9.8	188	20.9	15
10	100	10	108	10.8	208	20.8	20

In the table, the variable cost rises from 0 to 108 as output expands from 0 to 10. Total cost—the sum of fixed plus variable costs—rises from 100 to 208 as output expands to 10. Marginal cost equals the increase in total costs that results from producing an additional unit of output. It initially falls, reaches a minimum of 6 at 3 units of output, and then rises.

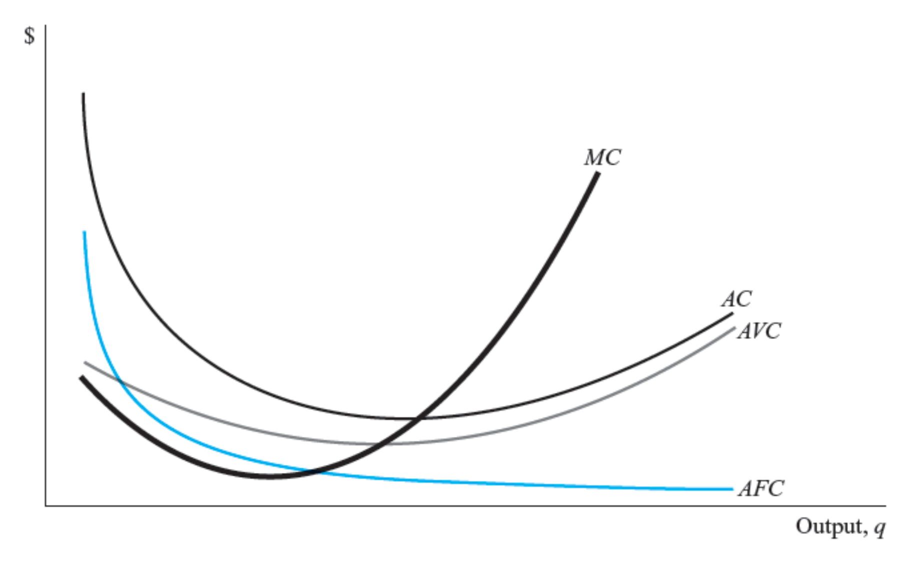
Average variable cost equals total variable cost divided by output, and average total cost equals total cost divided by output. Average total cost always exceeds average variable cost, but as shown, marginal cost may be less than, equal to, or greater than average total or average variable cost.

There is a geometric relationship between *MC*, *AVC*, and *AC* as depicted in Figure 2.3. When *MC* is below *AVC*, the *AVC* curve is falling. When *MC* is above *AVC*, the *AVC* curve is rising. When *MC* equals *AVC*, the *AVC* is at its minimum. A similar relationship exists between *MC* and *AC*. Figure 2.3 also shows that, as output increases, average fixed cost (*AFC*) approaches zero and *AVC* and *AC* get closer together.

The apples example can be used to illustrate why *AC* rises if *MC* exceeds it or falls if *MC* is below it. If you add an apple that is heavier than the average apple, the average weight of the basket increases. Conversely, if you add a lighter-than-average apple, the average weight falls.

In general, total costs depend on the amount of output produced as well as the prices of the factors of production (for example, wages of workers and the price of raw materials). Figure 2.3 illustrates how a typical (short-run) average cost curve varies with output: Average cost eventually rises as output expands because it becomes more

FIGURE 2.3 Cost Curves



costly to produce as output increases within a given plant. The curves are based on the assumption that the prices of the factors of production (for example, wages of employees) are held constant. If, for example, wages were to rise, then the entire average cost curve would shift up. It is *not* necessarily true that the curve would shift straight upward because the minimum-cost output may well change. That is, the size of a firm that yields minimum average cost depends on the wages of labor and costs of all other factors of production.

A cost curve summarizes an enormous amount of information. For instance, knowing how the cost curve changes as wages or other factor prices change, one can infer a firm's production technology: the relationship between inputs and output reflecting the maximum possible output that can be produced from a given set of inputs. In other words, knowing the cost function of a firm and knowing its technology are equivalent.[18] For example, suppose the wage rate is $10 per hour, and workers are the only input used to produce corn (seeds are free). To plant 1 bushel of corn costs $10, to plant 2 bushels costs $20, and so on. From this information on costs and

[18] Let $\mathbf{x}$ = an input or vector of inputs (for example, labor and raw materials), q = output, $\mathbf{w}$ = wage rate (and other unit prices for inputs), $F(\mathbf{x})$ = the production function (output as a function of inputs). The cost function $C(q, \mathbf{w})$ is derived by solving the following problem: Minimize the cost of producing q units subject to the constraint that q units are produced according to the engineering relationship between q and $\mathbf{x}$. Knowledge of $C(q, \mathbf{w})$ allows one to infer $F(\mathbf{x})$ under reasonable assumptions (using "duality theory," which is explained in Varian 1992, Ch. 6).

wages, we can infer that the production technology is that one worker can plant 1 bushel of corn per hour.

Cost Concepts

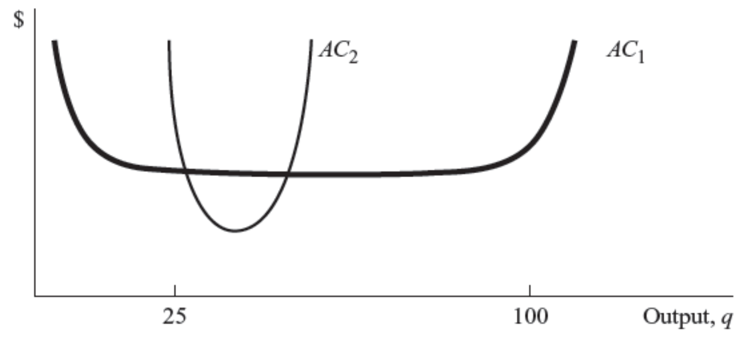
Although the definitions of the various cost concepts may seem straightforward, several complicated issues are associated with them. We now explore the most important ones.

Cost Factors in Addition to Output. A firm's costs depend on how much it produces for any given set of input prices. But factor prices are generally not the only influence on cost (Alchian 1959). The costs of production depend not only on how much is produced but also on *how fast*. Producing something quickly is more costly than producing it slowly. Moreover, variation in the rate of production over time matters. For example, steady production of 60 units/hour for 10 hours might involve lower costs than 100 units/hour for 2 hours plus 50 units/hour for 8 hours, even though total production is 600 units in either case.

It might be cost effective for a firm to spend money to make its plant highly adaptable to different levels of production. If a business is seasonal (for example, New Year's cards), the relevant cost is not the cost of producing a specific output but rather the cost of producing the range of outputs experienced during the year. If output fluctuates between 25 and 100 units per month, then a plant with a cost curve like AC_1 in Figure 2.4 might well be more efficient (that is, have lower total cost) than one with the curve AC_2, even though the minimum of the AC_2 curve is lower than that of the AC_1 curve.

The Short Run Versus the Long Run. The short run is a time period so brief that some factors of production cannot be costlessly varied. The long run is a period of time sufficiently long that all factors of production can be costlessly varied. For example, at the end of the year, the lawyer who rented an office is free to renew the lease or lease a new space. However, during the course of the year, the lease may not be broken without cost (there are sunk costs). In this example, the short run is less than one year, whereas the long run is one year or longer.

FIGURE 2.4 Cost Curves of Different Technologies



Another example illustrating the difference between short and long run has to do with installed machinery, which is costly to move and reinstall. If machines last for one year and must then be replaced, the number of machines can be regarded as predetermined in the short run of one year, though not in the long run. More generally, the short run is that time period during which the number of machines and physical space (the plant) are fixed and cannot be varied except at so substantial a cost that it is never profitable. In the short run, the firm must make do with its current plant and stock of machines. In the long run, the firm can alter its capital: It can buy new machines, discard old ones, and even move into a different plant designed to allow production of any given level of output at minimum cost.

The distinction between short and long run is not precise. Indeed, there is a continuum of runs, with increasingly more adjustment possible as the length of the run increases. The firm must incur greater costs, **adjustment costs**, as it increases the speed at which it adjusts its operations.[19]

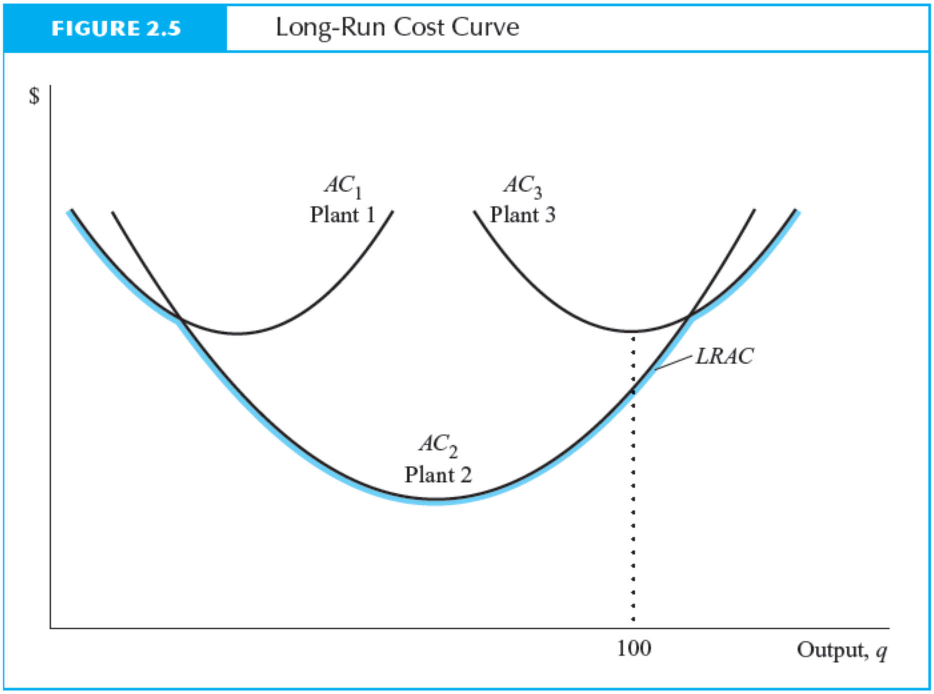
A firm can configure itself in any way it wants in the long run, but in the short run its choices are constrained. Therefore the long-run average cost is always at least as low as the short-run average cost. This relationship between long-run and short-run costs implies that the long-run curve is the *envelope* of the short-run curves; that is, the long-run average cost curve (*LRAC*) is the relevant section of whichever short-run average cost curve (*SRAC*) is lowest at that particular quantity, as Figure 2.5 shows. In the short run, suppose that the firm can only have a single plant size. In the figure, there are three possible *SRAC* curves, AC_1, AC_2, and AC_3. Notice that the *LRAC* is not always the minimum point of a short-run average cost curve. In Figure 2.5, the least expensive way to produce 100 units is to use Plant 2, even though that is not the output that minimizes average cost in Plant 2 but *is* the output that minimizes average cost in Plant 3. In textbooks, one typically draws the long-run average cost curve so that it eventually rises as output expands, which means that the firm's efficient size (the largest output that minimizes average cost) is finite.

Opportunity Cost. As Adam Smith said, "The real price of everything is the toil and trouble of acquiring it." That is, an action's **opportunity cost** is the value of the best forgone alternative use of the resources employed in that action. For example, if a firm hires three workers at the going wage of $10 per hour, then its labor cost is $30 per hour. In this example, the opportunity cost and the actual out-of-pocket costs are the same. Suppose, instead, that one of the three workers is the firm's owner, who does not receive a wage. An economist still measures the opportunity cost of the three workers at $30 per hour: The labor used by the firm is worth $30 because another firm would value the labor at that amount.

We can use opportunity costs to determine whether it is profitable to continue an activity. To return to the example, suppose that each worker produces 1 unit of output per hour, which sells for $9. The owner calculates the profits earned in one hour as the revenue of $27 minus the cost (using opportunity cost as the measure) of $30 for a net

[19] See **www.aw-bc.com/carlton_perloff** "Adjustment Costs" for details.

FIGURE 2.5 Long-Run Cost Curve



loss of $3. The presence of a loss shows that the owner should cease production and work for someone else at $10 per hour. Clearly, the owner is better off earning $10 per hour than earning $7 (= $27 − $20) in wages.

The opportunity cost concept is very useful in determining whether a firm should continue to use an asset that it owns when that asset could be rented readily. Consider a firm that owns the building it occupies. If the building could be rented to another tenant for $1,000 per month, then the firm should count that amount as its cost of occupying the building. It is the forgone earnings of not renting out the building. If the firm cannot afford to pay itself rent (because doing so would result in a negative profit), then the firm should realize that its use of the building is not the most profitable—it would be better to go out of its current business and rent the building.

Surprisingly, if all costs are valued at their opportunity cost, then profit need only be *zero* to make remaining in business worthwhile. Opportunity cost values *all* resources used at the highest value they could receive elsewhere. If revenues just cover costs, then all resources (for example, the owner's time, the firm's building) are being used in an efficient manner and would not be worth more if used elsewhere. Because opportunity cost values each resource at its most profitable alternative use, economists sometimes say that opportunity cost attributes a normal profit (best possible profit from an alternative use of the resource) to all of the firm's resources.

Expensing Versus Amortizing. Suppose that a firm rents a machine by the month for $100 and then decides to purchase the machine outright for its market price of $10,000.

Should it count all $10,000 as a fixed cost incurred in its month of purchase, or should it spread the cost over the months the machine will be used? When costs are counted as they are incurred, they are said to be **expensed**; when they are spread out over the useful life of the machine, they are said to be **amortized**. If the firm amortizes the cost of the machine, how much should it charge itself? The answer clearly affects how the firm judges its performance.

The simple answer to any question about the appropriate cost to assign a durable asset is that the relevant cost is the *rent* that the owner could earn by renting the asset to someone else. This calculation is often easy—when a firm owns an office building and uses only some of the space for its own needs, it determines the appropriate market rent when it rents space. In other cases, the appropriate rent may not be available; for example, there is no rental market for blast furnaces. How should the cost of such assets be treated? One answer is to calculate the cost of owning an asset as the lost interest on its value (if it were sold for $100, that $100 could be earning interest) plus the depreciation on the asset. Economic **depreciation** is the decline in the value of an asset during the year (for example, using a machine causes it to wear out and fall in value). Even when installed assets cannot be resold, one can still use this method to calculate a rent.[20] The resulting profit calculation reveals whether the firm's decision to install the machine was a good one and whether further investment would be profitable.

Economies of Scale

A firm's average costs may remain constant, rise, or fall as its output expands. If average cost falls as output increases, the firm is said to have **economies of scale** (or **increasing returns to scale**); if average costs do not vary with output, it has **constant returns to scale**; and if average cost rises with output, the firm is said to have **diseconomies of scale** (or **decreasing returns to scale**). In Figure 2.3, the firm first enjoys economies of scale, then (at least for one output level) it has constant returns, and then it suffers from decreasing returns to scale. If a firm enjoys economies of scale at all output levels, then it is efficient for one firm to produce the entire market output (see the discussion of "natural monopoly" in Chapter 4).

Reasons for Economies of Scale

There are many reasons to expect a firm's average costs to decline, at least initially, as its output expands. One is that fixed setup costs do not vary with the level of output. For example, a publishing company typically incurs substantial costs to have a book written. Editors must be paid and the plates for printing made. If 100 rather than 50 books are produced, the cost does not rise by a factor of 2 because the additional books require few additional costs. Another example is an automobile stamping facility. Typically, special dies must be made to press the parts into their unique shapes. The more parts produced with each die, the lower the average total cost of production.

[20] See **www.aw-bc.com/carlton_perloff** "Depreciation."

Average costs tend to fall with increased output for a second reason. As output expands, a firm can use its labor in more specialized tasks. For example, at low levels of business, one lawyer may handle both divorce and bankruptcy cases. As the law firm expands, one lawyer may specialize in divorce, while another specializes in bankruptcy, and each one can develop expertise in one area. If a training cost is associated with developing expertise in each task, only a firm that requires frequent repetition of each task finds it worthwhile to train separate workers for each task (see Example 2.3).

If a firm manufactures several products in one plant, the length of the production run could increase as output expands. Consider a paper manufacturer that sells three

EXAMPLE 2.3 *Specialization of Labor*

Why doesn't everyone work individually and sell finished products to others as needed? One answer is that it can be more efficient to break down production processes into several small steps in which workers specialize. Two examples illustrate the advantages of breaking production into several tasks.

Writing at about the time of the American Revolution, Adam Smith (1937, 4–5) offered an example to show that the division of labor can have important advantages in the "very trifling manufacture" of pin-making:

> [A] workman not educated to this business . . . nor acquainted with the use of machinery employed in it . . . could scarce, perhaps, with his utmost industry, make one pin a day, and certainly could not make twenty. But in the way in which this business is now carried on, not only the whole work is a peculiar trade, but it is divided into a number of branches, of which the greater part are likewise peculiar trades. One man draws out the wire, another straightens it, a third cuts it, a fourth points it, a fifth grinds it to the top for receiving the head; to make the head requires two or three distinct operations; to put it on, is a peculiar business, to whiten the pins is another; it is even a trade by itself to put them into the paper; and the important business of making a pin is, in this manner, divided into about eighteen distinct operations, which, in some manufactories, are all performed by distinct hands, though in others the same man will sometimes perform two or three of them. I have seen a small manufactory of this kind where ten men only were employed, and where some of them consequently performed two or three distinct operations. . . . [T]hey could, when they exerted themselves, make among them about twelve pounds of pins in a day [or] upward of forty-eight thousand pins in a day.

Similarly, Henry Ford became the largest automobile manufacturer in the early 1900s, and probably the most profitable, by developing mass production. He adapted the conveyor belt and assembly line so that he could produce a standardized, inexpensive car in a series of tasks in which individual workers specialized. He achieved cost savings despite paying wages that were considerably above average.

grades of paper. To produce each grade requires a separate setup of the production line. If the firm is small and has only one production line, then two switchovers are needed to produce the three grades daily. But if the firm triples in size, it can have one production line for each grade and thus avoid switching costs.

Certain physical laws generate scale economies; the best known concerns the relationship between volume and surface area. Suppose a chemical firm plans to make a certain liquid in a spherical container. The volume of the sphere is $(4\pi r^3)/3$ where r is the sphere's radius. The cost of the sphere depends on how much steel it takes to make it. That cost is related, not to volume, but to the surface area of the sphere, which equals $4\pi r^2$. Doubling the radius raises volume (and output) by a factor of 8, but raises surface area by only a factor of 4.

Similarly, there is a natural economy of scale in the holding of inventories and replacement parts because of the law of large numbers. This statistical law holds that random events tend to cancel out if there are enough of them, so that a firm's inventory as a fraction of its sales shrinks as the firm grows.[21]

Total Costs Determine Scale Economies

Even if economies of scale characterize some functions of a firm, diseconomies of scale may characterize other functions. Whether the firm experiences economies of scale overall depends on the contribution of each function to overall cost. For example, just because an individual plant has economies of scale in production, one should not conclude that it is most efficient to have only one plant producing. Such a false conclusion ignores other types of costs, such as monitoring costs and transportation costs.

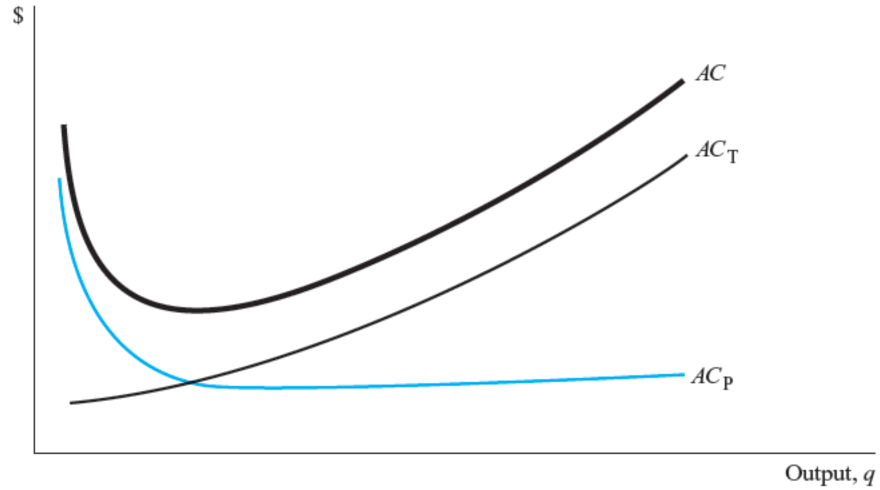
Suppose that a firm produces pasteurized milk and delivers it to grocery stores. The fewer the plants, the farther, on average, the milk has to be shipped, and the higher the transportation costs. Even if there are substantial economies of scale in production, it is not efficient to have one plant if transportation costs are very high. The relevant average cost curve is the sum of the cost of producing the milk and the cost of transporting it to customers.

Figure 2.6 shows the *AC* curve of production, AC_P. It slopes downward initially, indicating economies of scale in production. The average cost of transporting raw materials to the plant and transporting the milk to customers is AC_T. As more milk is produced in one location, it must be shipped farther, and so average transportation costs rise. The sum of these two curves is the overall average cost, *AC*, which is the relevant curve for determining the cost of operation. The output at the minimum of the *AC* curve would be smaller if the transportation costs increase so that AC_T becomes steeper. That means, all else equal, that the optimal size of the plant becomes smaller as transportation costs become more important. Many small-scale plants are common in industries characterized by high transportation costs.

The location of a plant is influenced by the relative costs of transporting raw materials to the plant compared to transporting the output from the plant to customers.

[21] See **www.aw-bc.com/carlton_perloff** "Scale and Inventories" for a detailed explanation.

FIGURE 2.6 Total Average Cost



The higher the cost of transporting raw materials, the closer the plant will be to their source. For example, shipping bromine is expensive (and dangerous) compared to shipping bromide fluids (which are made from bromine). Therefore, bromide fluid plants tend to be located close to facilities that make bromine.

Conversely, if raw materials come from many different locations, or if they are readily available in many locations, the differences across locations in the transportation costs for obtaining raw materials may be insignificant, so the plant tends to locate close to its customers. For example, cement is costly to transport, and the main raw material, limestone, is widely available. Therefore, cement plants tend to locate close to their customers.

The decision of how many plants a firm should have depends on both the cost of transporting raw materials and finished products and the economies of scale in production.[22] The more important the economies of scale in production, the more likely that production is concentrated in only a few plants. The greater the transportation costs (and the greater the dispersion of customers), the more likely that production is decentralized in several plants.

A Measure of Scale Economies

Scale economies exist if average cost falls as output expands. As long as marginal cost is below average cost, economies of scale exist; if marginal cost exceeds average cost, there are diseconomies of scale. This relationship suggests that a natural measure of scale

[22] See Scherer et al. (1975) for an extended analysis.

economies is the ratio of average to marginal cost.[23] If $s = AC/MC$, then economies of scale exist if $s > 1$, constant returns to scale exist if $s = 1$, and diseconomies of scale exist if $s < 1$ (see Appendix 2A).[24]

Empirical Studies of Cost Curves

Economists often estimate firms' cost curves and economies of scale. Because economies of scale refer to cost savings that arise as output increases, it is important in any study of economies of scale to verify that output is the *only* variable accounting for cost differences among firms (or for the same firm over time). Large firms may differ from small firms in many ways; for example, they may produce more products or perform different functions, such as marketing.

Cost differences between firms are due to economies of scale only if the two firms studied produce the same products and perform the same functions. If one firm markets its product itself, whereas the other, smaller firm does not, an analysis that failed to account for this difference would find diseconomies of scale: Average costs would appear to rise as output expands, when in fact the opposite may be the case.

Some studies focus on whether economies of scale characterize certain specific functions, such as purchase of equipment and operating costs. Other studies ask the more general question of whether economies of scale characterize the entire operations of a firm or group of firms (see Example 2.4).

Economies of Scale in Total Manufacturing Costs

Some firms have U-shaped long-run average cost curves. At the lowest point of the curve, output q^*, the average cost curve is flat. Empirical studies of manufacturing firms often find that cost curves are L-shaped: As output rises, the average cost curve slopes down sharply, slopes down more slowly, and finally is flat. That is, for small output levels, there are large economies of scale, but for large outputs, those economies are exhausted and average costs are constant. On L-shaped cost curves, we can determine the lowest output level, q^*, such that the long-run average cost curve is essentially flat.

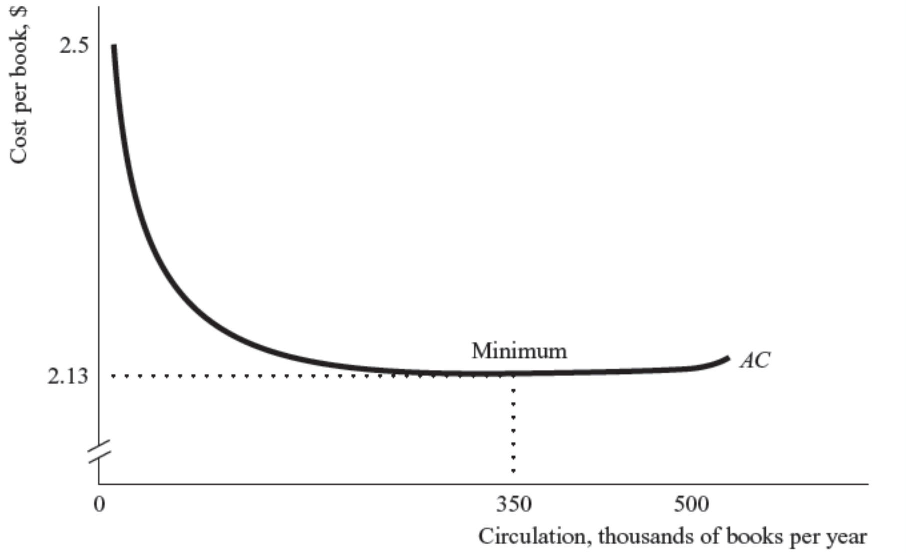
[23] This discussion of economies of scale is based on the cost function, which answers the question: What is the minimum cost of producing a given amount of output? A natural measure of scale economies is the ratio of average cost to marginal cost. Instead, one could base the definition of economies of scale on the production function, which answers the question: How much output is produced with given amounts of labor and raw materials? That is, economies of scale exist if an equal percentage increase in the use of all factors of production results in a proportionately greater expansion of output. So, for example, if a firm increases its use of its two inputs, labor and raw materials, by 10 percent, economies of scale exist if output rises by more than 10 percent. A natural measure of scale economies, therefore, is the percentage of output expansion generated by an increase of 1 percent in the use of all inputs. One can show that these two measures are equivalent in a competitive industry.

[24] If a firm makes zero profit, then s measures the ratio of costs to revenues.

EXAMPLE 2.4 *Indiana Libraries*

According to DeBoer (1992), Indiana libraries have U-shaped average cost curves when circulation is used as the measure of output. However, most libraries operate on the strictly declining portion of the *AC* curve. Average cost (including costs of labor, books, utilities, and equipment, but not capital costs such as rent and debt-service expenditures) is $3.62 for a small library with a circulation of 2,000 per year, $2.59 at a circulation of 10,000, and reaches its minimum of $2.13 at a circulation of 350,000. As the circulation rises beyond 350,000 books, average cost rises slightly.

A government can use this information to determine how much more it costs to have several branch libraries rather than a single central library. For example, the average cost is 5.5 percent higher with four branch libraries with 50,000-book circulations rather than a central library with a 200,000-book circulation.



A plant's minimum efficient scale (*MES*) is the smallest output (q^*) it can produce such that its long-run average costs are minimized. The size of the *MES* plant, especially in relation to the overall market, is useful for judging how many firms could operate in a market.

A useful measure of the importance of scale economies is the measure of the cost disadvantage incurred by a plant that is smaller than the *MES*. If this disadvantage is small, then economies of scale are unimportant.

Table 2.3 lists some engineering estimates of *MES* for various industries in the United Kingdom together with the cost disadvantage that would be incurred if plants equal in size to 50 percent of the *MES* were built. Pratten (1971) found that in only about 25 percent of the cases examined was the cost disadvantage of producing in the

TABLE 2.3 Estimates of Minimum Efficient Scale (MES)

Product	*MES* (physical output per year)	*MES* as a Percent of U.K. Market	Percent Increase in Unit Cost Incurred by a Plant of 50% *MES*
Oil	10 million tons	10	5
Chemicals			
Ethylene	300,000 tons	9	25
Dye	large	100	22
Sulfuric acid	1 million tons	30	1
Beer (brewery)	at least 1 million barrels	3	9
Steel production	9 million tons	33	5–10

Source: Pratten (1971) as reported in Siberston (1972, 380).

suboptimal size plant more than 10 percent. Weiss (1976) used Pratten's results to show that, for most industries, the size of the *MES* plant is typically a small percentage of total U.S. output. This work implies that for most industries, plant economies of scale are not so significant as to preclude having many firms in an industry.

Survivorship Studies

Another approach to measuring economies of scale is due to Stigler (1968b), who made use of the following simple but powerful observation: If a particular plant size is efficient, eventually all plants in the industry should approach that size. Any plant or firm size that survives for a long time is efficient. Accordingly, Stigler classified the fraction of output from petroleum-refining plants of various sizes, as Table 2.4 shows. Stigler uses these data to conclude that the very smallest and the very largest plants are inefficient, because their share of industry output declined over time.

TABLE 2.4 Distribution of Petroleum Refining

	Percent of Industry Capacity		
Plant Size (percent of industry total)	1947	1950	1954
under .1	8.22	7.39	6.06
.1–.2	9.06	7.60	7.13
.2–.3	5.45	4.99	7.28
1.5–2.5	17.39	23.64	22.45
2.5–4.0	21.08	16.96	15.54

Source: Stigler (1968b, 69).

TABLE 2.5 Number of Beer Plants

Annual Capacity (thousands of barrels)	1959	1971	1979
0–25	11	2	2
26–100	57	19	8
101–250	51	19	6
251–2000	88	67	26
2001–3000	5	9	6
3001–4000	3	3	7
4001+	2	7	20

Source: Elzinga (1986, 215).

If all firms face similar cost conditions, a survivorship study reveals the efficient plant size as the industry replaces its obsolete plants. If firms face different costs or produce *different* products, their optimal scales will vary, and a survivorship study can only identify the *range* of efficient plant sizes. In other words, economies of scale measure how costs fall as output expands, *holding all else constant.* If other factors are not constant across plants, a survivorship study does not reveal the efficient plant size but merely describes the range of efficient plant sizes.

Since Stigler's initial study, there have been numerous applications of his survivorship method to other industries (see, for example, Rogers 1992). For example, there has been a dramatic decline in the number of beer plants since 1947. The number of plants fell from 465 in 1947, to 253 in 1958, 108 in 1974, 96 in 1978, and 80 in 1983. Table 2.5 illustrates the composition of beer plants by size (annual capacity). It shows that, from 1959 to 1979, the share of the smallest plants diminished and the share of the largest plants grew. The data suggest that, in the beer industry, economies of scale at the plant level have become increasingly important. In the 1980s and 1990s, however, many microbreweries opened, suggesting the development of a new technology.

Cost Concepts for Multiproduct Firms

Most firms do not produce a single product; it is typical for a firm to make several different, perhaps related, products. For example, a doughnut shop produces filled doughnuts and plain doughnuts, a doctor treats sore throats and skin rashes, and a plumber fixes sinks and bathtubs. A firm that produces many different products is called a *multiproduct firm.*[25] The multiproduct nature of firms does not materially affect the analyses in most of this text. However, it is important to remember that treating

[25] See Baumol, Panzar, and Willig (1982) and Panzar (1989) for detailed studies of multiproduct firms.

firms as multiproduct is more realistic than not and that, in some cases, ignoring the multiproduct characteristics of firms can lead to improper conclusions or regulations (see Chapter 20).

Adaptation of Traditional Cost Concepts for a Multiproduct Firm

If a firm produces two or more products, one cannot measure *the* average cost or *the* marginal cost because there is no one measure of output. One can, however, define cost concepts that are analogous to those in a single-product environment. For example, if q_1 units of Product 1 and q_2 units of Product 2 are produced, the marginal cost of producing Product 1 is the additional cost incurred by increasing q_1 to $q_1 + 1$, holding the output of Product 2 constant at q_2. In this definition, the marginal cost of Product 1 depends not only on the level of output for Product 1 but also on q_2. Marginal cost for Product 2 is defined analogously.

Unlike marginal cost, average costs are not as easy to define in a multiproduct context. The problem arises in trying to decide whether to divide total cost by the output of Product 1, q_1, or Product 2, q_2. Perhaps total cost should be divided by the sum, $q_1 + q_2$. There is no single right answer, but several relevant average cost concepts have been suggested (see Appendix 2A).

Aside from the extrapolation of the concepts of marginal and average cost to a multiproduct environment, there are some cost concepts that arise only in a multiproduct setting. The most important such cost concept is economies of scope (see Appendix 2A for some others).

Economies of Scope

When it is cheaper to produce two products together (joint production) rather than separately, there is an economy of scope (Baumol, Panzar, and Willig 1982; Panzar and Willig 1977a). For example, a steer produces beef and hide. Although it is possible to use some steers just for hide and others just for beef, it would be inefficient with current technology.

Economies of scope imply that it is efficient to produce two or more products together; they do not necessarily imply that these products should be produced by a single firm. For example, consider how steel is made. First, iron ore is melted down into pig iron in a blast furnace; the molten pig iron is then run into a steel-making furnace.

It is possible to conceive of two separate firms, side by side, one of which makes pig iron and the other steel, with a pipe carrying the molten pig iron between the two firms.[26] As we discuss in Chapter 12, when firms must rely heavily on each other, transaction costs are high and each firm is liable to be exploited. High transaction costs

[26] Pipes are used to connect two separate firms in some industries. For example, a lead additive used to be added to gasoline to prevent engine knock. In the early 1980s, an Exxon gasoline refinery in Louisiana was located beside a plant owned by the Ethyl Corporation that made the lead additive and delivered it by pipe to Exxon.

explain why only a single firm typically produces all the products for which economies of scope exist.

Many possible factors contribute to economies of scope, and one of the most important is the use of common inputs. In the example of producing beef and hide, it is easy to see why it might be best to produce both simultaneously rather than using one steer for beef and another for hide.

Knowledge is one of the most important common inputs for producing and selling related products. Information about one product is likely to be relevant for another closely related product. For example, knowing how to market steel bars efficiently (knowing among other things where customers are located) might help in marketing steel sheets. Or knowing how to manufacture steel bars efficiently (knowing where to obtain low-price iron ore) might contribute to the efficient manufacture of steel sheets. In such situations, it is efficient to produce and sell these products together. Otherwise, resources like information would have to be duplicated wastefully. Moreover, because it is difficult to buy and sell information, a single firm often produces related products.

A final example of using common input arises when a person's physical presence is required for certain services. Consider a plumber who handles a wide variety of plumbing problems and can fix sinks as well as bathtubs. It might be that a plumber who repairs only sinks could service them better than a more versatile plumber. For that matter, there might be gains from specializing further and having one plumber repair sink washers and another repair sink stoppers. But a homeowner would have to call several plumbers to diagnose the problem before finding the right specialist. In other words, because of the *indivisibility* involved in diagnosing a problem (you need one person physically present to do it), it would be inefficient if that person were unable to fix a wide range of plumbing problems. If the gains from specialization in plumbing were great, it might be worth having specialists—perhaps even a specialist at diagnosing problems. But as long as the gains are small, such specialization is unlikely.

Economies of Scale and Economies of Scope

Firms often produce many products to gain economies of scope in marketing and distribution. A salesperson who sells white bread to a store can also sell rolls. A store may prefer to deal with one person who can satisfy all its needs rather than with several different salespeople. A firm that produces and sells many products can specialize production by plant, thereby obtaining economies of scale in production while maintaining a full product line. The disadvantage of such specialization is that transportation costs may rise as individual products must be shipped farther. See Examples 2.5 and 2.6.

Specialization in Manufacturing

Firms often produce different products in the same plant. The U.S. Bureau of the Census publishes a measure of how specialized each plant's output is for each industry. The specialization ratio for an industry equals shipments of products in the particular industry divided by total shipments of all products for all plants listed as being in the industry. For example, suppose there is only one plant that makes steel bars and the

EXAMPLE 2.5

The Baking Industry

The baking industry provides an excellent example of multiplant specialization. Until recently, bakeries typically produced a wide range of products (breads, rolls, cakes) and served relatively small geographic areas. Because bakery products are perishable, shipping distances were limited in earlier times. The development of improved preservatives extended the shelf life of baked goods with the result that shipping distances could be increased. Bakeries began to acquire nearby bakeries and use *reciprocal baking* to produce their products. Reciprocal baking means that plants become specialized in particular products and then ship their products to each other, so that each geographic area is served by a full line. Reciprocal baking allows bakery firms to take advantage of scale economies and still preserve the economies of scope in marketing that come from having a full product line.

plant also makes steel wire. If the plant sells $100 worth of steel bars and $50 worth of steel wire, it is classified as producing in the steel bar industry. The specialization ratio for this industry is 2/3, or 66.7 percent. The specialization ratio for an industry is typically in excess of 80 percent. This high share indicates that individual manufacturing plants (not necessarily firms) are relatively specialized.

A tabulation of the number of different industries in which one firm operates indicates that 146 of the top 200 manufacturing firms (in terms of shipments) operated in 11 or more different industries in 1968 (Scherer 1980, 76). Dunne, Roberts, and Samuelson (1988) study all manufacturing firms (which numerically are dominated by very small firms) and find that in 1982, firms on average produced between one and two separate products. Multiplant firms on average produced between two and three separate products.

An Example of an Industry with Economies of Scope

Friedlaender, Winston, and Wang (1983) estimate a multiproduct cost function for each of the four U.S. auto makers. They postulate that costs depend on prices of various inputs (for example, wages, raw materials) and various outputs (for example, small cars, large cars, and trucks). Their statistical procedure also adjusts for the differing physical specifications of small and large cars and trucks.

They use a generalization of the scale economy measure s for a multiproduct firm (Equation 2A.1 in Appendix 2A). They find that $s = 1.23$ for General Motors at a typical point, suggesting that GM has economies of scale. That is, if GM expanded output of small cars, large cars, and trucks by 10 percent, costs would rise by about 8 percent (10/1.23).

They are also able to measure the degree of economies of scope (see Equation 2A.2 in Appendix 2A), which depends on the group of outputs being considered. For example, the economy of scope, SC, of producing large cars together with small cars and trucks is defined as

EXAMPLE 2.6

Electricity Minimum Efficient Scale and Scope

There are both scale and scope economies in electricity distribution. Yatchew (2000) estimates that the minimum efficient scale is achieved by utilities with about 20,000 customers in Ontario (and about 30,000 customers in New Zealand). He observed Canadian utilities with between 600 and 220,000 customers. Consequently, he concludes that the merger of utilities that increases their size is unlikely (in most cases) to produce savings in distributing electricity due to scale economies. He also finds that those Canadian utilities that deliver electricity and other municipal services (46 percent of utilities deliver other services such as water and sewage) had costs that were 7 to 10 percent lower than those that delivered only electricity, indicating economies of scope.

$$SC = \frac{\begin{array}{c}C(\text{large cars alone}) + C(\text{small cars} + \text{trucks}) - \\ C(\text{large cars} + \text{small cars} + \text{trucks})\end{array}}{C(\text{small cars} + \text{trucks})},$$

where C stands for the total cost of producing the indicated outputs. SC indicates the percentage increase in costs that would occur if large cars were produced separately from small cars plus trucks. For GM, this number equals 25 percent, which indicates that there are substantial benefits from combining the production of large cars with small cars plus trucks. Surprisingly, no economies of scope arise from producing trucks together with small and large cars; it appears that truck production could occur in a separate firm with no loss of efficiency.

SUMMARY

In Western countries, most output is produced by firms and most of these firms are profit maximizers. Large firms in the United States and elsewhere are organized as corporations with limited liability. Corporations typically raise money by issuing debt and equity (stock). The corporation must make sure that its managers operate to maximize profits and do not pursue different goals that would adversely affect other concerned parties, such as debt holders and shareholders.

Large corporations are run by managers and not by the owners. If managers do not run firms efficiently or maximize profits, the firm may be driven out of business or taken over by others. Not all acquisitions and mergers of firms necessarily lead to greater efficiency and profitability, though. Since the 1980s, merger activity has reached a relatively high level (though not as high as at the turn of the century). Much, but not all, empirical analysis indicates that takeovers create economic value and that shareholders of the acquired companies capture the lion's share of the gains.

To maximize profit, a firm must minimize the cost of producing a given level of output. An economist's definition of cost is based on the concept of an opportunity cost, which includes a normal profit. A cost function shows how much it costs the firm to produce various amounts of output, or, in the case of a multiproduct firm, various combinations of different outputs. A cost function depends not only on the output produced, but also on the price of the factors of production such as the wages of workers and the price of raw materials.

There are many different types of costs: sunk costs, fixed costs, variable costs, avoidable costs, marginal costs, average variable costs, and average total cost. Some cost functions exhibit economies of scale, while others do not. A typical manufacturing process exhibits economies of scale at least initially. But the other functions of the typical firm, such as administration, monitoring, marketing, and delivery, may entail costs that exhaust all scale economies and lead to an optimal firm size.

When a firm produces several different products, an analysis of costs requires the development of cost concepts analogous to those used with a single-product firm, and the development of new cost concepts such as economies of scope. Cost concepts for a multiproduct firm explicitly recognize that the cost of producing one product depends on the amount of other products that are produced.

PROBLEMS

1. A firm purchased copper pipes a few years ago at \$10 per pipe and stored them. These were used only as the need arose. The firm could sell its remaining pipes in the market at the current price of \$9. What is the opportunity cost of each pipe and what is the sunk cost?
2. A refiner produces heating fuel and gasoline from crude oil in virtually fixed proportions. What can you say about economies of scope for such a firm? What is the sign of its measure of economies of scope (SC)?
3. For each situation below, discuss how two separate firms could carry out the activities. Identify those areas in which transaction costs are highest and you would expect to see only one firm.
    a. Oil pipelines, once built, cannot be moved. A pipeline ends at an electric power facility, which buys oil.
    b. A golf course locates beside a hotel.
    c. A postcard manufacturer wants a readily available supply of custom-made paper.
    d. A candy manufacturer needs to purchase sugar daily.
4. *(Difficult)* The managers of Firm A recommend that Firm A purchase Firm B because the purchase will diversify the business of Firm A. Diversification of risks is a desirable strategy for individual shareholders, but if shareholders can diversify their risks by holding stock in Firm B, is there any reason for Firm A to purchase Firm B? Suppose labor turnover is costly; could that provide an efficiency saving to support the proposed purchase? (*Hint:* If output is less variable, labor employment can be steadier.)
5. In the very short run, practically all costs are fixed. Does that mean that marginal cost is zero?
6. If there are economies of scope and if the price for each product equals marginal cost, is it possible for a firm to cover all its costs? If the firm's average cost of production declines the more it produces, can a price equal to marginal cost ever cover all its costs?
7. Suppose the cost of producing q_1 cars and q_2 trucks is $10{,}000 + 70q_1 + 80q_2$. Calculate the marginal cost of producing cars and the measure of scope economies when $q_1 = 100$ and $q_2 = 200$.
8. Why can the measure of economies of scope not exceed one as long as marginal costs are always positive?
9. Suppose there are a wide range of plant sizes in an industry. What do you conclude about the shape of the average cost curve if the plants are in the same area? Assume plants in the same area face similar costs. How does your answer change if the plants are located in different countries?

Answers to the odd-numbered problems are given at the back of the book.

SUGGESTED READINGS

Coase (1937) first asked why firms exist and gave a clear answer. Williamson (1975, 1985), Alchian and Demsetz (1972), and Klein, Crawford, and Alchian (1978) expand on this topic. See the interesting exchange between Coase and Klein and others in the April 2000 issue of the *Journal of Law and Economics*. Most of the discussions in these books and articles are relatively nontechnical. Calvo and Wellisz (1979) and Holmstrom (1979) explain the role of supervision and observability more formally. The articles in Auerbach (1988a) and Kaplan (2002) present the major theories and empirical evidence on mergers and acquisitions.

APPENDIX 2A

Cost Concepts for a Multiproduct Firm

When moving from a single-product to a multiproduct environment, one must adapt some of the definitions of cost and develop some new concepts to characterize cost.[1]

Total Costs

Suppose that $C(q_1, q_2)$ represents the cost of a firm that produces q_1 units of Product 1 and q_2 units of Product 2. The marginal cost of producing Product 1 at any given output level is defined, as in the single-product case, as the incremental cost of producing one more unit of Product 1—except now it is necessary to specify not only how much of Product 1 is being produced but also how much of Product 2. In mathematical terms, the marginal cost of Product 1 is just the (partial) derivative of $C(q_1, q_2)$ with respect to q_1.

Average Costs

What meaning can be given to the concept of average cost? The answer is that there is no unambiguous measure of average cost. Although total cost is well defined, there is no one unique output level to choose when two products are produced. One could define total output as $q_1 + q_2$, but that literally would be akin to adding apples and oranges. In fact, there is no reason why any linear combination of output, $a_1q_1 + a_2q_2$, is better than any other, where a_1 and a_2 are any two numbers.

If one specifies the proportions in which Products 1 and 2 are produced, it is possible to define an average cost concept, called *ray average cost* (RAC). Let λ_1 and λ_2 be the proportions in which Products 1 and 2 are produced, so $q_i = \lambda_i q$ implicitly defines q, a scale of output measure. Then, RAC is defined as total costs divided by q. That is,

$$RAC(q) = \frac{C(\lambda_1 q, \lambda_2 q)}{q}.$$

Using $RAC(q)$, one can define increasing ray average costs, constant ray average costs, and decreasing ray average costs. $RAC(q)$, of course, depends on the values of λ_1 and λ_2. If λ_1 and λ_2 are arbitrarily given, the multiproduct case reduces to the single-product case. For any given value of λ_1 and λ_2, we can calculate RAC and then find the scale, q, that minimizes RAC—just as in the single-product case. However, the scale at which RAC is minimized along different rays (different combinations of λ_1 and λ_2) generally differs.

[1] See Baumol, Panzar, and Willig (1982, Ch. 3, 4) and Panzar (1989) for a detailed treatment of these topics.

For example, consider an automobile company that makes small and large cars. If it is required to have a 50 percent mix, its average production cost may be minimized at 1 million units of each type of car. However, if the mix is 25:75 percent, its average production costs may be minimized at 1 million small cars and 3 million large cars.

It is possible to show that $RAC(q)$ falls, rises, or is constant as q increases, depending on whether s (a measure of scale economy) is above, below, or equal to 1, where[2]

$$s = \frac{C(q)}{q_1\frac{\partial C}{\partial q_1} + q_2\frac{\partial C}{\partial q_2}}. \tag{2A.1}$$

That is, s is the multiproduct analogue of the ratio of average to marginal cost. As in the single-product case, if firms are pricing at marginal cost, then s is the ratio of costs to revenues. In the single-product case, if s exceeds 1 so that AC exceeds MC, AC decreases with q, whereas if s is below 1 so that AC is less than MC, AC rises with q. Similarly, in the multiproduct case, if s exceeds 1, RAC falls with q, whereas if s is below 1, RAC rises with q. Thus, s can be viewed as measuring the proportionate increase in total costs from a percentage increase in the amount of *all* outputs. If s exceeds 1, costs increase by less than the percentage increase in output.

In addition to RAC, there are several cost concepts that do not have a clear analogy to the single-product case. Consider the cost of producing q_2 units of Product 2:

- The *incremental costs* of increasing Product 2 from 0 to q_2 holding Product 1 constant is $IC_2 = C(q_1, q_2) - C(q_1, 0)$.
- The *average incremental costs* of increasing Product 2 from 0 to q_2 holding Product 1 constant is $AIC_2 = [C(q_1, q_2) - C(q_1, 0)]/q_2$.

The incremental cost of producing q_2 units of Product 2 includes any fixed cost associated with the production of q_2 and depends on the assumed production of q_1.

Economies of Scale

The *product-specific economies of scale* (PS_i) of q_i, holding the other output, q_j, constant is defined using the AIC:

$$PS_i \equiv \frac{AIC_i}{MC_i}.$$

[2] *Proof:*

$$\frac{dRAC(q)}{dq} = \frac{1}{q}\left[\lambda_1\frac{\partial C}{\partial q_1} + \lambda_2\frac{\partial C}{\partial q_2}\right] - \frac{1}{q^2}C(q) = \frac{1}{q^2}\left[\lambda_1 q\frac{\partial C}{\partial q_1} + \lambda_2 q\frac{\partial C}{\partial q_2} - C(q)\right]$$
$$= \frac{1}{q^2}\left[q_1\frac{\partial C}{\partial q_1} + q_2\frac{\partial C}{\partial q_2} - C(q)\right].$$

Hence, $dRAC(q)/dq > 0$ if and only if $\Sigma q_i\ \partial C/\partial q_i > C(q)$, or $1 > C(q)/(\Sigma\ q_i\ \partial C/\partial q_i)$, or $1 > s$.

PS_i is the same as the scale measure *s* defined earlier for the particular case where all outputs except q_i are held fixed. The *AIC* cost function is like a typical single-product average cost function. The multiproduct cost function is converted into a single-product function by fixing the level of all outputs except one.

Economies of Scope

Most firms produce more than one product because it is cheaper to produce them together rather than separately. The term *economy of scope* refers to the savings that result from doing so. Consider the production of q_1 units of Product 1 and q_2 units of Product 2. The cost of producing each separately is $C(q_1, 0) + C(0, q_2)$; the cost of producing them together is $C(q_1, q_2)$. Economies of scope, *SC*, are measured as

$$SC = \frac{[C(q_1, 0) + C(0, q_2) - C(q_1, q_2)]}{C(q_1, q_2)}. \tag{2A.2}$$

SC measures the relative increase in cost that would result if the products were produced separately.[3] If *SC* is everywhere positive, it is cheaper to produce the products together. If marginal costs are positive, *SC* cannot exceed 1.[4]

When a firm increases its output of several products, it takes advantage of both economies of scope and economies of scale if they exist. It is possible for these two types of economies to have offsetting effects. A cost function is *trans-ray convex* at a given point if the cost of producing a linear combination of any two appropriately chosen output vectors is less than the weighted cost of producing the outputs separately.[5]

An Example

Suppose that it costs \$100 to rent a machine that can produce either red balloons or blue balloons. Let q_1 be the number of red balloons and q_2 the number of blue balloons. Suppose the cost function is $C(q_1, q_2) = 100 + q_1 + 2q_2$.

[3] If $\partial^2 C/\partial q_1 \partial q_2 < 0$, then Products 1 and 2 have *weak cost complementarity*. Increased production of one product lowers the marginal cost of the other. Here, economies of scope must necessarily exist (Panzar 1989).

[4] *Proof:* If $SC > 1$, $C(q_1, 0) + C(0, q_2) - C(q_1, q_2) > C(q_1, q_2)$ or $[C(q_1, 0) - C(q_1, q_2)] + [C(0, q_2) - C(q_1, q_2)] > 0$. But each term in the parentheses must be negative if the marginal cost is positive. Therefore, the inequality cannot hold, and *SC* cannot exceed 1.

[5] The formal definition of trans-ray convexity (Baumol, Panzar, and Willig 1982, Ch. 4, Def. 4D1) is: "A cost function $C(q)$ is trans-ray convex through some point $q^* = (q^*_1, \ldots, q^*_n)$ if there exists any vector of positive constants $w_1, \ldots, w_n$ such that for every two output vectors $q^a = (q^a_1, \ldots, q^a_n)$ and $q^b = (q^b_1, \ldots, q^b_n)$ that lie on the hyperplane $\Sigma\, w_1 q_i = w_0$ through point q^* (so that they satisfy $\Sigma\, w_i q^a_i = \Sigma\, w_i q^b_i = \Sigma\, w_1 q^*_1$), for any k such that $0 < k < 1$ we have

$$C(kq^a + [1 - k]q^b) \leq kC(q^a) + [1 - k]C(q^b)."$$

Trans-ray convexity can be related to conditions associated with natural monopoly (Panzar 1989).

The cost function shows that it costs \$1 to produce an additional red balloon but \$2 to produce an additional blue balloon, after the machine is purchased. Now several of the cost concepts that have been discussed can be illustrated.

The marginal cost of Product 1 is the derivative of $C(q_1, q_2)$ with respect to q_1. In this case, the marginal cost of Product 1 is constant and equals 1. The marginal cost of Product 2 is also constant and equals 2.

Next, we turn to the ray average cost. Suppose $\lambda_1 = .5$ and $\lambda_2 = .5$. Then

$$C(.5q, .5q) = 100 + .5q + 2 \times .5q = 100 + 1.5q.$$

Hence

$$RAC(q) = \frac{(100 + 1.5q)}{q} = \frac{100}{q} + 1.5.$$

In this example, *RAC* falls as q increases.

The measure of scale economies, s, is

$$s = \frac{C(q_1, q_2)}{\left(q_1\frac{\partial C}{\partial q_1} + q_2\frac{\partial C}{\partial q_2}\right)} = \frac{100 + q_1 + 2q_2}{q_1 + 2q_2} = \frac{100}{q_1 + 2q_2} + 1.$$

Thus, in this example, the measure of scale economies must exceed 1, so scale economies are always present.

If q_1 were produced separately, the cost would be $C(q_1, 0) = 100 + q_1$. Similarly, if q_2 were produced separately, the cost would be $C(0, q_2) = 100 + 2q_2$. The cost of producing q_1 and q_2 separately is

$$C(q_1, 0) + C(0, q_2) = 200 + q_1 + 2q_2.$$

This latter cost is clearly greater than the cost of producing them together, $C(q_1, q_2)$. Using equation (2A.2), we can calculate economies of scope as

$$\begin{aligned} SC &= \frac{[C(q_1, 0) + C(0, q_2) - C(q_1, q_2)]}{C(q_1, q_2)} \\ &= \frac{[(100 + q_1) + (100 + 2q_2) - (100 + q_1 + 2q_2)]}{[100 + q_1 + 2q_2]} \\ &= \frac{100}{100 + q_1 + 2q_2}. \end{aligned}$$

Because *SC* exceeds zero everywhere, it is cheaper to produce the two goods together rather than separately.

By fixing the level of output of one of the products, say q_2, we can calculate $AIC(q_1)$ as $[C(q_1, q_2) - C(0, q_2)]/q_1$ or

$$AIC(q_1) = \frac{[(100 + q_1 + 2q_2) - (100 + 2q_2)]}{q_1} = \frac{q_1}{q_1} = 1.$$

Thus, *AIC* is constant and equals 1. (Notice that the marginal cost of Product 1 is also constant and equals 1.) Because *AIC* is constant, there are no product-specific economies for q_1 (or for q_2), yet there are overall scale economies.

PART TWO

Market Structures

CHAPTER 3

Competition

Thou shalt not covet; but tradition approves all forms of competition.
—Arthur Hugh Clough

Perfect competition provides a benchmark against which the behavior of other markets is judged. The chapter starts by examining perfect competition, even though its strong assumptions apply to only a few markets. Next, two useful tools of analysis, elasticities and residual demand curves, are discussed. Then the chapter shows that competition has desirable efficiency and welfare properties. These properties, however, depend crucially on the assumptions of free entry and exit and on no externalities (firms bear the full costs of their actions). The chapter discusses the adverse effects if these conditions do not hold. The chapter concludes with examples of industries that most economists would characterize as reasonably competitive.

In this chapter, we stress five key points:

1. Perfect competition has many desirable properties.
2. Free entry and exit is a crucial factor in determining whether a market is perfectly competitive and efficient.
3. One important measure of welfare is maximized under perfect competition.
4. The desirability of perfect competition is reduced in the presence of externalities such as pollution.
5. Even if some of the necessary conditions for perfect competition do not hold, markets can come close to achieving the desirable properties of perfect competition.

Perfect Competition

Even though perfect competition is rarely, if ever, encountered in the real world, we study the perfect competition model because it provides an ideal against which to compare other models and markets. In later chapters, we examine how actual markets deviate from perfectly competitive ones, and determine which markets are likely to have the greatest deviations. The desirable properties of a perfectly competitive economy explain why economists generally favor competition. That a market deviates from the perfectly competitive model does not necessarily mean that the performance of a market can be improved, however, as is discussed at length throughout the book.

Assumptions

We define perfect competition as a market outcome in which all firms produce a homogeneous, perfectly divisible output; producers and consumers have full information, incur no transaction costs, and are price takers; and there are no externalities. That is, the main assumptions of perfect competition are:

- *Homogeneous Perfectly Divisible Output.* All firms sell an identical product. Consumers view the products of various firms as the same and hence are indifferent between them.
- *Perfect Information.* Buyers and sellers have all relevant information about the market, including the price and quality of the product. Firms can produce and consumers can buy a small fraction of a unit of output. As a result, the amount of output demanded or supplied varies continuously with price. This technical assumption avoids problems caused by large discrete changes in either supply or demand in response to small price changes.
- *No Transaction Costs.* Neither buyers nor sellers incur costs or fees to participate in the market.
- *Price Taking.* Buyers and sellers cannot individually influence the price at which the product can be purchased or sold. Price is determined by the market, so each buyer and seller takes the price as given.
- *No Externalities.* Each firm bears the full costs of its production process. That is, the firm does not impose externalities—uncompensated costs—on others. For example, pollution produced by a firm is an externality because the firm does not recompense the victims.

Some economists also assume that a perfectly competitive market has a large number of buyers and sellers. If there are many similar firms, no one firm can charge a price above the market price without losing all its customers, so the firm views the price at which it can sell as beyond its control. Similarly, consumers cannot find a firm willing to sell below the market price, so consumers must view the market price as beyond their control. Moreover, even if there are relatively few firms in a market, no firm can raise its price above the market price without losing all its customers if another firm can quickly enter the market and underprice it. Thus, because we assume firms and consumers are price takers, we do not also assume either that there are a large number of firms, or free entry and exit.[1] Competitive markets typically have a large number of firms and consumers, but industries can have all the properties of perfect competition even though there are few firms in those industries.

The Behavior of a Single Firm

Let us first examine the incentives of a typical firm. Suppose a firm has the short-run cost curves in Figure 3.1 and faces a market price of p_0. How much should it produce? Indeed, should it produce anything at all?

Profit Maximization. The objective of any firm, including a competitive firm, is to maximize its profits (or, equivalently, minimize its losses). The competitive firm's profits, π, are

$$\pi = pq - C(q),$$

where p is price, q is output, and $C(q)$ is total cost. As a result of the price-taking assumption above, the firm can sell all it wants at price p. (For example, the firm is too small a part of the market to influence the market price). That is, the firm faces a horizontal demand curve at price p.

It is profitable for the firm to expand output as long as the extra revenue from selling an additional unit exceeds the extra cost of producing that unit. The extra revenue from selling an additional unit is *price,* and the extra cost is the *marginal cost* (MC). That is, the optimal (profit-maximizing) production rule for a competitive firm is to expand its output until its marginal cost, MC, equals price, p.

Figure 3.1 illustrates the profit-maximizing decisions of a competitive firm facing a market price p_0. If the firm were producing a quantity greater than q_0, then p_0 would be less than MC, and the firm could increase its profits by reducing its output. If the firm were producing less than q_0, then p_0 would be greater than MC, and the firm could increase its profits by expanding its output. At output q_0, p equals MC, and profits are maximized.[2] In Figure 3.1, the shaded box represents profits.[3]

[1] We could derive the result that firms are price takers from these other assumptions. We make price taking an assumption for simplicity of presentation.

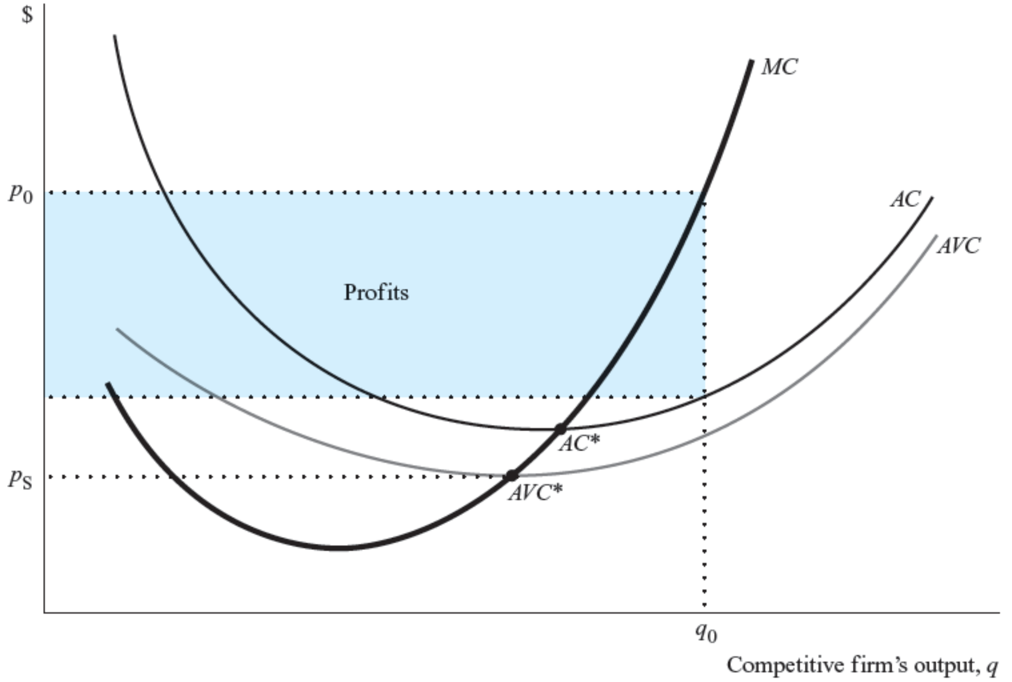
[2] The firm's objective is to

$$\max_q \pi = pq - C(q).$$

Its first-order condition is found by differentiating π with respect to q and setting that equal to zero: $p - C'(q) = 0$, where $C'(q) = \mathrm{d}C(q)/\mathrm{d}q$ is MC. This first-order condition—price equals marginal cost—is a necessary condition for profit maximization. The second-order condition is $C''(q) < 0$. That is, the second-order condition, which is a sufficient condition for profit maximization, is that the MC be upward sloping at the equilibrium.

[3] A firm's profits are total revenue minus total cost: $\pi = pq - C$, where pq (price times quantity) is total revenue. We can rewrite profits as average profits per unit (average revenue, p, minus average cost, $AC = C/q$) times the number of units sold (q), or $\pi = (p - AC)q$. Thus, profits can be shown graphically as a box with a height equal to average profits per unit, $p - AC$, and a length equal to the number of units, q, the firm sells.

FIGURE 3.1 Cost Curves and Profit Maximization



If the price, p, rises above p_0, the firm earns higher profits at its current output, but it earns even higher profits if it expands output until $p = MC$. If price falls below p_0, the firm earns lower profits at q_0, but it suffers less of a reduction if it reduces its output until $p = MC$. Thus, as the price rises, the firm moves up its marginal cost curve, and its profits rise; as the price falls, the firm moves down its marginal cost curve to minimize the reduction in its profits. Increases and decreases in profits signal a firm either to expand or contract output respectively.

Shutdown Decision. A firm produces only if doing so is more profitable than not producing. It produces only if the revenues from producing exceed **avoidable costs**: the costs that are not incurred if a firm ceases production. The revenues earned in excess of avoidable cost are called **quasi-rents**, which are the payments above the minimum amount necessary to keep a firm operating in the short run.

For simplicity, assume that all fixed costs are sunk. An example of a sunk cost given in Chapter 2 is that a firm is not refunded its incorporation fee if it ceases operating. In this case, avoidable costs are the same as variable costs. Thus, the rule for deciding whether to remain in business is: Produce and sell only if revenues are at least as great as total variable cost. Equivalently, the firm should produce and sell at price p only if p equals or exceeds average variable cost (AVC).

Minimum average cost (the lowest point on the *AC* curve), AC^*, is greater than minimum *AVC*, AVC^*, in the short run, because average costs are average variable costs plus average fixed costs. Thus, a firm finds it more profitable to produce than to shut down if price is below minimum average cost, $p < AC^*$, but above minimum average variable cost, $p > AVC^*$. It is more profitable to produce and earn some revenue in excess of variable cost than to shut down and earn no revenues (which can help offset the fixed costs). That is, the firm chooses to produce even though it is losing money when all costs are considered. Consider an example to clarify this apparent contradiction.

Suppose a firm's fixed cost is $200 and sunk. Its marginal cost (*MC*) is constant at $10 at quantities less than 100 units. At more than 100 units, *MC* is extremely high. If the price is $10, the firm produces and sells 100 units. The firm just covers its production cost and makes no contribution to the $200 fixed cost: It loses $200.

If the price is $9, the firm is better off not producing at all, because it loses an additional $1 for every unit it produces and would lose $300 if it produced 100 units. It is better to shut down and lose only $200 than to produce and suffer greater losses.

If the price is $11, by producing 100 units, the firm now more than covers variable cost: It earns $100 above variable cost. It still loses money overall ($-200 + 100 = -100$) because of the fixed cost of $200, but it is better to lose $100 than $200. The point of the example is that the decision to produce or not is *independent* of the fixed sunk cost. If fixed costs are sunk (incurred whether the firm produces or not), they should be ignored in deciding whether to produce.

If all fixed costs are sunk, a firm operates if *p* is greater than or equal to AVC^* but not if *p* is less than AVC^*. The price at which a firm ceases production is the **shutdown point**, which is p_s in Figure 3.1. That is, if price exceeds AVC^*, the firm operates along its *MC* curve. The **firm's supply curve** reflects the quantity that a firm is willing to supply at any given price. The competitive firm's supply curve, then, is the portion of the *MC* curve above AVC^*, the shutdown point.

If a firm suffers losses in the short run (the period in which costs are sunk), should it continue to operate and remain unprofitable in the future?[4] No. In the long run, a firm that is losing money will not reinvest—it will not continue to sink costs. Short-run losses are a signal that the firm should not invest further to replace plant and equipment. In the long run, a rational firm shuts down if it expects to have losses in each period forever. It prefers to cease production rather than invest in new facilities or maintenance and lose even more.

When a firm loses in the short run, its revenues are below the long-run opportunity cost of its resources. Because opportunity cost includes a normal profit, a firm that is making a loss may not literally be paying out more money than it receives; it is simply earning less than it could have earned had it invested its (already) sunk costs elsewhere.

[4] As described in Chapter 2, the short run and the long run are useful abstractions, but in reality adjustment costs determine how fast an industry can adjust to change. The time needed to adjust to any change depends on the current state of the industry and the size of the needed adjustment. See **www.aw-bc.com/carlton_perloff** "Adjustment Costs."

If fixed costs are not sunk, the shutdown decision depends on whether revenues exceed *avoidable* costs. An example (in Chapter 2) of an avoidable cost is the lawyer who can pay a penalty to break a lease. If some fixed costs are avoidable, a price equal to *AVC** is not high enough to prevent the firm from shutting down. Use the numbers from above and suppose that the fixed cost of \$200 represents a yearly rental payment and that, for a \$100 penalty fee, the landlord will release the lawyer from the obligation to pay \$200. The firm compares losing \$100 for sure (the penalty fee) with producing and earning revenues minus production costs minus the \$200 rental payment. If price is \$10, the firm earns \$0 per sale and is stuck paying the \$200 of fixed cost; therefore, it prefers to pay the \$100 penalty and go out of business. Even if price were \$10.50 so that the firm would make 50¢ on each of its 100 units sold, it would still be better to pay the \$100 penalty and go out of business.

The price at which shutdown occurs is above average variable cost and closer to average cost the greater the proportion of fixed costs that are avoidable. In the extreme, when there are no sunk costs (all fixed costs are avoidable), the shutdown point coincides with the minimum point on the *AC* curve. Thus, if it has no sunk costs, a firm shuts down before it incurs economic losses.

The Competitive Market

Given the behavior of individual competitive firms, we can derive a market supply curve. The intersection of the market supply curve and the market demand curve determines the competitive equilibrium.

The Short-Run Equilibrium. We start by supposing that there are n identical firms and that all fixed costs are sunk in the short run. The short-run **market supply curve**, S in Figure 3.2b, is the horizontal sum of the supply curves of each firm, the *MC* curve above the minimum of the *AVC* curve in Figure 3.2a. The horizontal portion of the market supply curve reflects (1) that no output is forthcoming if price is below the shutdown point and (2) that at a price slightly above the shutdown point, all firms produce.

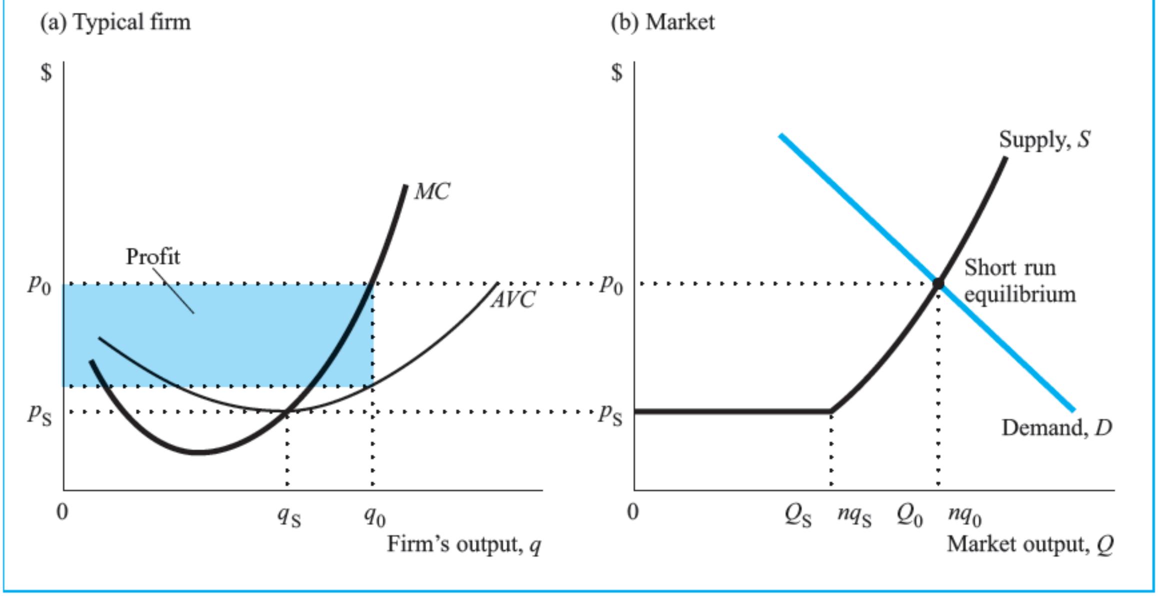
The intersection of the demand curve with the short-run market supply curve determines the *short-run equilibrium price,* p_0, and quantity, Q_0. The amount that firms want to supply at the equilibrium price exactly equals the amount that consumers demand at that price. There are no unsatisfied buyers and no unsatisfied sellers. All buyers pay and all sellers receive the same price.

In the short-run equilibrium in Figure 3.2, a typical firm may earn a profit, which provides an incentive for firms to enter the market. However, such entry cannot occur in the short run because firms cannot build new plants in the short run.

The Long-Run Equilibrium. In the long run, firms can adjust their levels of capital so that they can enter this market. Short-run profits or losses induce firms to enter or leave the market until price is driven to the minimum long-run average cost, *AC**, in the long run.

In Figure 3.2, firms are making a positive profit at the short-run equilibrium price p_0, which is determined by the intersection of the market demand curve and the original short-run market supply curve. In the long run, these profits induce new firms to

FIGURE 3.2 Short-Run Equilibrium



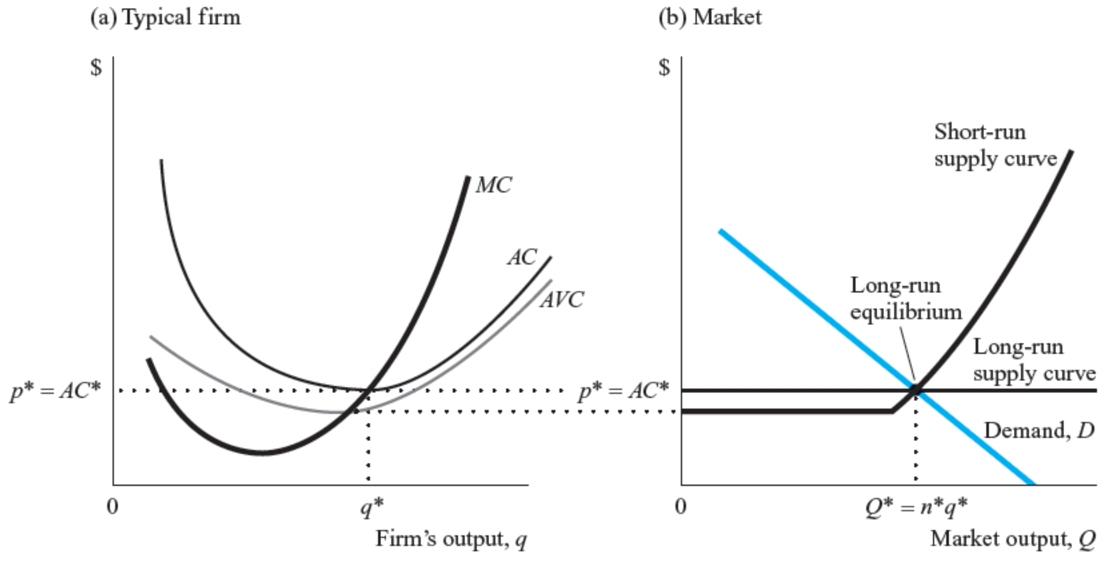
enter this market. If the number of firms that can potentially produce at the same cost is very large, the long-run supply curve is horizontal at the minimum of the average cost curve, AC^*, as Figure 3.3 shows. The long-run equilibrium is determined by the intersection of the demand curve and the long-run market supply curve. In Figure 3.3, the market is in a new short-run *and* long-run equilibrium because the demand curve, D, intersects both the long-run supply curve and the new short-run supply curve corresponding to the equilibrium number of firms, n^*. The equilibrium price is $p^* = AC^*$, and equilibrium output is $Q^* = n^*q^*$. In this long-run equilibrium, firms make zero profit.

Similarly, short-run losses induce firms to leave the market and reduce output until price rises again to yield normal (zero) profits. In long-run equilibrium, firms receive economic profits of zero, which is just enough to induce them to remain in the market.

The Slope of the Long-Run Supply Curve. In this last example, a very large number of firms could enter the market and produce at the same marginal and average costs as the existing firms. Consequently, the long-run perfectly competitive supply curve was perfectly flat at AC^*, which is the minimum average cost of production. However, the long-run supply curve need not be flat.

If an expansion of output causes the prices of some key inputs to rise, the long-run supply curve tends to be upward sloping. As the output of wheat produced increases, farmland becomes more valuable, and the land rents (or the opportunity cost of owning the land) increase. As rents increase, the average cost curve of each farmer rises so the minimum average cost, AC^*, increases. Thus, the long-run supply curve for the

FIGURE 3.3 Long-Run Equilibrium



wheat market (whose height is traced out by the minimum average cost points) rises as output expands.

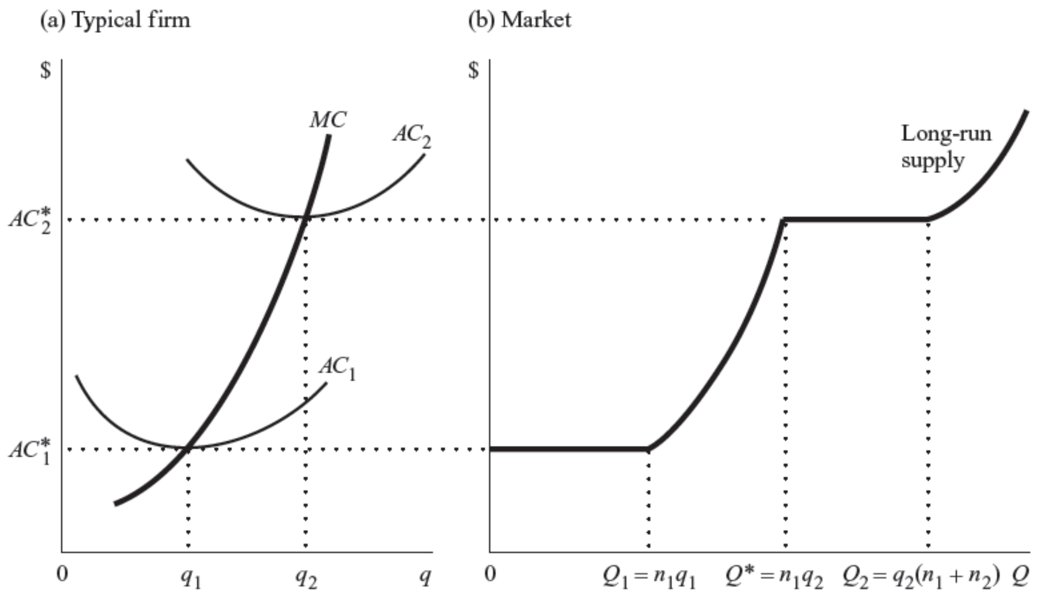
Whenever some factors of production (such as fertile land) are in fixed supply, their price gets bid up as market output expands. Prices of key inputs may fall as output expands if there are economies of scale. If input prices fall as output expands, the long-run market supply curve could slope down. The long-run supply curve of a market tends to be flat as long as the market accounts for only a small fraction of any one factor's total employment.

Another reason why the long-run supply curve may be upward sloping is that there are only a few firms that can produce at low costs. For market output to increase, less efficient firms have to enter the market. In Figure 3.4a, there are n_1 low-cost, efficient firms with marginal cost curve MC and average cost curve AC_1. The minimum point on AC_1 is AC^*_1, which is obtained if the firm produces q_1 units of output. For market output levels up to $Q_1 = n_1 q_1$, these low-cost firms can produce at the minimum average cost, AC^*_1, so the long-run supply curve is flat at AC^*_1 up to Q_1, as Figure 3.4b shows. If less than Q_1 is demanded, some of these n_1 firms exit the market.

If the market demand is slightly larger than Q_1, the average cost of production must rise. The market supply curve is the horizontal sum of the supply curves of the n_1 firms: Their marginal cost curves above AC^*_1. Thus, because there are no more low-cost firms, the market supply curve rises beyond Q_1.

Now suppose that there are n_2 other firms that can produce this product with the same marginal cost curve as the first n_1 firms but with an average cost curve, AC_2, with a higher minimum average cost, AC^*_2 ($> AC^*_1$). That is, these high-cost firms have larger fixed costs than do the low-cost firms.

FIGURE 3.4 Upward-Sloping Long-Run Supply Curve



If the quantity demanded is slightly greater than $Q^* = n_1q_2$, the price is AC^*_2 and some high-cost firms enter the market. Increases in market demand beyond this point are met by additional high-cost firms entering the market and producing q_2 at an average cost of AC^*_2. When the quantity demanded can no longer be met by entry by additional high-cost firms, the long-run supply curve again rises, tracing out the sum of the marginal cost curves of all the firms in the market. That is, the long-run supply curve rises for output greater than $Q_2 = Q^* + n_2q_2 = q_2(n_1 + n_2)$.

If the quantity demanded exceeds n_1q_1 but is less than Q^* (that is, the second group of firms has not entered the market), the low-cost firms earn an unusual return (profit) to their scarce knowledge or other scarce resource that enables them to produce at relatively low costs. That is, they earn a rent: a payment to the owner of an input beyond the minimum necessary to cause it to be used. If the quantity demanded exceeds Q_2, both types of firms earn rents on their scarce know-how or other scarce input.[5]

[5] In some markets where firms must incur substantial sunk costs so that only a few firms can efficiently produce, a competitive equilibrium may be impossible. In such markets where no competitive equilibrium exists, the market exhibits instability, including price wars and bankruptcies. Here, firms may temporarily make above-normal profits; these profits attract other firms, the additional competition results in all the firms in the market making losses, some firms exit, the remaining firms make above-normal profits, and the process repeats. The study of when interactions between firms and consumers will lead to stability is called the theory of the core. See Clark (1923), Telser (1978), and **www.aw-bc.com/carlton_perloff** "Nonexistence of Competitive Equilibrium."

Elasticities and the Residual Demand Curve

Throughout the rest of this book, we make repeated use of two related concepts to analyze both competitive and noncompetitive industries: (1) the price elasticity of demand or supply, and (2) the demand curve facing a single firm, which is called the firm's *residual demand curve*. The price elasticity of supply or demand aids in understanding how a market responds to changes in either demand or supply. The residual demand facing a single firm allows an analyst to comprehend the behavior of a single firm. We now examine how the elasticity of the residual demand curve is related to the assumption that a competitive firm cannot affect price.

Elasticities of Demand and Supply

If either the demand or the supply curve shifts, the competitive equilibrium changes, and the shapes of the demand and supply curves influence how the new equilibrium compares to the old. For example, if the demand curve is perfectly flat, the competitive price remains unchanged, even if the supply curve shifts radically.

One concept used to characterize the shape of demand or supply curves is the price elasticity of demand or supply (often the word *price* is omitted). The **elasticity of demand** is the percentage change in quantity demanded in response to a given small percentage change in price.[6] Similarly, the **elasticity of supply** is the percentage change in quantity supplied in response to a given small percentage change in the price. The elasticity of demand is always a negative number, and the elasticity of supply is usually, but not always, positive.

If a 1 percent increase in price leads to a more than a 1 percent reduction in the quantity demanded (so that the total amount paid in the market falls), a demand curve is called **elastic**. That is, an elastic demand curve has an absolute value of the elasticity of demand greater than 1 (the absolute values of 1 and −1 are both 1). It is common to omit the phrase *the absolute value of* when discussing the price elasticity of demand. The statement, "The price elasticity of demand is 2," is interpreted to mean that the price elasticity is −2.

When the absolute value of the elasticity of demand is 1, the demand curve is said to have **unitary elasticity**. In that case, a 1 percent change in price causes a 1 percent change in the quantity demanded, and the total amount paid (total revenues) remains constant. If the absolute value of the elasticity of demand is less than 1, the demand curve is **inelastic**: A 1 percent increase in price causes less than a 1 percent decline in the quantity demanded, and the total amount paid rises.

[6] The price elasticity of demand at price p and quantity Q is the percentage change in quantity divided by the percentage change in price (if that change is small): $(\Delta Q/Q)/(\Delta p/p) = (p/Q)(\Delta Q/\Delta p)$. Because the elasticity is the ratio of two percentage terms, the elasticity is invariant to changes in scale of either price or quantity (it is a *pure* number—without scale itself). For example, if price is measured in cents rather than in dollars, the elasticity is unchanged even though the slope of the demand curve, $\Delta Q/\Delta p$, does change. The technical definition of the price elasticity is $(p/Q)(\mathrm{d}Q/\mathrm{d}p)$.

In general, the elasticities of demand and supply depend upon many economic factors, such as the level of output, the availability of substitute products, and the ease with which suppliers can alter production. For example, as more substitute products are available, consumers find it easier to substitute for a product if its price rises, which makes its demand curve more elastic. Similarly, the more flexible the production process of a firm, the more likely it is that the firm can greatly increase production in response to a price increase, which tends to increase the elasticity of supply.

The Residual Demand Curve of Price Takers

Competitive firms are often described as *price takers*. They believe that they cannot affect the market price and must accept, or take, it as given. There are three equivalent ways to describe a firm's inability to affect price, all of which are used in this chapter:

- A competitive firm is a price taker.
- The demand curve facing a competitive firm is horizontal at the market price.
- The elasticity of demand facing a competitive firm is infinite.

A firm is a price taker if it faces a horizontal demand curve, because a horizontal demand curve has an infinite price elasticity of demand. If a firm facing an infinite price elasticity raises its price even slightly, it loses all its sales. Equivalently, by lowering its quantity, the firm cannot cause the price to rise. In contrast, a firm facing a downward-sloping demand curve can raise its price by decreasing its output.

If the number of firms in a market is large, the demand curve facing any one firm is nearly horizontal (elasticity of demand is infinite) even though the demand curve facing the market is downward sloping with a low elasticity. Indeed, for most market demand curves, there do not have to be very many firms in a market for the elasticity of demand facing a particular firm to be large.

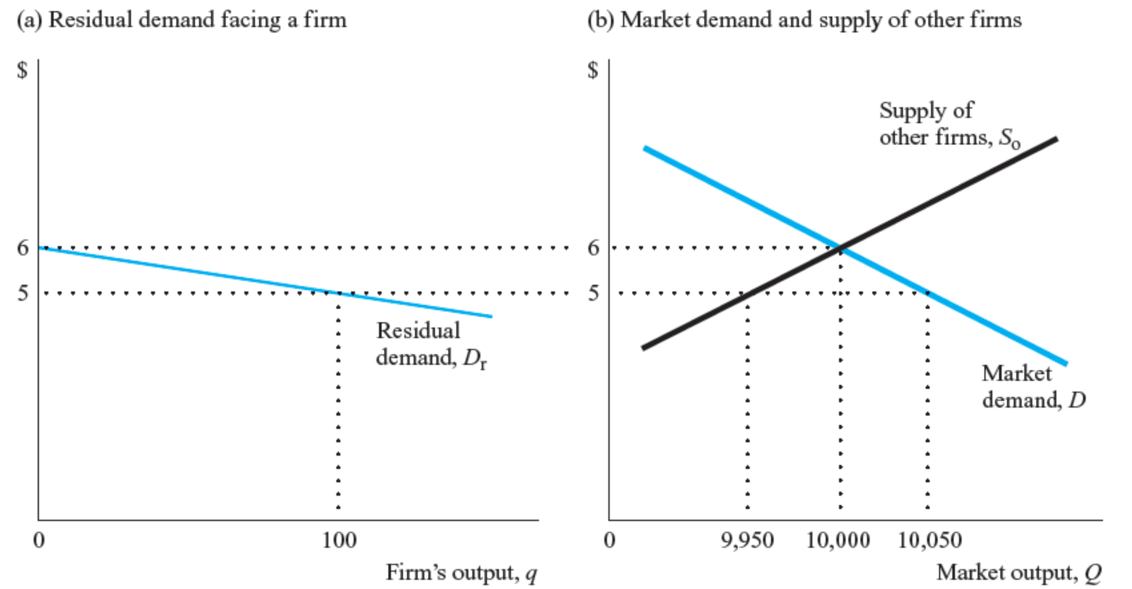
To show this result, it is necessary to determine the demand curve facing a particular firm: the **residual demand curve**. A firm sells to people whose demands are not met by the other firms in the market. For positive quantities of residual demand, the residual demand, $D_r(p)$, is the market demand, $D(p)$, minus the supply of other firms, $S_o(p)$:

$$D_r(p) = D(p) - S_o(p).$$

If $S_o(p)$ is greater than $D(p)$, $D_r(p)$ is zero.

Figure 3.5b shows the market demand curve and the supply curve of all the firms except one. Figure 3.5a shows the residual demand curve facing a particular firm, which is the horizontal difference between the quantity demanded by the market at a given price minus the supply of other firms at that price. For example, at a price of $5, market demand is 10,050 units and the supply of the other firms is 9,950 units in Figure 3.5b. Thus, market demand exceeds their supply by 100 units at $5, so the remaining firm faces a residual demand of 100 units at that price.

FIGURE 3.5 Derivation of Residual Demand Curve



At $6, the supply of other firms equals the market demand. The residual demand facing the firm in Figure 3.5a is zero. Were the price to rise even higher, other firms would be willing to supply even more than is demanded. Thus, at any price greater than or equal to $6, the firm in panel a sells no units.

The residual demand curve facing the firm, Figure 3.5a, is much flatter than the market demand curve in Figure 3.5b. Similarly, the single firm's demand elasticity is much higher than the market elasticity. For example, the elasticity of demand for the individual firm is −11 at $5.50, whereas the corresponding market elasticity of demand is approximately −0.027.[7] In other words, the firm's residual demand curve is over 400 times as elastic as the market demand curve at this price.

More generally, if there are n identical firms in the market, then the elasticity of demand facing Firm i is

$$\epsilon_i = \epsilon n - \eta_o(n - 1), \qquad (3.1)$$

[7] For a linear demand curve, $Q = a - bp$, the elasticity of demand is the slope of the demand curve, $dQ/dp = -b$, times p/Q. Thus, the elasticity of demand of the residual demand curve is $-100(5.50/50) = -11$ and the elasticity of demand of the market demand curve is $-50(5.50/10{,}025) \approx -0.027$.

where ϵ is the market elasticity of demand (a negative number), η_o is the elasticity of supply of the other firms (a positive number), and $(n - 1)$ is the number of other firms.[8]

Thus, for a given market elasticity, as the number of firms in a market, *n*, increases, the elasticity facing a single Firm *i*, ϵ_i, grows large in absolute value (more negative). Similarly, the larger the elasticity of supply of the other firms, η_o, or the more other firms there are, the larger in absolute value (more negative) is the elasticity of demand facing Firm *i*.

Table 3.1 shows how the elasticity of demand facing a single firm varies with the number of firms and the market elasticities, given that the supply of other firms is completely inelastic ($\eta_o = 0$). For example, if the market elasticity is unitary ($\epsilon = -1$) and there are 50 firms, then $\epsilon_i = -50$. That is, if a firm were to increase its price by 1 percent, the quantity it sells would fall by about 50 percent. If the market demand elasticity is -0.5 and there are 1,000 firms, $\epsilon_i = -500$, so that if the firm were to raise its price by a tenth of a percent, the quantity it sells would fall by 50 percent. Thus, even if the supply of the other firms is completely inelastic, if there are a large number

TABLE 3.1 **Price Elasticity for a Single Firm**

	Market Elasticity		
Number of Firms *n*	Inelastic $\epsilon = -0.5$	Unitary $\epsilon = -1$	Elastic $\epsilon = -5$
10	−5	−10	−50
25	−12.5	−25	−125
50	−25	−50	−250
100	−50	−100	−500
500	−250	−500	−2,500
1,000	−500	−1,000	−5,000

Note: Because the supply of the other identical firms is assumed to be perfectly inelastic ($\eta_o = 0$), the elasticity of demand facing a particular firm is $\epsilon_i = n\epsilon$.

[8] The residual demand curve facing any one firm is $D_r(p) = D(p) - S_o(p)$. Differentiating $D_r(p)$ with respect to *p*, we obtain

$$\frac{dD_r}{dp} = \frac{dD}{dp} - \frac{dS_o}{dp}.$$

Let the quantity produced by one firm be $q = Q/n$ and the total quantity produced by all the other firms be $Q_o = (n - 1)q$. Multiplying both sides of the expression above by p/q and multiplying and dividing the first term on the right-hand side by Q/Q and the second term by Q_o/Q_o, this expression is

$$\frac{dD_r}{dp}\frac{p}{q} = \frac{dD}{dp}\frac{p}{Q}\frac{Q}{q} - \frac{dS_o}{dp}\frac{p}{Q_o}\frac{Q_o}{q},$$

where $q = D_r(p)$, $Q = D(p)$, and $Q_o = S_o(p)$. This expression is rewritten as $\epsilon_i = \epsilon n - \eta_o(n - 1)$ in Equation 3.1.

EXAMPLE 3.1 *Are Farmers Price Takers?*

In most U.S. agricultural markets there are a large number of farms, and no farm has as much as even 1 percent of total sales. As a result, the elasticity of demand facing each farm is extremely large. Farms are price takers.

We can roughly calculate the residual demand price elasticity facing an individual farm. For simplicity, we assume that other farms have an inelastic supply ($\eta_o = 0$), which may be a reasonable assumption in the short run. Less accurately, we assume that all farms are approximately the same size, so that each farm's share of the market is equal to 1 divided by the number of farms. The following table shows the approximate elasticity of demand facing each farm.

Crop	Estimated Market Demand Elasticity	Number of Farms	Each Farm's Residual Demand Elasticity
Fruits			
apples	−.20	28,160	−5,620
grapes	−1.03	19,961	−20,560
peaches	−.82	14,459	−11,856
Vegetables			
asparagus	−.65	2,672	−11,140
cucumbers	−.30	6,821	−2,046
dry onions	−.16	3,296	−527
sweet peppers	−.25	6,271	−1,568
tomatoes	−.38	14,366	−5,459

Sources: Number of Farms: U.S. Department of Commerce, Bureau of the Census, 1997 Census of Agriculture; Survey of Elasticities: You; Epperson, and Huang (1998).

Thus, each farm faces a gigantic price elasticity. For example, were a grape farm to increase its price by as little as 0.001 percent (one-thousandth of one percent), the quantity demanded from the farm would fall by 21 percent. Each farm is a price taker.

of firms in the market, the elasticity of demand facing a single firm is very large, as Example 3.1 illustrates.

Efficiency and Welfare

The welfare of the people is the chief law. —*Cicero*

The competitive equilibrium has desirable efficiency and welfare properties. Indeed, no one can be made better off without making someone else worse off in the competitive equilibrium.

Efficiency

The competitive equilibrium of price and quantity has two desirable efficiency properties. First, production is efficient in the sense that there is no possible rearrangement of resources (such as labor, machines, and raw materials) among firms that can increase the output of one product without reducing the output of at least one other product.

Second, consumption is efficient. The value that a buyer places on consuming the good is exactly equal to the marginal cost of producing that good (remember, the competitive price equals the marginal cost of production). Moreover, no rearrangement of goods among consumers can benefit a consumer without harming at least one other consumer.

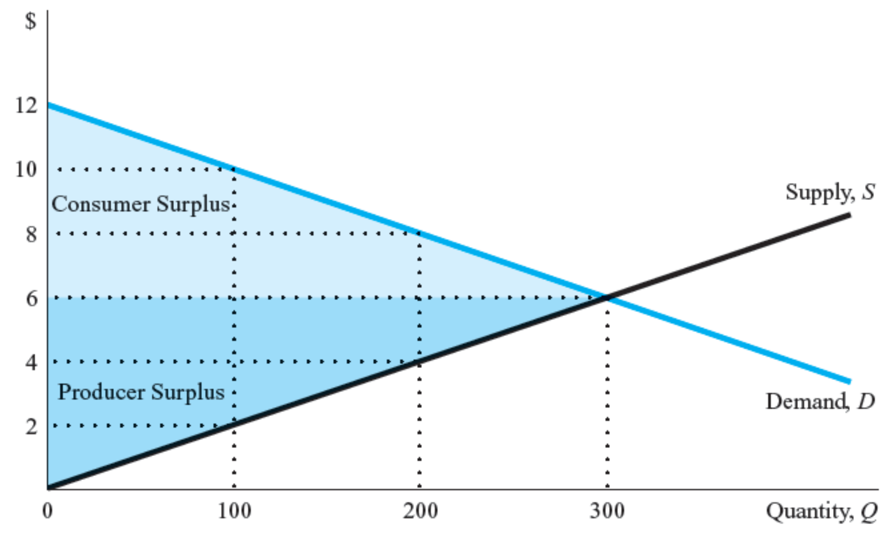
Welfare

We now describe a common measure of welfare, show that competition maximizes this measure of welfare for any given distribution of income, and illustrate that departures from competition lower welfare. In particular, in the next section, we show that restrictions that prevent firms from entering a market lower welfare.

Consumer Surplus. Typically, consumers value the goods they purchase above the amount they actually pay for them. Consumer surplus is the amount above the price paid that a consumer would willingly spend, if necessary, to consume the units purchased.

A good's demand curve reflects the value that consumers place on consuming additional units of a good. For example, the demand curve in Figure 3.6 indicates that

FIGURE 3.6 Consumer Surplus and Producer Surplus



consumers would pay $10 for 100 units of the good, $8 for 200 units, and $6 for 300 units.

In the competitive equilibrium in Figure 3.6, consumers pay $6 for 300 units. They would have been willing to spend $4 more for the first 100 units, $2 more for the first 200 units, and no extra amount for 300 units. The total consumer surplus is the shaded area below the demand curve and above the equilibrium price of $6 up to the equilibrium quantity of 300 units.[9] This area equals $900 (= [$12 − $6] × 300/2).

In the competitive market, consumers paid $1,800 to buy the 300 units. In this example, consumer surplus is 50 percent of the amount they actually pay. If consumers could have had the choice of buying 300 units or none, they would have been willing to spend up to $2,700 (the $1,800 they spent, plus the extra $900 in consumer surplus) to purchase the 300 units.

Producer Surplus. Similarly, firms may receive more for the goods they sell than it costs them to produce those goods. **Producer surplus** is the largest amount that could be subtracted from a supplier's revenues and yet the supplier would still willingly produce the product.

We can use information from the supply curve to calculate firms' producer surplus. A supply curve represents the marginal cost of producing output. For example, in Figure 3.6, it costs firms $2 to produce 100 units, $4 to produce 200 units, and $6 to produce 300 units. The producer surplus is the area above the supply curve and below the market price up to the quantity sold. The producer surplus is the area equal to $900, which is above the supply curve and below the price of $6 up to 300 units. That is, firms would be willing to pay $900 for the right to sell 300 units of the good at $6 rather than selling none at all.

Welfare. One common measure of welfare from a market is the sum of consumer surplus and producer surplus. This measure of welfare is the value that consumers and producers would be willing to pay to purchase the equilibrium quantity of output at the equilibrium price.

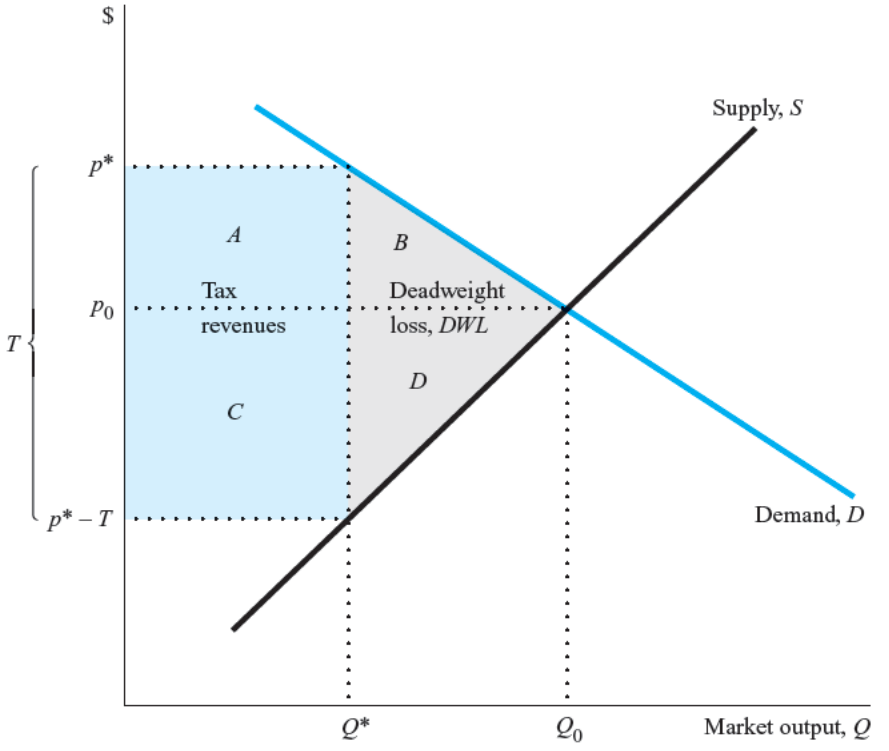
Figure 3.6 illustrates that this measure of welfare is maximized at the competitive equilibrium. For example, if fewer units were produced, welfare would fall, as we now show.

Deadweight Loss. The cost to society of a market's not operating efficiently is called **deadweight loss** (*DWL*). It is the welfare loss—the sum of the consumer surplus and producer surplus lost—from a deviation from the competitive equilibrium.

For example, the competitive equilibrium is at price p_0 and quantity Q_0 in Figure 3.7. At Q_0, the value that a consumer places on additional consumption equals the marginal cost of producing the good. If the government taxes this good or restricts its

[9] Consumer surplus is an accurate measure of consumer well-being if there are no income effects (a change in a consumer's income leaves demand unchanged). Even when there are income effects, changes in consumer surplus can provide a close approximation to changes in welfare (Willig 1976).

FIGURE 3.7 Deadweight Loss from a Tax



sale, this link between the value the consumer places on an additional unit and the cost of producing it is broken, which lowers welfare.

Suppose that the government collects a tax of T per unit of goods sold. If a customer pays p, the government takes T of that, and the firm receives $p - T$. Thus, the tax creates a wedge of T between the value that the marginal demander places on the good (as shown by the demand curve) and the cost that the marginal supplier is willing to incur to produce the good (as shown by the supply curve). The imposition of the tax reduces the quantity sold from Q_0 to Q^*. The price consumers pay rises to p^*, and the price firms receive falls to $p^* - T$.

In this after-tax equilibrium, the amount sold, Q^*, is less than in the competitive equilibrium, Q_0, and the value consumers place on consuming an additional unit, p^*, now exceeds the marginal cost of producing it by T. Consumers suffer a loss of consumer surplus equal to areas A and B (the area between p^* and p_0, to the left of the demand curve). Suppliers suffer a loss of producer surplus equal to areas C and D (the area between p_0 and $p^* - T$ to the left of the supply curve). The government receives tax revenues equal to TQ^*, boxes A and C. Thus, the transfer from consumers and producers to the government (the tax revenues = boxes *A* and *C*) is less than the combined loss of the consumers and producers. This extra cost to society due to reduced output is the deadweight loss, which equals the sum of the triangles *B* and *D* in Figure 3.7.[10]

The deadweight loss triangle is the total loss to society if the government makes good use of the tax revenues. The *DWL* triangle is an efficiency loss because the marginal cost of producing a good is less than the marginal willingness of consumers to pay for it.

As long as the government makes efficient use of this money, the tax revenues are not an efficiency loss. Rather, the tax revenues reflect a redistribution of income from buyers and sellers of this good to those who benefit from the government's use of these funds.

Entry and Exit

As we show throughout the rest of this book, ease of entry and exit plays a critical role in determining market structure and the subsequent performance of firms. If firms that are as efficient as those already in the market cannot easily enter, existing firms may be able to exercise market power by setting prices above marginal cost.

Restrictions on Entry

In many industries, governments or groups of firms collectively set licensing requirements that restrict entry (see Example 3.2). An example of an entry restriction is the limit on the number of taxicabs allowed in many cities throughout the world. Such entry restrictions elevate prices above competitive levels.

Figure 3.8 illustrates how a restriction on entry leads to a price above the long-run competitive equilibrium price. In this market, a large number of firms could produce with identical cost curves, as panel a shows.

Figure 3.8b shows two long-run supply curves for a market where all firms have identical costs. In the absence of a government restriction on entry, there are 150 firms in this market. The competitive equilibrium is determined by the intersection of the supply curve for 150 firms and the market demand curve. The equilibrium price is p_0, and each firm is producing at the minimum of its long-run average cost curve, AC^*.

[10] The deadweight loss triangle can be expressed in terms of the elasticities of demand and supply. For simplicity, assume that the supply curve is perfectly horizontal (infinitely elastic). Then the deadweight loss triangle = $-1/2\Delta p\Delta Q$, where $\Delta p = p^* - p_0$ and $\Delta Q = Q^* - Q_0$. The elasticity of demand, ϵ, is (approximately) equal to $(\Delta Q/Q_0)(p_0/\Delta p)$. Define t as $\Delta p/p_0$, which is the percentage change in the price due to the tax. The deadweight loss triangle equals

$$-\frac{1}{2}\Delta p\Delta Q \approx -\frac{1}{2}\frac{\Delta p}{p_0}p_0Q_0\left(\frac{\Delta Q}{Q_0}\frac{p_0}{\Delta p}\right)\frac{\Delta p}{p_0} \approx -\frac{1}{2}t^2R\epsilon,$$

where *R* is the revenue, p_0Q_0, and ≈ means *approximately equal*. Thus, the deadweight loss depends on the size of the market, *R*, in addition to *t* and ϵ.

EXAMPLE 3.2 *Restrictions on Entry Across Countries*

Most countries restrict new businesses from entering. Virtually every country requires that a potential new firm fill out certain forms and pay fees to become a legal business. Some countries prohibit entry in certain industries.

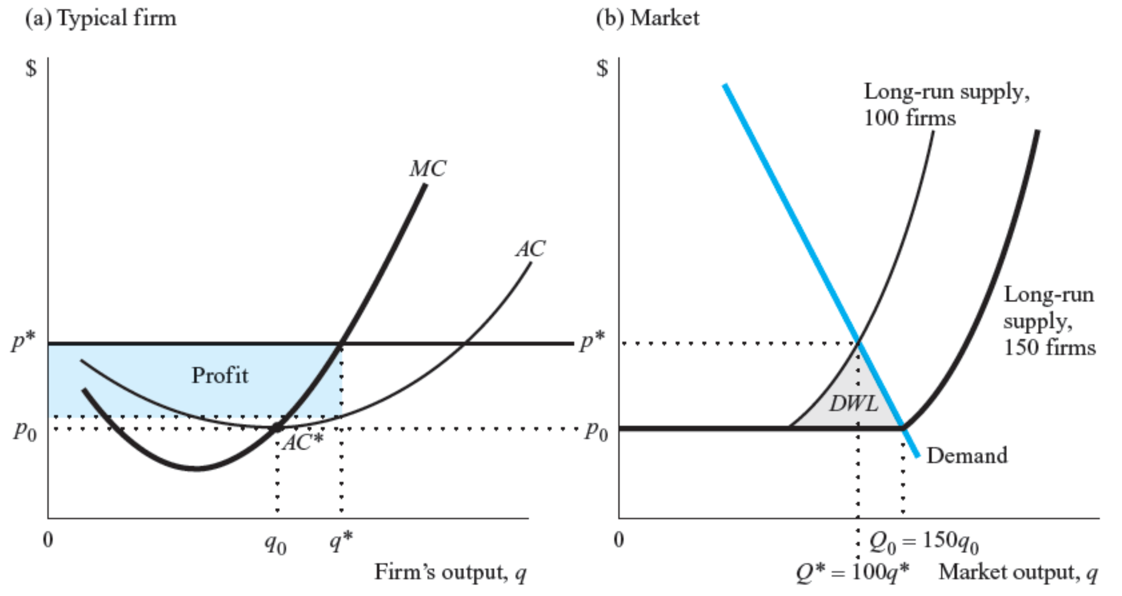
A World Bank survey of entry restrictions in 85 countries found that the ease of entry varied substantially across countries. It takes two business days to enter a typical business in Australia or Canada, but 152 days in Madagascar. The average for the 85 countries surveyed was 47 days.

To determine the cost of entry, the researchers calculated the ratio of fees plus the cost of time in applying as a percentage of per capita annual gross domestic product (GDP). Again, the range across countries was enormous: The lowest ratio was less than 0.5 percent for the United States, the highest was more than 4.6 times per capita GDP in the Dominican Republic, and the average was 47 percent of per capita GDP.

Are there any patterns to help explain which countries have the most onerous restrictions? Yes. In general, rich countries have fewer restrictions than poor countries. Countries with greater political freedoms, less corruption, and smaller illegal sectors tend to have fewer entry restrictions. Leaders of governments in poor, underdeveloped countries generally set rules that protect existing businesses from competition, perhaps benefiting their friends, their relatives, or themselves, all of whom may have business interests.

Source: Djankov et al. (2002).

FIGURE 3.8 Long-Run Equilibrium with an Entry Restriction



If the government restricts the number of firms in the market to 100, the long-run supply curve lies to the left of the original one. With this restriction on entry, the new equilibrium price is p^*. The entry restriction, therefore, results in consumers' paying a price, p^*, higher than the unrestricted competitive price, p_0, and consuming only Q^*, which is less than the unrestricted competitive quantity, Q_0.

The shaded area *DWL* in Figure 3.8b is the lost welfare from restricting entry. It reflects the loss of consumer surplus from paying p^* rather than p_0 that is not captured by the firms.

The entry restriction is inefficient for two reasons. First, there is a loss in efficiency due to restricting output from Q_0 to Q^*. Second, the average cost of production is greater with entry restrictions. With free entry, each firm produces q_0 and the average and marginal cost of production is p_0. With the restriction, firms produce q^* units at a marginal cost of p^* and an average cost above p_0 in Figure 3.8a. The area between the two supply curves to the left of Q^* in Figure 3.8b is a measure of this increased cost. The entire shaded area in Figure 3.8b is the deadweight loss caused by both sources of inefficiency from the entry restriction.

A firm that is among the 100 firms allowed into the market is better off than if there were no entry restrictions. The elevated price raises the profits of each of these 100 firms (the shaded area in Figure 3.8a) above the level that would have existed had the equilibrium number of firms, 150, been allowed to enter. With free entry, each firm produces at minimum average costs, and profits are zero. Thus, entry restrictions are like a tax on the consumption of a good. A tax, however, transfers money from consumers and producers to the government. In contrast, the entry restrictions transfer money from consumers to firms that were able to operate in this market. Consequently, the Federal Trade Commission (FTC) opposes many such barriers: see Example 3.3.

EXAMPLE 3.3 *FTC Opposes Internet Bans That Harm Competition*

Preventing Internet shopping may raise the price of some goods. In 2003, the Federal Trade Commission (FTC) issued a report concluding that removing bans on interstate wine sales over the Internet would save consumers as much as 21 percent on relatively expensive wines and increase consumer choice. In 26 states, including New York, Florida, Massachusetts, and Pennsylvania, laws (many dating from the Prohibition era) ban direct-to-consumer shipping from out of state, in part to prevent sales to minors. However, the FTC concluded that shipping wine directly to homes does not lead to more underage drinking. One reason is that many states require an adult to sign to accept wine deliveries.

The FTC has worked to reduce barriers to online trade in other markets. For example, they argued before Connecticut's Board of Examiners for Opticians against regulations making it difficult for online vendors to sell contact lenses to consumers. Similarly, they opposed barriers to online sales of caskets in Oklahoma.

Sources: Federal Trade Commission, **www.ftc.gov/os/2003/07/winereport2.pdf, www.ftc.gov/opa/2003/07/wine.htm**.

Competition with Few Firms—Contestability

In some markets, total output is small relative to the efficient size of a firm. In other words, the economies of scale in production and sales are important so that only one or a few firms can efficiently produce in the market. Even in such a setting, it is possible for competition to work (Demsetz 1968, Baumol, Panzar, and Willig 1982), although the process is different from the competitive markets we have been analyzing.

If many identical firms are capable of entering the market and producing, no firm is able to earn more than the normal level of profits in the long run. If there is free entry into and exit from a market instantaneously (no sunk costs), firms have an incentive to enter whenever price exceeds average cost. Markets with free instantaneous entry and exit are called perfectly *contestable* (Baumol, Panzar, and Willig 1982).

Residential garbage collection may be an example of a contestable market. There are economies of scale in providing residential garbage collection in a single town (see Example 20.4). It would be inefficient for more than one firm to traverse the same route to pick up garbage. A town can solicit bids from garbage collection firms and choose the lowest bidder. Competition among bidders ensures that the town is served at the lowest possible cost, even though only one firm actually provides the service.[11]

In contrast, in a market that is protected from entry, price remains above marginal cost because no firm can enter the market and drive down price. Thus, restrictions on entry are the reason that many markets are not perfectly competitive, so that prices are above marginal cost.

Definition of Barriers to Entry

Is entry difficult in most markets? To answer this question, we need to define what we mean by a *barrier to entry*. A common definition of a **barrier to entry** is anything that prevents an entrepreneur from instantaneously creating a new firm in a market. This definition may not be useful, however, as it implies that virtually every market has a barrier to entry. Under this definition, the cost of hiring labor or the cost of building a plant is an entry barrier. Moreover, it implies that any market in which entry takes time has a barrier to entry.

Unfortunately, the term *barrier to entry* is often used to refer to both the costs of entering and the time it takes for entry to occur. Because the term has several meanings, confusion has resulted sometimes, even though the proponents of the various definitions agree that higher costs of entry raise prices. See Carlton (2004).

[11] A complication arises with economies of scale. If the bidding competition drives profits to zero so that price equals average cost and if there are economies of scale, price *exceeds* marginal cost. If price does not equal marginal cost, the allocation of resources is inefficient. A more efficient way to finance garbage collection service is to charge a fee per pickup that reflects marginal cost, and to charge a separate fee for the right to have garbage picked up at all. Under this scheme, if residents want garbage picked up twice a week, they would pay the fixed fee plus twice the fee per pickup. Such two-part pricing schemes are studied in Chapter 10.

The economic theories discussed in this book predict the erosion of profits by entry only in the long run. Thus, one reasonable approach is to focus on long-run barriers to entry, which prevent new firms from entering a market even though an existing firm earns long-run profits.

If there are many firms that can enter with identical cost curves and face identical prices, then no one firm can succeed in the long run at earning profits that exceed costs without inducing additional entry. Only by having some advantage over new entrants can a firm earn persistently higher profits than other actual or potential firms. Because long-run profits can only persist if a firm has an advantage over potential entrants, a logical definition of a **long-run barrier to entry** is a cost that must be incurred by a new entrant that incumbents do not (or have not had to) bear.[12]

Entry Barriers. A good example of a long-run barrier to entry is a patent. Under most patent systems, the government grants an inventor the monopoly right to sell the invention for a fixed period of time. A patent creates a legal monopoly through a long-run barrier to entry. To compete against an incumbent firm with a patent, a potential entrant has to either invent around the patent or license it from the incumbent firm.[13] Because the incumbent firm has the right to exclude anyone from using the patent, it can prevent entry. The incumbent probably had to invest in research and development in order to acquire the patent. If the same avenue of research and development is not available once a patent has been granted, the potential entrant faces a cost that is greater than that of the incumbent.[14]

An incumbent may use a variety of strategies designed to raise the cost of entry, all of which require the incumbent to exploit some asymmetry between it and a potential entrant in order to raise the cost to a potential entrant above its own. When it is successful, the incumbent firm can create a long-run barrier to entry. These strategic responses are studied in detail in Chapter 11.

Exit Barriers. An important consideration in understanding a firm's incentive to enter a market is, paradoxically, the firm's ability to exit the market. If it is costly to exit a market, the incentives to enter are reduced. It is costly to exit a market if there are sunk costs that cannot be recovered. For example, suppose that a firm in a market must have very specialized equipment that is difficult to resell. A firm contemplating entry into that market realizes that if the unusual profit opportunities in the market are short lived, it may not pay to enter. In contrast, if there are no costs to enter or exit, then instantaneous entry and subsequent exit, sometimes called *hit-and-run entry,* by outside

[12] This definition is adapted from Stigler (1968a). See also von Weizsäcker (1980), who adds the condition that an entry barrier must lower consumer welfare. See McAfee, Mialon, and Williams (2004) for a discussion of various definitions of barriers to entry.

[13] As discussed in Chapter 16 on patents, the holder of a patent may allow other firms to use the invention in return for a license fee. That fee, however, is the barrier to entry, for it is the amount that an entrant must pay that the incumbent does not pay.

[14] As discussed in Chapter 16, patents have an offsetting benefit. Were it not for the unusually high profits a patent holder obtains from its monopoly, firms would not sink large amounts into the research and development that leads to discoveries. Thus, society probably would be worse off in the absence of this barrier to entry.

firms guarantees that prices will not exceed costs at each instance.[15] Therefore, costs of exit serve to *prevent entry*, just as do costs of entering a market.

General Evidence on Entry and Exit. Agriculture, construction, wholesale and retail trade, and services are generally thought to have easy entry and exit. In contrast, in some manufacturing industries, mining, and in certain regulated industries (public utilities and some insurance industries), entry and exit may be more difficult. According to the Economic Report of the President (2003, Table B 12), the composition of the U.S. gross domestic product in 2001 by sector is agriculture, 1 percent; construction, 5 percent; mining, 1 percent; manufacturing, 14 percent; transportation and public utilities, 8 percent; wholesale trade, 7 percent; retail trade, 9 percent; finance/insurance, and real estate, 2 percent; services, 22 percent; and government, 13 percent.

There are several interesting recent studies on entry and exit, including Berry (1992), Bresnahan and Reiss (1988, 1990, 1991), Dixit (1989), Geroski (1991), Lieberman (1990), Pakes and Ericson (1999), Mazzeo (2002), and Schary (1991). For example, Bresnahan and Reiss examine markets with only a few producers of professional services (such as rural markets for doctors) and ask how large a market must become before a single firm enters. They then ask how much more the market has to grow for two, three, and more firms to enter. They find that competition acts very quickly to reduce price and profits. Although initial entrants can charge a high price, the entry of two additional firms appears to produce competition. They also obtain a measure of sunk costs by comparing the size at which exit occurs to the size at which entry occurs. The larger a sunk cost, the smaller is the market size that triggers exit compared to the market size that induces entry.

The empirical literature suggests that there is much entry and exit and that entrants tend to be small. These high rates of entry and exit are roughly equal in stable industries (Caves 1998). Dunne, Roberts, and Samuelson (1988) find that entrants are much smaller than the average firm in a manufacturing industry. They produce 17 percent of the output level of existing firms and account for about 11 percent of industry output on average.[16] Similarly, exiting firms, which produce 11 percent of industry output, only produce one-fifth the output of the average firm. Despite entry, the four largest firms in an industry stay in that group on average for over 10 years (Caves 1998). Birch (1987), using Dun and Bradstreet data for all sectors (not just

[15] Baumol, Panzar, and Willig (1982) emphasize this point and are responsible for popularizing the concept of hit-and-run entry and relating it to contestability. See also Eaton and Lipsey (1980). Weitzman (1983) shows that hit-and-run entry is equivalent to a horizontal supply curve.

[16] There are not many economic models about how new firms enter and grow in an industry. Simple application of the competitive model suggests no difference between a new entrant and an existing firm. More realistic models based on differences in knowledge can generate specific growth processes for new firms and a distribution of firm sizes. See Jovanovic (1982), Jovanovic and MacDonald (1984), Hopenhayn (1992), and Ericson and Pakes (1995). See also Evans (1987a, b), Hall (1987) and Syverson (2003). Sutton (1997) and Caves (1998) report that the variability of a firm's growth rate first falls and then levels off with firm size, that a firm's average growth rate diminishes with size, and that a firm's probability of survival increases with age and size. See Sutton (1997) for a discussion of Gibrat's Law, which postulates a log-normal distribution for firm size.

manufacturing), finds that about half of all new entrants fail within five years and that despite their high rate of failure, entrants over a period of a few years are a significant source for the creation of (net) new jobs. However, employment by entrants accounts for a disproportionately small share of total employment and does not generate a disproportionately higher share of employment growth (Davis, Haltwanger, and Schuk 1996).

In many new industries, a massive entry of small firms is followed by a shakeout that eliminates the weakest firms. The surviving firms then grow both in size and in functions until the industry eventually goes into decline. This pattern is explained by the models of entry and exit discussed in this chapter in which firms learn whether they are efficient and, if not, exit.

The beer, auto, and tire industries followed this pattern. For example, in the beer industry, massive entry in the 1870s doubled the number of firms. Massive exit in the 1880s reduced the number of firms by 40 percent. Firms that entered just prior to the shakeout had higher survival probabilities than firms that entered after the shakeout had begun.

The spike in entry followed by a spike in exits can be pronounced. The number of tire manufacturers rose from about 170 in 1915 to about 270 in 1921, fell back down to about 150 around 1925, and dropped to approximately 50 by the beginning of the Depression in the early 1930s. Similarly, the number of auto firms increased from about 150 in 1905 to 250 in 1910, fell back to 150 by 1915, and dropped to below 30 by 1930.[17]

Identifying Barriers to Entry

Bain (1956) pioneered the modern approach to analyzing barriers to entry. He identified three such barriers:

- Absolute cost advantage
- Economies of large-scale production that require large capital expenditures
- **Product differentiation**: related products that have varying characteristics so that consumers do not view them as perfect substitutes (for example, Apple computers are not perfect substitutes for IBM computers)

An absolute cost advantage allows an incumbent firm to earn excess profits without fear of new firms entering the market. For example, suppose Firm A can produce at a constant cost of $2 per unit, whereas all other potential firms could produce at a cost of $5. Firm A can set its price at $4, which is above its per unit cost, earn an unusually high profit, and yet not fear entry. Because it is less clear that the other two barriers fit our definition of long-run entry barriers, let us examine them in more detail.

If both an incumbent and a new entrant can enjoy the same benefits of economies of scale, why should an incumbent be able to earn excess profits? Some argue that a new entrant would have difficulty raising money (or be unwilling to invest its own money) to finance a large expenditure. It is not necessarily true that it is more difficult

[17] Klepper (2000) and Hovarth, Schivardi, and Woywode (2001) [statistics read from their graphs].

to raise money for large than small projects.[18] If capital markets work properly (banks and others are willing to loan money for profitable activities), raising capital should be no more difficult for a profitable large-scale project than for a profitable small-scale project. There should be many investors for good projects.

But is it reasonable that the scale of a firm has no effect on the incentives to enter? If large *sunk* costs are associated with entry and if entry is unsuccessful, the entrant's losses are large. In such a setting, threats of strategic behavior (for example, vigorous price cutting) may prevent new entry. The greater the risk of encountering strategic behavior and the greater the potential loss, the more potent is the threat of strategic entry deterrence. In such a case, the need for large-scale investment that involves large sunk costs could well provide a disincentive for a potential entrant because it would have so much to lose (see Chapter 11).

Product differentiation (firms produce similar but not identical products) can create a long-run barrier to entry. For example, consumer goodwill toward established brand names may make it more difficult for a new brand to enter. Of course, an advantage may accrue to the first firm to introduce a new product. That firm may have a **first-mover advantage**: the first firm to enter incurs lower marketing costs because it faces no rivals (see Chapter 11). Later firms face higher marketing costs because they must compete against the first.[19] If the presence of the incumbent raises the marketing costs of the second firm to enter, then the first firm has a permanent advantage—a long-run barrier to entry—and can maintain high prices.[20] For example, because the product of the first firm in the market is familiar to customers, they may be reluctant to switch to a new brand (Schmalensee 1982).

The Size of Entry Barriers by Industry

A number of methods are used to assess long-run barriers to entry. Some economists use subjective judgments to predict how difficult it would be for a new firm to enter a market. These estimates can be based on how frequently entry has occurred in the past.

Other measures of barriers to entry are based on the answers to questions like the following: What would be the cost disadvantage if a new entrant's plant was half the size of the incumbent's? How much higher are the entrant's costs because of the incumbent's patents or acquired expertise? Tables 3.2 and 3.3 reproduce Bain's characterization of the extent of barriers to entry in certain industries.

Harris (1976) examined the rate of entry into those industries that Bain and later Mann (1966) considered difficult to enter, and found that several of these industries had significant entry. The entry barriers identified by Bain and Mann that did seem to

[18] Dunne, Roberts, and Samuelson (1988) find that existing firms that choose to enter a new business enter at a larger scale than do newly created firms. Possibly, capital market imperfections are more easily overcome by existing firms than by entrepreneurs without track records or existing firms are more confident of their likely success.

[19] Sometimes the second firm to enter has lower marketing costs than the first firm, which had to spend money educating consumers about the use and desirability of the new type of product.

[20] Caves and Porter (1977) stress the importance of *mobility barriers* that prevent firms in an industry from moving into different segments of that industry.

TABLE 3.2 Bain's Barriers to Entry

Industry	Scale Economy	Product Differentiation	Absolute Cost	Capital Requirement
Automobiles	3	3	1	3
Cigarettes	1	3	1	3
Liquor	1	3	1	2
Shoes	2	1–2	1	0
Soap	2	2	1	2
Steel	2	1	3	3
Tractors	3	3	1	3
Tires and tubes	1	2	1	2
Meat packing	2	2	2	0–1
Cement	2	1	1	2
Flour	1	1–2	1	0

Note: Higher scores indicate greater entry barriers.

Source: Bain (1956, 169)

TABLE 3.3 Bain's Overall Barriers to Entry

Industry	Overall Barriers
Automobiles	Very high
Cigarettes	Very high
Liquor	Very high
Soaps	Substantial
Steel	Substantial
Tractors	Very high
Flour	Moderate to low
Cement	Moderate to low
Meat packing	Moderate to low
Tires	Moderate to low
Rayon	Moderate to low

Note: Industries with very high barriers could elevate price 10 percent or more above competitive levels. Substantial and moderate-to-low entry barriers allow prices to be in excess of competitive levels by 7 percent and 4 percent respectively.

Source: Bain (1956, 170).

restrict entry were those having to do with product differentiation. Only long-run barriers to entry can prevent prices from eventually falling to equal marginal cost.

From a practical point of view, if the long run is very long, knowing that profits will eventually be driven to zero may not matter much to incumbent firms. Large short-run profits are still desirable. The time it takes for entry to expand output enough to eliminate unusually high profits may be more informative than the size of the long-run entry barrier.

The speed with which entry into various industries erodes profits may differ across competitive and noncompetitive industries. Most researchers find that profit erosion takes longer in concentrated industries (Stigler 1963) and in high-profit industries (Connolly and Schwartz 1985). Dunne, Roberts, and Samuelson (1988) find that differences in entry and exit rates across manufacturing industries persist over time. Moreover, there is considerable dispersion in entry and exit rates across industries. They find that the rates of entry and exit in a market are highly related. Industries with high rates of entry also have high rates of exit. In roughly half the manufacturing industries, entrants account for 7 to 25 percent of industry value, and exiting firms account for 8 to 25 percent of value.

Externalities

A competitive market may lack desirable welfare properties for reasons in addition to inefficient taxes and restrictions on entry. The competitive equilibrium is nonoptimal when a valuable good has no price or the wrong price. Unfortunately, many *goods* (such as information and fresh air) or *"bads"* (such as pollution and garbage) may not be priced in our economy. An externality occurs when consumers or firms do not bear the full cost (benefit) from the harm (good) their actions do to others.

Pollution is one of the most important examples of a negative externality, which is an uncompensated action that harms someone. Pollution is a bad that has no price. In the absence of government regulations, manufacturing firms do not pay for the pollution they create, so they ignore the cost to society of pollution in deciding how much output to produce. That is, their *private marginal cost* (their out-of-pocket production cost) of making one more unit is less than the *social marginal cost* (the private marginal cost plus the damage from the pollution). As a result, they produce more than is socially optimal. Such distortions, or inefficiencies in production due to improper pricing, are referred to as market failures.

An uncompensated action that benefits others is a positive externality. For example, when you plant a beautiful garden in view of your neighbor, your neighbor receives a free benefit. Two important examples of positive externalities involve the generation and dissemination of information, which can benefit many people at once. When Henry Ford developed the assembly line, other firms benefited from his innovation without compensating him. Many consumers who don't buy *Consumer Reports* learn of its ranking of automobiles and benefit from this in-formation. Whereas too much is typically produced in industries with negative externalities, too little is typically produced in industries with positive externalities.

Information is also described as a **public good**: a commodity or service whose consumption by one person does not preclude others from also consuming it. Another example of a public good is national defense. In addition to being an externality, pollution is a public bad (an undesirable public "good"). In contrast, a *private good,* such as a hot dog, is consumed by only one consumer—it cannot be consumed by another as well. An externality can be either a private or public good or bad. We now show that externalities arise when property rights are not clearly defined (Coase 1960).

You have **property rights** when you own or have exclusive rights to use some asset such as a good or service. Others must compensate you if they wish to use your property. For example, you may have property rights to a particular car, but no clearly defined area of a highway belongs to you alone. You share the highway with others. Each driver claims a temporary property right in a portion of the highway by occupying it (thereby preventing others from occupying the same space). Competition for space on the highway can lead to congestion (a negative externality), which slows up every driver on a highway. (See Example 3.4).

EXAMPLE 3.4 *Increasing Congestion*

According to the annual reports of the Texas Transportation Institution (TTI) at Texas A&M University, U.S. highway congestion and its costs have increased substantially over the last two decades. TTI estimates the cost from congestion in 75 areas around the United States at $67.5 billion in 2000, which reflects 3.6 billion hours of delay and 5.7 billion gallons of excess fuel consumed. In Los Angeles—the most congested city—the annual delay per peak-hour traveler was 136 person-hours, up from 47 hours in 1982. For all areas, the average annual delay increased from 16 hours to 46 hours from 1982 to 2000. The annual congestion cost per peak road traveler was $2,510 in Los Angeles and averaged $1,160 across the country. In addition, stalled traffic causes increased air pollution.

The only way to reduce this congestion is to decrease the number of simultaneous users of the highways or to increase the highways' capacity. According to the TTI, preventing congestion from growing between 1999 and 2000 would have required 1,780 new lane-miles of freeway and 2,590 new lane-miles of streets, or an average of 6.2 million additional new trips per day taken by either carpool or transit, or other operational improvements that allowed 3 percent more travel to be handled on the existing systems, or some combination of these actions.

One approach to controlling the congestion externality would be to impose a toll on users who travel during peak hours, where the toll equals the externality cost. Such a toll would force drivers to take into account the cost they impose on others when they travel.

Source: **mobility.tamu.edu/ums/study/short_report.stm, mobility.tamu.edu/ums/study/appendix_A/**

Where property rights are clearly defined so there are no externalities, competitive markets are efficient. For example, growing wheat does not affect a farmer's neighbors or anyone else usually, and farmers have the right to produce and sell the wheat as they choose. In contrast, if property rights are not well defined, markets may be inefficient. For example, if a software company cannot protect its property right to its computer programs and prevent other firms from selling them, these programs provide a positive externality (the company is not compensated for use of its programs). As a result, too few resources are devoted to producing software. Unfortunately, externalities are common. Specific examples include pollution and fisheries.[21]

Limitations of Perfect Competition

Some markets satisfy most of the assumptions of the perfect competition model. For example, on the New York Stock Exchange, many people buy and sell shares of stock of individual firms like IBM. Individuals who own the stock but want to sell it ask their brokers to sell, and those who want to buy ask their brokers to buy. There are many well-informed participants in the stock market, and the price of a particular stock is determined by the forces of supply and demand. Most individuals correctly believe that they have no effect on the price of the stock. Although this market comes close to satisfying the assumptions for perfect competition, most do not.

The model of perfect competition is directly relevant to only a few markets, and much of this book is devoted to analyzing the consequences of more realistic models of economic behavior. In these more realistic models, firms are able to influence price and their rivals' actions, engage in advertising and other marketing activities to inform and influence consumers, and undertake research and development to produce more efficiently.

Many people observe that, even in perfectly competitive markets, welfare is nonoptimal if the distribution of income is "unjust." Individual wealth depends on assets (for example, money, machines) and skills. Competition does not necessarily reward the deserving. It rewards those who are the most productive and those who own productive assets.

If the distribution of income is not just, why does anyone care about efficiency? After all, efficiency means only that one person cannot be made better off without another being made worse off. If there is only $10,000 to distribute and Debbie gets $9,999 and Rebecca gets $1, that is efficient because all $10,000 is distributed. But so is $1 for Debbie and $9,999 for Rebecca. There are many efficient points; in fact, *any* division of the $10,000 between Debbie and Rebecca is efficient. The particular efficiency point that the competitive equilibrium produces depends on the initial owner-

[21] See **www.aw-bc.com/carlton_perloff** "Pollution" for a more extensive discussion of pollution and externalities, and of how the assignment of property rights to either firms or consumers can lead to an efficient outcome.

ship of assets. Public policy may assert that it is unjust for either Debbie or Rebecca to receive only $1 while the other receives $9,999, and might prefer $5,000 to Debbie and $5,000 to Rebecca. It might even find $4,999 to Debbie and $4,999 to Rebecca preferable to the policy of $9,999 for one and $1 for the other. That means that an *inefficient* policy may be preferred to an efficient one! Why then do economists seem to stress efficiency?

One answer is that the morally just (however defined) distribution of income can be achieved by competition plus a system of appropriate income redistribution. That is, the government could assign wealth initially according to society's moral values, and then competition would lead society to an efficient outcome. One interpretation of this result is that it is up to the government to achieve the moral distribution of wealth through nondistorting taxation, and it is up to the competitive process to achieve efficiency.

Economists can objectively discuss whether economic efficiency is achieved, but they are no better equipped than others to discuss the best or most moral income distribution. They may analyze how the distribution of income changes as a result of certain policies, but they cannot scientifically determine whether one distribution is ethically superior to another.

The Many Meanings of Competition

We have been specific about the definition of perfect competition. Many non–economists and even many economists use the term *competition* loosely to apply to markets that we refer to as noncompetitive.

Some people use the term *competition* to refer to a market in which a few price-setting firms compete vigorously for sales. Each firm tries to obtain customers for itself at the expense of its rivals. In this interpretation, competition is used to describe rivalry between firms that can affect market price. This use of the term differs from our definition, where a perfectly competitive firm is a price taker that can sell all it wants at the market price.

Even though few industries fit the requirements of perfect competition, economists often speak of certain types of industries as being reasonably competitive if they have certain characteristics. Price-taking behavior, many firms, and free entry and exit are often used as criteria to judge the competitiveness of a market. Free entry and exit typically result in firms eventually earning zero profits. For example, in some states, it is very easy to become a barber, and there are often many independent barber shops in an area. Even though barbers are not identical in either quality or prices and all consumers are not aware of all barber shops, most economists would describe the provision of haircuts by barber shops as a reasonably competitive market.

Another example of a reasonably competitive market is steel scrap. Firms in that market collect used steel, process it, and sell it to steel firms. Entry into the steel-scrap market is easy and quick, the product sold is fairly homogeneous, and there are published quotations of prices. Even though transaction prices undoubtedly vary from firm to firm, the large number of sellers acting independently would cause most economists to label this market as reasonably competitive.

Some discussions involving public policy use the term *competitive* in a still different way: A competitive market is one that requires no intervention to improve its performance; a noncompetitive market is one that has some defect that should be corrected. This usage of the terms *competitive* and *noncompetitive* can be confusing. The confusion arises because intervention can sometimes improve the performance of industries that satisfy all the assumptions of perfect competition—as can occur, for example, when the government encourages inventive activity. Conversely, the failure of a market to satisfy all the assumptions of perfect competition does not necessarily mean that some intervention can improve market performance.

SUMMARY

In perfect competition, all firms produce homogeneous, perfectly divisible output; producers and consumers have full information, incur no transaction costs, and are price takers; and there are no externalities. If all these conditions hold, use of resources is efficient. Welfare defined as consumer surplus plus producer surplus is maximized.

Government interventions in competitive markets such as taxes and restrictions on entry and exit reduce the efficiency of these markets. Government intervention may be helpful, however, if some of the assumptions of perfect competition do not hold. For example, where property rights are not clearly defined or high transaction costs prevent negotiated solutions, polluting competitive firms do not pay for the damage they cause, and produce too much pollution. The optimal government policy reduces the pollution.

The assumptions of perfect competition do not hold in many industries. Subsequent chapters explore firms' behavior and the consequences of departures from perfect competition.

PROBLEMS

1. If the inverse demand function for toasters is $p = 60 - Q$, what is the consumer surplus when the price is $30?
2. The government imposes a fixed fee per year on each firm that operates in a competitive market. What happens to output, the optimal scale of a firm, and price if there is free entry into the market?
3. Suppose a competitive market consists of identical firms with a constant long-run marginal cost of $10. (There are no fixed costs in the long run.) Suppose the demand curve at any price, p, is given by $Q = 1{,}000 - p$.
   a. What are the price and quantity consumed in the long-run competitive equilibrium?
   b. Suppose one new firm enters that is different from the existing firms. The new firm has a constant marginal cost of $9 and no fixed costs but can only produce 10 units (or fewer). What are the price and the quantity consumed in long-run competitive equilibrium? Are these the same as in (a)? Explain.
   c. Are positive economic profits inconsistent with a long-run competitive equilibrium?
   d. Identify the marginal cost of the last unit sold in (b). Is it $10 or $9? That is, if demand fell by 1 unit, would the new entrant or the other firms reduce output?
   e. How much profit do the less efficient firms in (b) earn?
   f. In the long-run competitive equilibrium, must the profit of the marginal entrant (the next firm to enter the market if demand expands or, alternatively, the next firm to leave the market if demand contracts) be zero?

4. If the market demand curve is $Q = 100 - p$, what is the market price elasticity of demand? If the supply curve of individual firms is $q = p$ and there are 50 identical firms in the market, draw the residual demand facing any one firm. What is the residual demand elasticity facing one firm at the competitive equilibrium?

5. When is a firm's shutdown point equal to the minimum point on its average cost curve?

6. The U.K. produces and imports eggs. Suppose that the government imposed a quota on imports: Foreign suppliers could export no more than Q eggs (regardless of price). What effect does this quota have on the foreign supply curve of eggs, the total U.K. supply curve of eggs, the equilibrium price, British consumers, and British producers?

Answers to the odd-numbered problems are given at the back of the book.

CHAPTER 4

Monopolies, Monopsonies, and Dominant Firms

Aeroflot Airlines: You Have Made the Right Choice.
—Ad campaign for the only airline in the then Soviet Union

A firm is a **monopoly** if it is the only supplier of a product for which there is no close substitute. A monopoly sets its price without fear that it will be undercut by a rival firm. A monopoly faces a downward-sloping demand curve and sets a price above marginal cost. As a result, less is sold than if the market were competitive (where price equals marginal cost) and society suffers a deadweight loss.

This chapter analyzes a monopoly's behavior and the consequences of that behavior. It also discusses how a monopoly is maintained and asks whether monopoly is always bad. The effects of externalities in a monopolized market are discussed next. The chapter then turns to two related topics. It examines monopsony, which is a monopoly on the buying side of the market. Then it discusses what happens to a monopoly if higher-cost competitive firms enter its market.

The six key questions we answer in this chapter are:

1. How does monopoly compare to competition in terms of prices and welfare?
2. How are monopolies created and maintained?
3. Are there markets in which there are benefits to monopoly?
4. Are all firms that earn profits monopolies, do all monopolies earn profits, and can monopolies earn profits in the long run?
5. How does a monopsony exercise its market power?
6. What happens to a monopoly if smaller, price-taking firms enter its market?

Monopoly Behavior

Because a monopoly faces a downward-sloping market demand curve, it can raise its price above marginal cost. To maximize its profit, it has an incentive to produce its output efficiently. A firm's behavior and government regulations influence the firm's ability to become and remain a monopoly.

Profit Maximization

> *Price, n. Value, plus a reasonable sum for the wear and tear of conscience in demanding it.*
> —*Ambrose Bierce*

Like a competitive firm, a monopoly sets its level of output to maximize its profits. Because the market demand curve is downward sloping, the more the monopoly sells, the lower the price it receives.

The market demand curve constrains the monopoly. In its quest to maximize profit, it can set only price or only quantity—not both. If the monopoly sets quantity, the market price is determined by the market demand curve. If it sets price, the quantity is determined by the market demand curve.

Given the demand curve in Figure 4.1, if the monopoly wants to sell Q_0 units of its product, it charges price p_0. If it wishes to sell one more unit, it has to lower its price to p_1.

If the monopoly lowers its price to p_1, its revenues may rise or fall. The monopoly gains revenue on the extra unit it sells at price p_1, Area B in Figure 4.1. To sell that extra unit, however, it must cut its price from p_0 to p_1 on the original Q_0 units, resulting in a loss of revenues of $(p_0 - p_1)Q_0$, Area A in Figure 4.1.

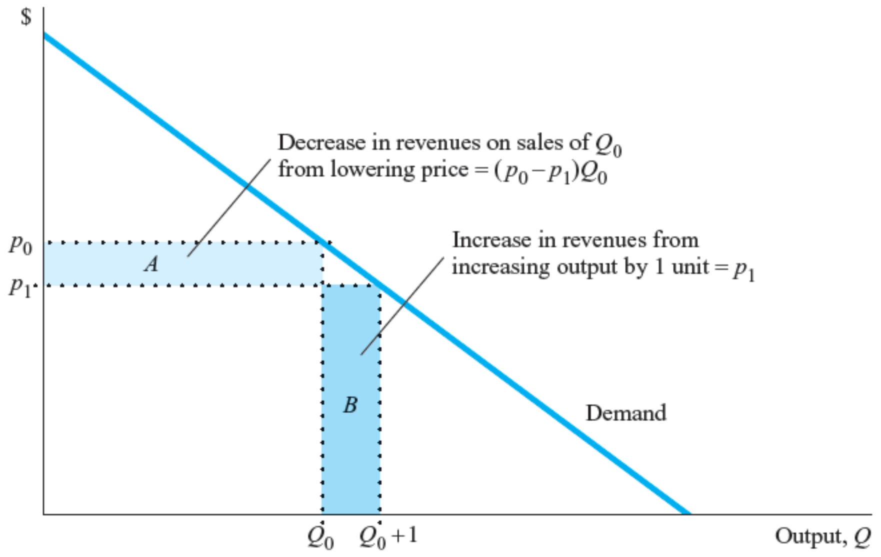
When discussing a competitive firm's behavior in Chapter 3, we did not have to consider this loss of revenue due to lower price. Because a competitive, price-taking firm faces a horizontal demand curve, the price it receives does not fall if it expands its quantity.

If Area B is larger than Area A in Figure 4.1, then selling the extra unit causes revenues to rise. The extra revenues, $p_1(Q_0 + 1) - p_0Q_0$, that a firm receives when it produces one more unit of the product is called the **marginal revenue**.[1] Hence, the marginal revenue equals area B minus Area A. If the monopoly did not have to lower its price to sell the additional unit, then the increment to revenues from selling an additional unit would simply be the initial price, p_0. But because the demand curve

[1] Marginal revenue, MR, is the change in revenue from selling an additional unit. Total revenue is $p(Q)Q$, where $p(Q)$ is the inverse demand curve (p is a decreasing function of Q). Marginal revenue is equal to $p + Q(\Delta p/\Delta Q)$, where $(\Delta p/\Delta Q)$ is the decline in price necessary to sell the additional unit. Using calculus,

$$MR = \frac{d(p(Q)Q)}{dQ} = p + \frac{dp}{dQ}Q.$$

FIGURE 4.1 Demand Curve Facing a Monopoly



is downward sloping, the monopoly must lower its price to sell more units. Therefore, the marginal revenue is always less than the price for a monopoly, as Figure 4.2a illustrates.[2] For a firm in a perfectly competitive market, marginal revenue equals price.

Marginal revenue and total revenue are closely related. When marginal revenue is positive, total revenue increases as output expands, but when marginal revenue is negative, total revenue falls as output expands. As a result, total revenues are maximized (Figure 4.2b) when marginal revenue equals zero (Figure 4.2a).[3]

A monopoly maximizes its profit rather than its revenue (just as a competitive firm does). Profit is maximized at a smaller quantity than is revenue, as Figure 4.2b illustrates.

[2] If a straight-line demand curve, as in Figure 4.2a, hits a horizontal line at Q, the corresponding marginal revenue curve is also a straight line and hits the horizontal line at $Q/2$. To prove this result, let the straight-line demand curve be $p = a - bQ$. Total revenue is $R = pQ = aQ - bQ^2$. The marginal revenue curve is obtained by differentiating R with respect to Q: $MR = a - 2bQ$. The demand curve hits the horizontal axis ($p = 0$) at $Q = a/b$. The marginal revenue curve hits the horizontal axis ($MR = 0$) at $Q = a/(2b)$.

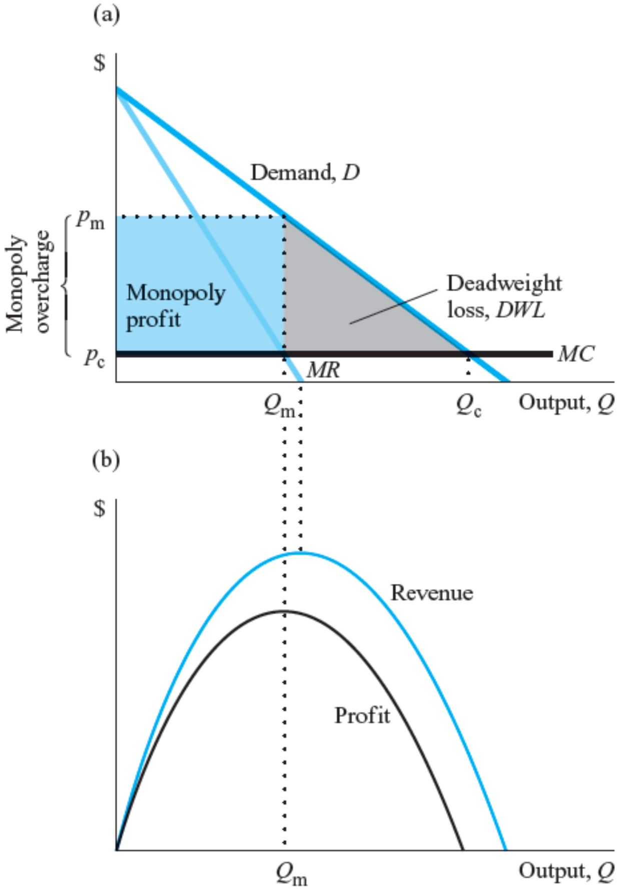
[3] If the monopoly wants to maximize revenues through its choice of Q,

$$\max_{Q} R = p(Q)Q,$$

it sets its marginal revenues equal to zero (first-order condition):

$$MR = p + \frac{dp}{dQ}Q = 0.$$

FIGURE 4.2 Monopoly Profit Maximization



A monopoly maximizes its profit when the extra revenue from selling one more unit just equals the extra cost of producing that last unit of output. That is, profit is maximized where *marginal revenue equals marginal cost*:[4]

$$MR = MC. \tag{4.1}$$

[4] If the monopoly wants to maximize profits through its choice of Q

$$\max_{Q} \pi = p(Q)Q - C(Q),$$

it sets its marginal profits equal to zero (first-order condition):

$$\frac{d\pi}{dQ} = MR - MC = \left(p + \frac{dp}{dQ}Q\right) - \frac{dC}{dQ} = 0.$$

(continues)

Figure 4.2a illustrates this profit-maximizing relationship. The profit-maximizing monopoly output, Q_m, is smaller than the competitive output, Q_c, determined by the intersection of the demand curve with the marginal cost curve (which we assume would be the supply curve if the market were competitive) at price p_c. The monopoly does *not* have a supply curve that can be specified solely as a function of price because the monopoly's output depends on marginal revenue (which depends on the slope of the demand curve) and marginal cost.

The properties of the demand curve determine the monopoly *overcharge*: the amount by which the monopoly price, p_m, exceeds the marginal cost or competitive price, p_c, in Figure 4.2a. A relationship exists between the monopoly overcharge and the price elasticity of demand.

The elasticity of demand is a characteristic of the demand curve and is defined as the percentage change in quantity that results from a 1 percent change in price. If the elasticity of demand is very high (a large negative number), then the curve is said to be *elastic*. With a very elastic demand, a small price change induces a very large change in the quantity demanded. If the elasticity is low (a number between -1 and 0), the demand curve is *inelastic*, and a price change of 1 percent has relatively little effect on the quantity demanded.

Marginal revenue can be written as[5]

$$MR = p\left(1 + \frac{1}{\epsilon}\right), \tag{4.2}$$

where ϵ is the elasticity of demand. Thus, marginal revenue is positive if the demand curve is elastic ($\epsilon < -1$). It is negative if the demand curve is inelastic ($-1 < \epsilon < 0$). The elasticity of demand, in general, depends on not only the particular demand curve but also the point (the price and quantity pair) on the demand curve. For example, the elasticity of demand could decrease as price becomes lower.

By substituting Equation 4.2 for *MR* in Equation 4.1, we can write the profit-maximizing condition for the monopoly as:

$$\frac{p - MC}{p} = -\frac{1}{\epsilon}. \tag{4.3}$$

Thus, it sets $MR = MC$. Another condition for profit maximization is that the marginal revenue curve cut the marginal cost curve from above, as in Figure 4.2a. That is, the second-order condition must hold:

$$\frac{d^2\pi}{dQ^2} = \frac{dMR}{dQ} - \frac{dMC}{dQ} < 0.$$

A monopoly uses the same shut-down condition as does a competitive firm. In the short run, if price is below average variable cost, the monopoly stops producing.

[5] Differentiating revenue, $R = p(Q)Q$, with respect to Q, we find that the marginal revenue is

$$MR = p + \frac{dp}{dQ}Q = p\left(1 + \frac{dp}{dQ}\frac{Q}{p}\right) = p\left(1 + \frac{1}{\epsilon}\right),$$

where ϵ is defined as $(dQ/dp)\,(p/Q)$.

The left-hand side of Equation 4.3 is the price-cost margin: the difference between price and marginal cost as a fraction of price, $[p - MC]/p$. As the equation shows, the price-cost margin depends on only the elasticity of demand the monopoly faces. The price-cost margin is also called the Lerner Index of market power (Lerner 1934).

Equation 4.3 shows that the monopoly's price is close to MC when the demand is very elastic, and the price increasingly exceeds MC as the demand becomes less elastic. For example, if the elasticity of demand is -2, price is twice marginal cost. If the elasticity is -100 (very elastic), price equals $1.01MC$. The higher the elasticity of demand, the closer is the monopoly price to the competitive price. Therefore, the key element in an investigation of market power is the price elasticity of demand. Where the elasticity of demand is relatively inelastic, a monopoly markup may be substantial, as Example 4.1 illustrates.

Market and Monopoly Power

In contrast to a price-taking competitive firm, a monopoly knows that it can set its own price and that the price chosen affects the quantity it sells. A monopoly can set its price above its marginal cost but does not necessarily make a supracompetitive profit. For example, if a monopoly incurs a fixed cost, its profit may be zero (the competitive level) even if its price exceeds its marginal cost.

It is common practice to say that whenever a firm can profitably set its price above its marginal cost without making a loss, it has *monopoly power* or *market power*. One might usefully distinguish between the terms by using *monopoly power* to describe a firm that makes a profit if it sets its price optimally above its marginal cost, and *market power* to describe a firm that earns only the competitive profit when it sets its price optimally above its marginal cost. However, people do not always make this distinction, and generally use the two terms interchangeably, sometimes creating confusion.

EXAMPLE 4.1

Monopoly Newspaper Ad Prices

When the *Houston Post* shut down in April 1995, the managing editor of the sole surviving paper, the *Houston Chronicle*, received dozens of calls from concerned *Post* readers worried about one thing: Would the *Chronicle* pick up the *Post's* comics? Local advertisers also were very concerned: What would happen to newspaper advertising prices?

Ad rates skyrocketed by nearly 62 percent from January 1995 (before the *Post* folded) to December 1996. The rate for a one-column inch ad in a daily paper rose from \$252.64 to \$409.00 per day, and Sunday rates jumped from \$294.84 to \$477.28. These rates increased by much more than readership, which rose 32 percent on weekdays and 23 percent on Sunday. Thus, a loss of competition resulted in a substantial increase in price.

Source: Iver Peterson, "New Realities of Life in a One-Paper Town," *New York Times*, December 30, 1996:C5.

The Incentive for Efficient Operation

> *Organized crime in America takes in over forty billion dollars a year and spends very little on office supplies.* —*Woody Allen*

The consequences of inefficient behavior are different for monopolies and competitive firms. An inefficient competitive firm may not be able to remain in business because it is unprofitable, but an inefficient monopoly can profitably remain in business. This observation has led some to conclude that the monopoly strives less hard to be efficient (called *x-inefficiency* by Leibenstein 1966) than does a competitive firm.

This argument is rejected by many economists who believe that monopolies, like other firms, prefer more to less. Monopolies want to maximize profits, and the only way a firm can do so is to minimize its costs at its chosen output level. Therefore, to postulate that monopolies want to maximize profits is to assume implicitly that they also minimize their costs. No firm—monopolistic or competitive—wants to throw money away. If improving the efficiency of operations increases profits, the firm should do it, whether it is a monopoly or a competitor.

A monopoly, however, may not have the same *ability* to produce as efficiently as a competitive firm. A firm in a market with many other firms can observe what other firms are doing. It can observe, for example, whether its own costs of production are above or below the market price. Because the market price reflects the efficiency of the other firms in the market, a competitive firm knows that it can improve its production efficiency if its costs of production are high relative to the market price. In contrast, a monopoly has no other firms to look at and may have no other standard by which to judge how efficiently it is operating. Therefore, a competitive firm may operate more efficiently than does a monopoly because it is more difficult for a monopoly to monitor internal efficiency than it is for a competitive firm.

Monopoly Behavior over Time

If demand is inelastic ($-1 < \epsilon < 0$), it is not possible to satisfy the profit-maximization condition of Equation 4.3. Thus, a monopoly never operates on the inelastic portion of its demand curve. If a monopoly were operating in the inelastic portion of its demand curve, it could increase its profits by raising its prices until it was operating in the elastic portion of its demand curve. In the inelastic portion of the demand curve, a 1 percent increase in the monopoly's price causes the quantity sold to fall by less than 1 percent, so that revenues increase. With reduced output, however, the monopoly's costs must fall, so that total profits must rise. Thus, if the monopoly is operating in the inelastic portion of the demand curve, it should keep increasing its price, obtaining ever more profits, until it is in the elastic portion of the demand curve.[6]

[6] What if there were no elastic portion of the demand curve? The monopoly would produce just a small amount of output, charge an infinite price, and make infinite profits. That this story is implausible underscores the empirical irrelevance of a monopoly's demand curve that is everywhere inelastic.

This observation, however, applies only in the context of a simple, timeless model. In actual markets, demand curves shift over time. As a result, a rational monopoly changes its price over time.

Consumers may have a more inelastic demand curve in the short run than in the long run. In the short run there are limitations on how fast consumers can substitute away from a product in the face of a price increase. Therefore, if a monopoly takes advantage of an inelastic portion of its short-run demand curve and raises its price, its consumers are more likely to substitute away from its product in subsequent periods. Thus, a monopoly may operate in the inelastic portion of its short-run demand curve to avoid long-run substitution.

The oil market provides an excellent example of the time it takes to substitute away from a product. When the Organization of Petroleum Exporting Countries (OPEC) raised the price of oil in the early 1970s, total consumption of energy changed very little in the first year. However, the quantity of oil demanded fell sharply over the next several years as consumers adjusted to the increased price and began to take energy-saving measures.

The Costs and Benefits of Monopoly

A monopoly is socially reprehensible in the hands of others.

If a monopoly restricts its output and raises its price above marginal cost, society suffers a deadweight loss. We first examine why such behavior leads to a deadweight loss. Then we use our understanding of how monopolies arise to show that, in certain circumstances, there are benefits associated with monopolies. Indeed, in certain situations, monopoly may be preferable to competition.

The Deadweight Loss of Monopoly

In order to maximize its profit, a monopoly sets its output where its marginal revenue curve intersects its marginal cost curve, as Figure 4.2a shows. The gap between the monopoly's price and marginal cost represents the difference between the value (price) that buyers place on the product and the marginal cost of producing it. This gap is similar to the one caused by a tax on a competitive market (Chapter 3). In both cases, price and output differ from their competitive levels, and there is a deviation between the demand price (as given by the demand curve) and the supply price (as given by the marginal cost curve).

If consumers must pay a monopoly price p_m that is above the competitive price p_c, they lose consumer surplus equal to the sum of the monopoly profits and the deadweight loss in Figure 4.2a. The monopoly profit is less than the consumer surplus loss. Thus, society suffers a deadweight loss (the *DWL* triangle in Figure 4.2a) that equals the consumers' loss less the monopoly's gain. This *DWL* triangle is the area below the demand curve, above the marginal cost curve, and to the right of the equilibrium monopoly quantity.

Thus, both monopoly and an inefficient tax cause a deadweight loss. However, who keeps the transfer from consumers differs: Tax revenues go to the government, whereas the monopoly keeps the monopoly profit. Even fairly small deadweight losses may be associated with a large redistribution of wealth, as the "monopoly profit" box in Figure 4.2a illustrates.

Many researchers have estimated the deadweight loss that monopoly imposes on the U.S. economy. In a pioneering paper, Harberger (1954) calculated that the deadweight loss is small: less than 0.1 percent of the gross national product (GNP: a measure of the value of all goods and services in our economy).[7] Later researchers repeated these calculations based on different assumptions. Worcester (1973), for example, also finds that the *DWL* is small: 0.4 to 0.7 percent. Kamerschen (1966) estimates the *DWL* at 6 percent and Cowling and Mueller (1978) estimate that it is between 4 and 13 percent.[8] Jenny and Weber (1983) find that the *DWL* in France is as high as 7.4 percent.

Rent-Seeking Behavior

The gods help those that help themselves. —*Aesop*

Some researchers contend that the efficiency loss to society is much larger than the *DWL* triangle. They argue that an amount equal to some or all of the monopoly profits is also an efficiency loss.

Monopoly profits can be regarded as a transfer from consumers to the monopoly, just as tax revenues are a transfer of income from consumers to the government. By itself, a transfer of income does not affect efficiency. Only if the monopoly restricts output below competitive levels is there an efficiency effect.

However, Posner (1975) argues that the monopoly profits may also represent a loss to society to the extent that it creates incentives for a firm to use real resources to become a monopoly. For example, suppose that a firm can become a monopoly by persuading the government to pass a law that restricts entry into the market. The use of a firm's resources to hire lobbyists, lawyers, and economists to argue its case before legislators is a cost to society, because these resources could have been productively employed elsewhere.

If there is a positive monopoly profit, as in Figure 4.2a, a firm would be willing to spend an amount up to these profits in order to become a monopoly. Of course, the firm would like to spend as little as possible, but the opportunity to earn monopoly profit could create the incentive to use valuable resources up to the amount of monopoly profits in order to secure the monopoly.[9] Because firms compete to earn the "rent"

[7] Stigler (1956) and Cowling and Mueller (1978) criticize Harberger's methodology on technical grounds.

[8] See, however, Masson and Shaanan (1984) for a critique of this last result.

[9] Whether the firm would dissipate the entire monopoly profit depends on the institutional details as to how the monopoly can be acquired (Fisher 1985).

(monopoly profits) from the monopoly, the expenditure of resources to attain government-created monopoly profits is called **rent seeking**.

If rent seeking occurs, the calculation of the deadweight loss from monopoly must include that part of the transfer that is dissipated by the firms seeking to become the monopoly. Thus, the cost of monopoly is greater than the *DWL* triangle that Harberger calculated: The loss equals the *DWL* triangle plus at least part of the monopoly profits.

Posner recalculates the deadweight loss from regulated and unregulated monopoly on the extreme assumption that the entire amount of monopoly profit is dissipated in rent-seeking activities. His estimates of deadweight loss as a percent of revenues exceed previous estimates. For example, Posner found deadweight losses of up to 30 percent of revenues for some of the industries he examined (such as motor carriers, physician services, and oil). His insight was that a great part of the loss to the economy from monopoly (or, more generally, noncompetitive pricing) is directly traceable to the existence of government institutions that insulate some firms from competition. If he's correct, the recent rescinding of many government regulations (see Chapter 20) will provide sizable benefits to society.

Monopoly Profits and Deadweight Loss Vary with the Elasticity of Demand

Monopoly profits and the *DWL* triangle depend on the shape of the demand curve. We illustrate how monopoly profits and deadweight loss vary with the elasticity of demand with a linear demand curve,

$$p = a - bQ.$$

The light demand curve in Figure 4.3 is for $a = \$60$ and $b = 0.5$. Given a constant marginal and average cost, $MC = AC = \$10$, the monopoly sells $Q_m = 50$ units at $p_m = \$35$, where the elasticity of demand is -1.4. The monopoly's profit is Area $A = \$1{,}250$, and the deadweight loss is area $D = \$625$.[10]

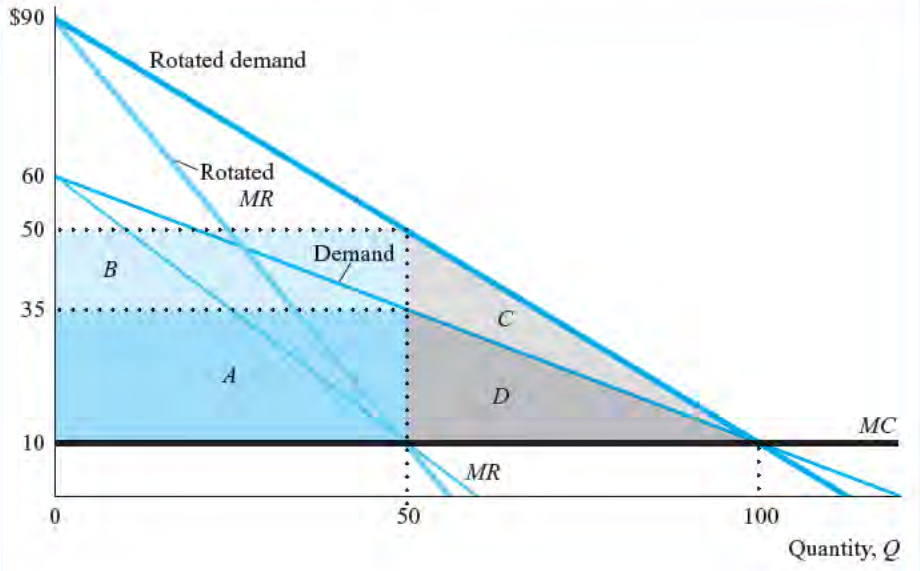
We now rotate the demand curve so as to vary the elasticity of demand. The demand curve is rotated around the point where it crosses the *MC* line, at 100 units. That is, for all the demand curves examined, if price were set efficiently at $MC = \$10$, consumers would buy 100 units. Because the demand curve is linear, the marginal revenue curve is also linear and crosses the horizontal *MC* line at half the distance that the demand curve does. Thus, the profit-maximizing monopoly equilibrium quantity of 50 units is unchanged as we rotate the demand curve.

[10] Let t (= $[p_m - p_c]/p_c$) be the monopoly markup above the competitive price. For small t, the monopoly *DWL* triangle can be approximated as

$$-1/2t^2R\epsilon,$$

where R would be the revenues if the product were sold at the competitive price (p_cQ_c) and ϵ is the elasticity of demand. *DWL* does not necessarily rise as the absolute value of ϵ increases because t is inversely related to ϵ, and as t changes, so does R. Holding R constant, *DWL* falls as the absolute value of ϵ increases.

FIGURE 4.3 Monopoly Profits and Deadweight Loss Vary with the Elasticity of Demand



The thick blue demand curve in Figure 4.3 shows a demand curve that has been rotated so that its intercept with the price axis, *a*, is $90, which is higher than the original $60 intercept. The thick blue demand curve intercepts the price axis at $90. The monopoly sells the same quantity as before, $Q_m = 50$, but at a higher price, $p_m =$ $50, so that the demand elasticity falls (in absolute value) from −1.40 to −1.25 (becomes less elastic), as Table 4.1 illustrates. The monopoly profit rises to $A + B =$ $2,000, and the deadweight loss increases to area $C + D =$ $1,000.

TABLE 4.1 Monopoly Profits and Deadweight Loss Vary with the Price Elasticity of Demand

Intercept of Demand Curve with Price Axis	Elasticity of Demand at *Q* = 50	Monopoly Price	Deadweight Loss	Monopoly Profit
$ 30	−2.00	$20	$ 250	$ 500
60	−1.40	35	625	1,250
90	−1.25	50	1,000	2,000
120	−1.18	65	1,375	2,750
150	−1.14	80	1,750	3,500

As the demand curve becomes less elastic at the monopoly equilibrium, people are less willing to do without this good: An increase in price causes the quantity they purchase to fall by less than if demand were more elastic. The monopoly, realizing this opportunity exists, increases its equilibrium price and earns a larger monopoly profit. As the demand curve becomes steeper at a given quantity (demand is more inelastic), the deadweight loss increases.

The Benefits of Monopoly

The welfare harms from monopoly may be offset by several benefits. These benefits are ignored in the static analysis above where we calculated deadweight losses. For example, the prospect of receiving monopoly profits may motivate firms to develop new products, improve products, or find lower-cost methods of manufacturing. Were it not for the quest to obtain monopoly profits, firms might innovate less.

The benefit of monopoly is most clearly recognized in research and development (see Chapter 16). If a firm succeeds in developing a new product, it can obtain a patent that prohibits other firms from using the patented technology for a fixed number of years—currently 20 years in the United States. Were it not for the patent, the innovative firm might discover that, within a matter of weeks, other firms had copied the new product. The innovative firm would then receive no more than the competitive level of profits and would not recover its expenditures on research and development. The firms that copied the product would have no research and development expenditure to recover. The ability of other firms to copy a new product removes the innovating firm's incentive to invest in research and development. The patent system attempts to deal with this problem by granting the innovating firm the sole property right to commercially exploit its innovation.

Naturally, if monopoly had no offsetting benefits, competition would be preferable. For example, if all firms in a competitive market decide to merge, and if the merger does not lead to a more efficient market, then the only result is the creation of a monopoly. As long as new entry takes time, the firms could price above their marginal cost. Because there is no benefit from this action, such behavior should be discouraged. One responsibility of the Department of Justice and the Federal Trade Commission is to scrutinize each merger carefully to make sure that its effect is not simply to raise prices to consumers.

Creating and Maintaining a Monopoly

There are several ways in which a firm may become and remain a monopoly. One possibility is that all the firms *merge* (combine into a single firm) or act in concert as a monopoly would. We address these possibilities in detail in Chapter 5 and in Example 4.2. Another possibility is that the firm takes strategic actions that prevent entry by other firms, as we discuss in Chapter 11 and in Example 4.3. Here, we examine three other reasons why a firm is able to create and maintain a monopoly:

EXAMPLE 4.2 *Monopolizing by Merging*

United States

In 2001, the Federal Trade Commission (FTC) accused the Hearst Corporation of illegally acquiring a monopoly over medical drug databases that are used by pharmacies and hospitals. Hearst bought Medi-Span, which was the only major competitor for Hearst's database company, First DataBank, according to the FTC. The FTC went on to contend that Hearst withheld information necessary for its premerger antitrust review. After Hearst acquired Medi-Span, it raised prices, doubling some and tripling others, according to the FTC and Express Scripts, a pharmacy-benefit management company. In its settlement with the FTC, Hearst agreed to return $19 million to customers. Later, Hearst paid more than $26 million to Express Scripts and other class-action plaintiffs in a private antitrust suit in 2002.

South Africa

South African Breweries controls 98 percent of South Africa's beer sales, with its 14 brands, including Castle, Lion, Heineken, Guinness, Amstel, and Carling Black Label. It was formed by a merger of two major competitors in 1979 because South Africa had virtually no antitrust laws. A company spokesman claims that the firm has little market power because the market is "fully contestable" with no legal barriers to entry. The firm's control of distribution channels may be responsible for its ability to maintain its high market share.

Sources: "FTC Accuses Hearst of Creating Monopoly," *San Francisco Chronicle*, April 15, 2001:D2; "Hearst Settles Dispute with FTC," *Milwaukee Journal Sentinel*, December 15, 2001:D1; Peter Shinkle, "Express Scripts Drops Antitrust Suit vs. Hearst; Maryland Heights Company with Share in FTC Settlement," *St. Louis Post-Dispatch*, May 23, 2002:C11; Donald G. McNeil, Jr., "In South African Beer, Forget Market 'Share'," *New York Times*, August 27, 1997:C1, C4; Bernard Simon, "Private Sector; An Old School Brewer for Miller," *New York Times*, February 2, 2003:3.2.

The firm may have special knowledge, the government may protect it from entry, or the market may only be large enough for a single firm to produce profitably.

Knowledge Advantage

A firm may be a monopoly because only it knows how to produce a certain product or it can produce the product at lower cost than other firms. A firm may have special knowledge that enables it to produce a new or better product that others cannot imitate. The firm may try to keep secret its special knowledge so as to prevent rivals from imitating it (see Example 4.4). A firm with an important secret faces a downward-sloping demand curve for its product and does not fear the entry of rival firms or the introduction of products that are close substitutes.

EXAMPLE 4.3 *Controlling a Key Ingredient*

In 2000, the attorneys general of 33 states and the Federal Trade Commission settled a lawsuit with Mylan Labs (and its suppliers) for $100 million. The suit contended that Mylan Labs cornered the market on the active ingredients for two drugs used to treat Alzheimer patients and then raised the price of the drug Clorazepate more than 3,000 percent (from about 2¢ a tablet to over 75¢) and increased the price of the drug Lorazepam more than 2,000 percent (from about 1¢ a tablet to over 37¢).

Source: **www.state.ia.us/government/ag/mylan.htm**.

Similarly, a firm may have special knowledge about production techniques that enable it to produce the same product at lower cost than other firms, which may be unable to discover the production technique of the efficient firm. We illustrate this possibility in Figure 4.4. Initially all the firms in a competitive market have a constant marginal cost, m_1, so the equilibrium price, p_1, equals m_1 and the equilibrium quantity is Q_1. One firm discovers a new production technique that it can keep secret and that lowers its marginal costs from m_1 to m_0. It faces a *residual demand curve* (the unmet demand after all other firms sell as much as they want at a given price) that is horizontal at p_1 (equal to m_1) up to Q_1 because many firms can produce and sell at price m_1. Beyond Q_1 (prices less than p_1), the residual demand curve coincides with the market demand curve because below p_1 no other firm can profitably produce.

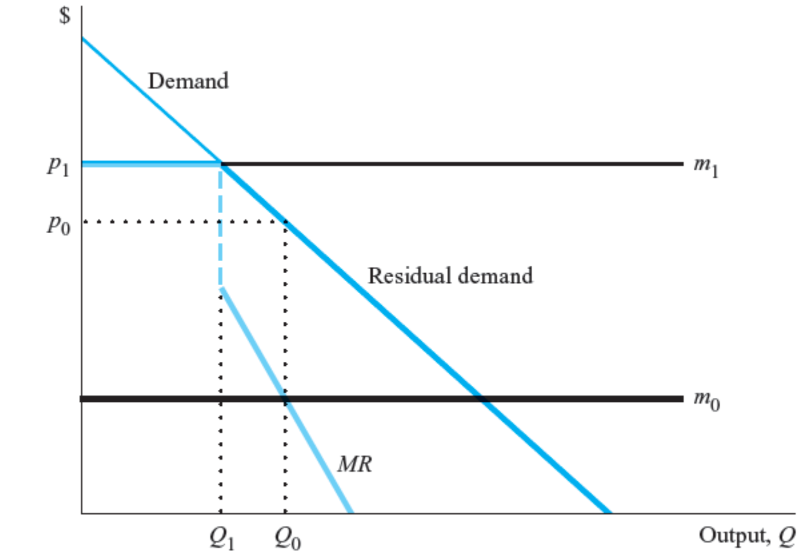
If m_0 is close to m_1, the firm may maximize its profit by selling at a price equal to p_1. However, in Figure 4.4, m_0 is enough less than m_1 that the profit-maximizing monopoly price is less than m_1 but above m_0. Because the residual demand curve has a kink in it at Q_1, the corresponding marginal revenue curve is discontinuous at the output, Q_1. The marginal revenue curve is horizontal where the residual demand curve is horizontal and slopes down where the residual demand curve is downward

EXAMPLE 4.4 *Preventing Imitation—Cat Got Your Tongue?*

Why are violin strings called *catgut* when they are really made of sheep intestines? An old Roman named Erasmo (c. 130 AD) started making strings for musical instruments out of sheep intestine. The demand grew. Because it was considered extremely bad luck to kill a cat, Erasmo identified his product as *catgut* so nobody would imitate it and ruin his monopoly.

Source: L. Boyd, "Grab Bag," *San Francisco Chronicle*, October 27, 1984:35.

FIGURE 4.4 Monopolization Through Efficiency



sloping. To maximize its profit, the firm with the secret process produces Q_0 units of output where its marginal revenue curve equals its marginal cost curve. The firm sets its price at p_0, which is less than $p_1 = m_1$, so no other firm remains in the market.

Government-Created Monopolies

A firm may be a monopoly because the government protects it from entry by other firms. For example, suppose a firm invents a new product and realizes that imitation *is* possible technically. In most countries, the original innovating firm can obtain legal protection to prevent entry for some period of time. The law on intellectual property, in particular the patent law, grants a legal monopoly to a firm that has discovered a new product or technique. A firm can obtain a patent (see Chapter 16) on a new product that prevents any other firm from copying its product and competing with it for a fixed period of years.

Aside from the patent laws, other types of government (or government sanctioned) restrictions on entry can serve to create and maintain monopolies. Generally, government restrictions on entry allow at least a few firms to produce, but they prevent the normal competitive forces from driving price and profits down to competitive levels (see Example 4.5).

For example, in many cities, one must purchase medallions, of which only a fixed number are sold by local authorities, to operate a taxicab. The United States, by granting exclusive (monopoly) rights to portions of the electromagnetic spectrum, gave broadcast television stations at least $40 billion in present-value terms for the first 30 years of television (Isé and Perloff 1997).

EXAMPLE 4.5 *Protecting a Monopoly*

An 1872 law established the U.S. Postal Service (USPS) monopoly on mail delivery. In 1971, the USPS started an express mail service. A 1979 amendment to the 1872 law broke the agency's monopoly on urgent mail, establishing as the definition of *urgent,* mail that must arrive by noon the next day or lose its value. However, the USPS has the right to decide what is urgent and what is not.

How serious was the Postal Service's competition on express mail? By 1994, the USPS's share of the express mail market had fallen to just under 15 percent. Worse, to the horror of postal officials, the federal government contracted with Federal Express for next-day delivery of government parcels at a price of $3.75, well below the Postal Service's overnight Express Mail rate of $9.95.

The USPS fought back. From 1990 through 1993, the Postal Service fined 21 companies for violating the USPS's legal monopoly on mail delivery, collecting more than $542,000 in fines from these companies that sent "nonurgent" mail by private couriers such as Federal Express, UPS, and DHL.

For example, an Atlanta-based credit-reporting company, Equifax Inc., was assessed a $30,000 penalty, making up the loss to the Postal Service for routine business mail it sent by express services. Postal officials say they recovered $4 in lost revenue for every $1 spent on enforcement.

In 1994, the USPS issued a postal inspectors' audit that found that five federal agencies—the General Services Administration (GSA) and the departments of Agriculture, Health and Human Services, Treasury, and Energy—routinely infringed on the USPS's monopoly on first-class mail by using Federal Express to ship materials that were not time sensitive.

The report warned that the agencies, which accounted for one-third of the 4.3 million government packages moved by Federal Express in the first two years of the contract, "are incurring a substantial liability for postage"—the revenue that would have otherwise gone to the Postal Service. The USPS did not demand any payments for the postage, but postal officials pressured the GSA to train federal mailroom personnel as to what kind of materials they can legally send by Federal Express.

Armed with news reports of USPS fines on private firms and pressure on federal agencies, outraged private companies went to Congress for legislation ending the Postal Service's practices. Smarting from bad publicity and congressional pressure, the USPS announced that it would cease its practice of raiding businesses to check up on their use of commercial overnight delivery services, and stopped complaining about federal agencies. Nice try, though.

Sources: Michael A. Goldstein, "Can the U.S. Postal Service Market Itself to Success?" *Los Angeles Times Magazine,* December 22, 1996:14; Bloomberg News, "UPS Aims to Curb Postal Service Monopoly," *The Dallas Morning News,* April 14, 1998:9D; Bill McAllister. "Must It Get There Overnight?: Agencies Improperly Bypassing Postal Service, Inspectors Report," *Washington Post,* January 12, 1994:A17; "Private Couriers and Postal Service Slug It Out," *New York Times,* February 14, 1994:D2.

Until recently, U.S. states required someone wishing to build an in-patient medical facility to obtain a certificate of need by demonstrating that a new facility was needed. Using these laws, an early entrant could make entry by potential competitors difficult. In part because of these laws, Community Psychiatric Centers, a chain of psychiatric hospitals in the United States and Britain, had annual earnings growth of 15 to 30 percent between 1969 when it went public and 1985.[11]

Similarly, trade barriers can be used to prevent entry. For example, in 1992, the Ontario government agency that monopolizes the sale of beer in that province, the Liquor Control Board of Ontario, announced a ban on American beer imports. Similarly, China places a 230 percent tariff (tax on foreign products) on foreign cigarettes to protect the China National Tobacco Corporation, which sells 1.75 trillion of the 5 trillion cigarettes sold throughout the world and accounts for 12 percent of the revenue of the Chinese government.[12]

Natural Monopoly

In some markets, it is efficient for only one firm to produce all of the output. When total production costs would rise if two or more firms produced instead of one, the single firm in a market is called a **natural monopoly**.

A firm is a natural monopoly if it can produce the market quantity, Q, at lower cost than can two or more firms. Let $q_1, \cdots, q_k$ be the output of the k (≥ 2) firms in a market that produce an identical product so that total market output equals the sum of the firms' output: $Q = q_1 + \cdots + q_k$. If each firm has a cost function $C(q_i)$ and one firm can produce Q at lower cost than the sum of the k firms,

$$C(Q) < C(q_1) + C(q_2) + \cdots + C(q_k),$$

then the least expensive (most efficient) way to produce is to have one firm produce all Q units. A cost function is said to be *subadditive* at Q if this inequality holds, so subadditivity is a necessary condition for the existence of a natural monopoly (Sharkey 1982; Baumol, Panzar, and Willig 1982).

A natural monopoly often has falling average costs and constant or falling marginal costs in the region in which it operates. A strictly decreasing average cost curve implies subadditivity (though the opposite does not necessarily follow).

Suppose that the average cost curve of a natural monopoly is downward sloping, and that the firm can produce 100 units at an average cost of $10 per unit. The firm's total cost of producing that many units is $1,000. Now suppose that a second firm with identical costs enters the market. If each of these two firms produces 50 units, their average cost of production is higher than before because the average cost curve is downward sloping. If their average cost is $15 per unit, for example, their combined

[11] See **www.aw-bc.com/carlton_perloff** "Model of Insanity."

[12] Glenn Collins, "U.S. Tobacco Industry Looks Longingly at Chinese Market, but in Vain," *New York Times*, November 20, 1998:A10.

total cost of producing 100 units is $1,500. Thus, a single firm can produce 100 units at lower cost than can two firms.

It is often argued (but may not be true) that electrical, gas, telephone, and cable television are natural monopolies. There is a relatively high fixed cost for running an electric power line or a phone line to a home or firm, but constant or falling marginal costs of supplying the service. As a result, marginal cost is constant or falls, and average cost falls as output increases.[13]

If production is characterized by economies of scale everywhere, then average cost declines as output increases, and it is always less costly for one firm to produce any given output than for several firms to produce that output. Therefore, when average cost falls with output, there is a natural monopoly. A natural monopoly can occur even if average cost is not declining everywhere with output. For example, if a U-shaped average cost curve reaches a minimum at an output of 100, it may be most efficient for only one firm to produce an output of 101 even though average cost is rising at that output. Therefore, economies of scale are a sufficient but not a necessary condition for natural monopoly.[14]

Profits and Monopoly

Many people associate high profits with monopoly or too little competition, normal profits with competition, and losses with excessive competition. Although each of these beliefs has some element of truth, none is correct. We now show why these beliefs do not hold in general by answering three questions: (1) Is anyone who earns positive profits a monopoly? (2) Does a monopoly always earn positive profits? (3) Should the government allow mergers that create monopoly in a market that was suffering short-run losses?

Is Any Firm That Earns a Positive Profit a Monopoly?

Although a monopoly may earn positive profits, it does not follow that any firm that earns a positive profit is a monopoly. The previous chapter discusses the possibility that certain scarce resources, such as land, can earn rents. For example, a wheat farmer who owns particularly productive land earns a large profit. This profit is attributable to the land that is owned and should properly be called a rent. The farmer behaves competitively, taking price as given and operating where price equals marginal cost. This

[13] The empirical literature, however, leaves some doubt as to whether many utilities exhibit increasing returns to scale, which implies downward-sloping marginal and average cost curves. Moreover, showing that there are scale economies in one range of output is not sufficient to demonstrate that a firm is a natural monopoly (that is, the cost function is subadditive). See, for example, Fuss and Waverman (1981) and Evans and Heckman (1982a, 1982b). Shin and Ying (1992) argue that local telephone exchange carriers were not natural monopolies prior to deregulation. Friedlaender (1992) finds evidence of substantial returns to scale for railroads.

[14] In Chapter 20, we examine how governments regulate natural monopolies and the conditions under which other firms will try to enter a market with a natural monopoly.

farm is a competitive firm; rents on factors of production do not indicate a monopoly. As long as output is not restricted so that price equals marginal cost, there is no market power. Scarce resources can command very high prices and those who own those resources benefit. For example, star athletes earn high salaries (rents) even though they are not monopolies that restrict output.

Does a Monopoly Always Earn a Positive Profit?

Although a monopoly earns a larger profit than a competitive firm would, it is not true that a monopoly always earns a positive profit. In the short run, a monopoly can make losses, just as a competitive firm can. A monopoly that faces a sudden decline in demand may continue to operate even though it makes a negative short-run profit (its price is less than its average cost) if its price is above its average variable cost. Losses in a market do not imply that it is competitive. In the long run, when there are no sunk costs, no firm continues to operate if there are only losses in the market.

As in competition, the length of time that losses will be earned by a monopoly depends on how long the short run lasts—how long it takes for the plant and equipment to wear out, forcing a decision on whether to replace them. In some markets, the short run may be very long. For example, railroad tracks can last for years or possibly decades. Therefore, one might expect that a monopoly railroad could earn a negative profit on its investments for a long time before deciding to exit the market.

Briefly, in the long run, a competitive firm makes zero economic profit, whereas a monopoly makes a zero or positive profit. In the short run, both competitive firms and monopolies may make losses or profits.

Are Monopoly Mergers to Eliminate Short-Run Losses Desirable?

A merger of firms into a monopoly can eliminate competition and allow the merged firm to exercise market power and raise price so that the losses are eliminated. Firms in a market where all firms are losing money often argue for a merger for this reason (see Example 4.6). This motivation for merger appears to have a certain logical appeal—if

EXAMPLE 4.6 *EU Allows Merger to Eliminate Losses*

In 2003, the European Union allowed Rupert Murdoch's News Corporation to merge Telepiu, which had two-thirds of all pay-TV subscribers in Italy, with its own Italian pay-TV firm, Stream, to create a new firm called Sky. EU Competition Commissioner Mario Monti conceded that his decision "will create a quasi-monopoly on the Italian market." He justified his actions by saying that a weak business environment allowed room for only one firm to survive in this market, as both Stream and Telepiu had been losing money.

Source: Raf Casert, "EU Commission Allows Murdoch's News Corp. to Forge 'Quasi-Monopoly' in Italian Pay TV," Associated Press, April 2, 2003.

the merger eliminates the losses, perhaps it is efficient for the merger to occur. However, such a merger harms society!

If a merger enables firms to set a price in the short run that is greater than the level at which they would have priced had they remained competitive, then the merger imposes a deadweight loss on society. The existence of sunk costs in the short run that cause short-run losses cannot be eliminated by merging firms. The merger only changes the amount of competition that firms face. Because the merger does not eliminate sunk costs, it is inefficient to allow firms to form a monopoly and thus allow the price to rise.

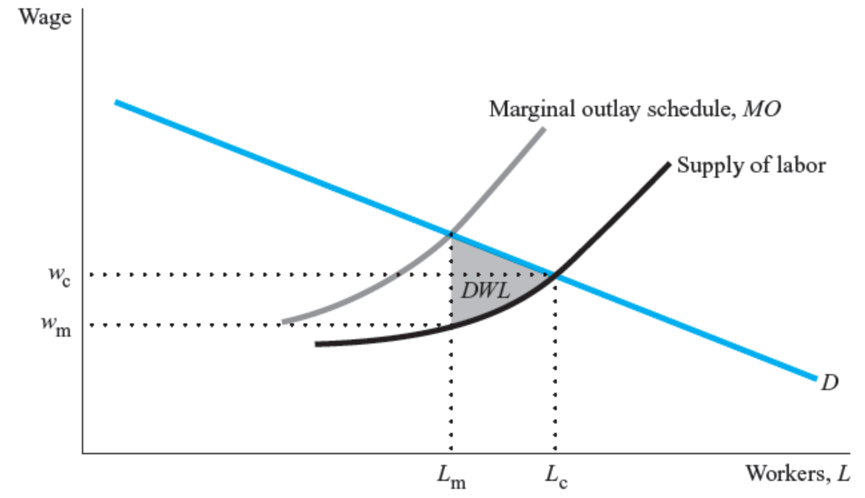
Monopsony

A single buyer in a market is called a **monopsony**. A monopsony's decision on how much to buy affects the price it must pay (just as a monopoly's choice of output affects the price it receives). The monopsony decides how much to purchase by choosing a price-quantity pair on the market supply curve. Monopsony is the flip side of monopoly. Both a monopoly and a monopsony recognize that their actions affect the market price.

A monopsony determines how much to buy in much the same way that a monopoly determines how much to produce. A monopsony buys more of the good as long as the value of the extra consumption as given by its demand curve equals or exceeds its marginal cost of consuming one more unit.

If there is a competitive labor market, each firm takes the wage rate as given, and the marginal cost of hiring one more worker is simply the wage rate. Now suppose there is only one local employer (buyer of labor services): a monopsony. In Figure 4.5,

FIGURE 4.5 Deadweight Loss from Monopsony



it faces an upward-sloping supply curve for labor. In order to hire an extra worker, the monopsony must not only pay that worker a slightly higher wage rate but also pay *all* its other workers a slightly higher wage rate, because only by raising the wage can extra labor be induced into the marketplace.

If the monopsony must raise its wage from, say, \$5 to \$6 to induce that last individual to work for the firm, the monopsony's extra cost of hiring the additional worker is not just \$6; it is \$6 plus the \$1 increase in wages that must be passed along to each of its original workers. If it originally had 100 workers, its total wage bill rises from \$500 to \$606: an increase of \$106. The monopsony recognizes that *its* marginal cost of hiring the additional worker is \$106 rather than \$6 and takes that into account in deciding whether to hire the additional worker. The monopsony hires an extra worker only if the marginal benefit as given by its labor demand curve exceeds its marginal cost of hiring an additional worker.

The marginal cost to a monopsony of buying additional units (hiring additional workers) is described by a **marginal outlay schedule**, which is analogous to a marginal revenue curve. As Figure 4.5 illustrates, the marginal outlay schedule lies above the upward-sloping supply curve because the monopsony must raise the wage for all its workers to hire an extra worker. A profit-maximizing monopsony hires L_m workers, where its marginal benefit, as given by its demand curve, equals its marginal outlay. Because the marginal outlay curve lies above the supply curve, the monopsony hires fewer workers, L_m, than would a competitive market, which hires L_c workers (determined by the intersection of the demand curve and the supply curve). In other words, a monopsony restricts output just as a monopoly does.

The monopsony wage rate, w_m, is below the competitive wage rate, w_c. Using a definition analogous to the one for market power, we can define monopsony power as the ability to profitably set wages (or other input prices) below competitive levels. At the monopsony solution (L_m, w_m) in Figure 4.5, there is a gap between the demand curve and the supply curve. A gap between the demand curve (which represents the marginal benefit to society of consumption) and the supply curve (which represents the marginal cost to society) reflects a loss in efficiency. The monopsony deadweight loss triangle (Figure 4.5) is analogous to the deadweight loss that results from monopoly (Figure 4.2a).

Most labor economists believe there are few monopsonized labor markets in the United States. Example 4.7 identifies one such market. The most frequent examples given of monopsony in the labor market concern single-company towns, local employment markets, and sports leagues. For example, a major league baseball player can work in the United States only if he plays for a team that belongs to either the American or National Leagues. Collectively, these teams are the sole buyer in the United States for the services of a major league baseball player. To the degree that the teams agree not to compete for players, they gain monopsony power. To offset such monopsony power, baseball players can form a union to obtain monopoly power in selling their labor services.

Monopsony is most likely in markets where resources are specialized to a few uses. Moreover, even if resources are initially specialized to one use, as with a piece of custom-designed machinery (or a plant in a specific location serving a single buyer),

EXAMPLE 4.7 *Priest Monopsony*

Newspapers repeatedly report a "shortage" of Catholic priests, citing sources both inside and outside the church hierarchy. Between 1960 and 2000, the number of priests declined 13 percent even as 55 percent more Catholics had joined their local parishes. Strikingly, other religious denominations are not suffering from a "shortage of clergy."

Why the difference between churches? Is it due to changing tastes among potential priests or some other factor? Daniel Condon (2002) attributes the difference to the fact that the Catholic Church exercises monopsony power, whereas other churches and synagogues permit an active, competitive labor market for clergy. Individual Catholic parishes do not compete for clergy, who are instead assigned by the central church authority (diocese). Wages vary little across parishes, although priests in wealthy parishes may receive larger fees for performing wedding and funeral services.*

Condon estimates that Catholic priests earn 41 percent less than non-Catholic officiants, controlling for education, experience, location, and whether they are provided rent-free housing. He concludes (p. 918) that the true differential is "more pronounced when one considers that Catholic clergy have a condition of employment (celibacy) that would require additional monetary compensation for most."

*Academics at a Catholic university may face an even greater problem. For them, it's publish or parish.

monopsony may not persist in the long run. The reason is that no one will make new custom-designed machinery (or new investments in a plant) for a specific buyer if they earn a depressed return compared to what they can earn from making other machines (or building a plant elsewhere). In other words, few resources are specialized in the long run, and therefore it is unlikely that monopsony can persist in the long run.

Another way to explain the preceding point is as follows. If resources are not specialized to a particular market in the long run, then the long-run supply curve tends to be flat (highly elastic). As Chapter 3 explained, a flat long-run supply curve is most likely to occur when the market in question uses only a relatively small fraction of the total consumption of its inputs. Long-run monopsony power is impossible if the long-run supply curve is flat because price cannot be lowered below the competitive price.

If the long-run supply curve is flat, there may not be any monopsony power even in the short run. Suppose that before a firm enters a market, it has many alternative uses for the resources it owns. After it enters the market, it specializes its machines so that it has very few alternative uses for its assets. Suppose that it will only enter a particular market if it receives $10 per unit of output (which is the long-run average cost). The sole buyer, the monopsony, agrees to pay $10. After the firm enters, it is committed, at least for some time, in the sense that its machines are specialized to this particular market. If the monopsony lowers its price to $9, it may not pay for the firm to exit immediately. But the firm will not replace the specialized machines when they wear out, and the monopsony may eventually have no one willing to supply the product. Even if the buyer again promises $10 per unit to induce a supplier to enter, no firm would believe the buyer in light of its previous behavior. So, for a buyer that is concerned about a long-run source of supply, it may not pay to exercise short-run monopsony power.

Dominant Firm with a Competitive Fringe

Where does the gorilla sleep?
Anywhere the gorilla wants to sleep.

What happens to a monopoly if other, higher-cost firms enter its market? Or, similarly, what happens if a lower-cost firm enters a market with many price-taking, higher-cost firms? After entry, the lower-cost firm has a relatively large share of the market. If one firm is a price setter and faces smaller, price-taking firms, it is called a **dominant firm**. It typically has a large market share. The smaller, price-taking firms, called **fringe firms**, each have a very small share of the market, though collectively they may have a substantial share of the market.

There are several industries in which one firm has a large share of the industry sales. For example, Kodak's share of the photographic film business has been estimated at 65 percent.[15] Hewlett-Packard is estimated to have 59 percent of laser printer sales.

We begin by discussing what makes a firm dominant. We then analyze how entry limits a dominant firm's market power. We examine two extreme cases. In the first, entry by other firms is impossible. In the second, entry by competing fringe firms can occur instantaneously. The analysis shows that a dominant firm's price-setting behavior depends on the ease of entry by fringe firms.

We draw two main conclusions. First, it is generally not in a profit-maximizing dominant firm's best interest to set its price so low that it drives all competitive-fringe firms out of the market. Second, the presence of competitive-fringe firms or the threat of entry by additional firms may force a dominant firm to set a price lower than the price a monopoly would set (see Example 4.8).

If a sufficiently large number of price-taking firms can enter the market, a dominant firm cannot continue to charge a price higher than the minimum average cost of these new firms. Indeed, if potential entrants' costs are as low as the dominant firm's, the dominant firm eventually has no more market power than any other firm.

[15] A firm's share of sales in an industry depends crucially on how the industry is defined, and hence is often controversial, especially in court proceedings.

EXAMPLE 4.8

Price Umbrella

It is often asserted that a dominant firm provides a *pricing umbrella* for smaller firms. As long as competing firms price at or below the level of the dominant firm, they will be able to find buyers. If their products are inferior (say because they are risky to use for legal reasons), the fringe firms have to set their prices substantially below the dominant firm's.

In many countries, phone monopolies charge rates that are more than twice those in the United States, where competition has kept rates relatively low. This price difference causes problems for the monopolies.

"Callback" services offer some customers a way to evade paying high monopoly prices. A callback service provides a "trigger" number connected to a computer in the United States. The customer calls that number using the monopoly service and hangs up before the phone is answered, paying nothing for the incomplete call. The computer calls the customers back and offers an American dial tone, which can be used to place a call anywhere in the world for rates well below the monopoly price. In some cases, the callback rates are less than the price of a local call. Hundreds of American companies provide these services, and the rate of use has grown exponentially over time. Ghana's monopoly is reported to lose $1 million each week to callback and Internet services.

To protect local monopolies, governments in many countries—including Argentina, Canada's Northwest Territories, China, Malaysia, Saudi Arabia, South Korea, and Uganda—try to stop these services. The U.S. operators believe they are beyond the reach of local laws. For example, when Uganda blocked all calls to the Seattle, Washington, area code where one service, Kallback, is based, the company routed the calls through a different area. When other countries tried to identify and block the services by picking up the touch-tone beeps used to complete calls, Kallback added a voice-recognition system. As a firm spokesman said, "It's a cat and mouse game. It's kind of fun."

Source: "Don't Call US," *The Economist*, 338(7947), January 6, 1996:55; **www.kallback.com**; "Telecom Loses $1m a Week, Communications Experts Say," *Ghanaian Chronicle*, February 7, 2003.

Why Some Firms Are Dominant

> *All animals are equal, but some animals are more equal than others.*
>
> *—George Orwell*

Why do some firms gain substantial market power, while others do not? At least three possible reasons are sufficient to create a dominant firm-competitive fringe market structure.

The first reason is that *dominant firms may have lower costs than fringe firms.* There are at least four major causes of lower costs:

- A firm may be more efficient than its rivals. For example, it may have better management or better technology that allows it to produce at lower costs. Such a technological advantage may be protected by a patent.
- An early entrant to a market may have lower costs from having learned by experience how to produce more efficiently.
- An early entrant may have had time to grow large optimally (in the presence of adjustment costs) so as to benefit from economies of scale. By spreading fixed costs over more units of output, it may have lower average costs of production than a new entrant could instantaneously achieve.
- The government may favor the original firm. The U.S. Postal Service does not pay taxes or highway user fees, which reduces its cost relative to that of competing package delivery services.

A second important reason is that *a dominant firm may have a superior product* in a market where each firm produces a differentiated product. This superiority may be due to a reputation achieved through advertising or through goodwill generated by its having been in the market longer.

A third reason is that *a group of firms may collectively act as a dominant firm.* As Chapter 5 shows, groups of firms in a market have an incentive to coordinate their activities to increase their profits. A group of firms that explicitly acts collectively to promote its best interests is called a *cartel.* If all the firms in a market coordinate their activities, then the cartel is effectively a monopoly; if only some of them do so, then the group acts as a dominant firm facing a competitive fringe of noncooperating firms.

One example of a dominant firm is the cartel consisting of Philippine coconut-oil-producing firms that act in concert but face a fringe of firms in other countries that act as price takers. With nearly four-fifths of the world's export market, the Philippine cartel has dominant-firm market power with a Lerner Index of 0.89 (Buschena and Perloff 1991).

Whether a dominant firm can exercise market power in the long run depends crucially on the number of firms that can enter the market, how their production costs compare to those of the dominant firm, and how fast they can enter. We now examine the dominant firm-competitive fringe model under two alternative extreme assumptions about the ease of entry.

The No-Entry Model

Consider a market with a dominant firm and a competitive fringe in which no additional fringe firms can enter. Two key results emerge from an analysis of this model: (1) It is more profitable to be the *gorilla* of a market than a mere fringe firm. (2) The existence of the fringe limits the dominant firm's market power—that is, it is more profitable to be the only firm in a market (a monopoly) than merely a dominant firm.

Assumptions. Five crucial assumptions underlie this no-entry model:

1. *There is one firm that is much larger than any other firm because of its lower production costs.* Although a market may be characterized by a small group of relatively large firms rather than a single dominant firm, we concentrate on the case of the single dominant firm for simplicity.
2. *All firms, except the dominant firm, are price takers,* determining their output levels by setting marginal cost equal to the market price (p).
3. *The number of firms (n) in the competitive fringe is fixed: No new entry can occur.* That is, the dominant firm knows that it can raise the market's price without causing new firms to enter the market or existing firms to build additional plants.
4. *The dominant firm knows the market's demand curve, $D(p)$.* Each firm produces a homogeneous product, so that there is a single price in this market.
5. *The dominant firm can predict how much output the competitive fringe will produce at any given price;* that is, it knows the competitive fringe's supply curve, $S(p)$.

The first three assumptions determine that this market has a dominant firm facing a competitive fringe with no more than n firms. The last two assumptions ensure that the dominant firm knows enough to be able to set its output level optimally.

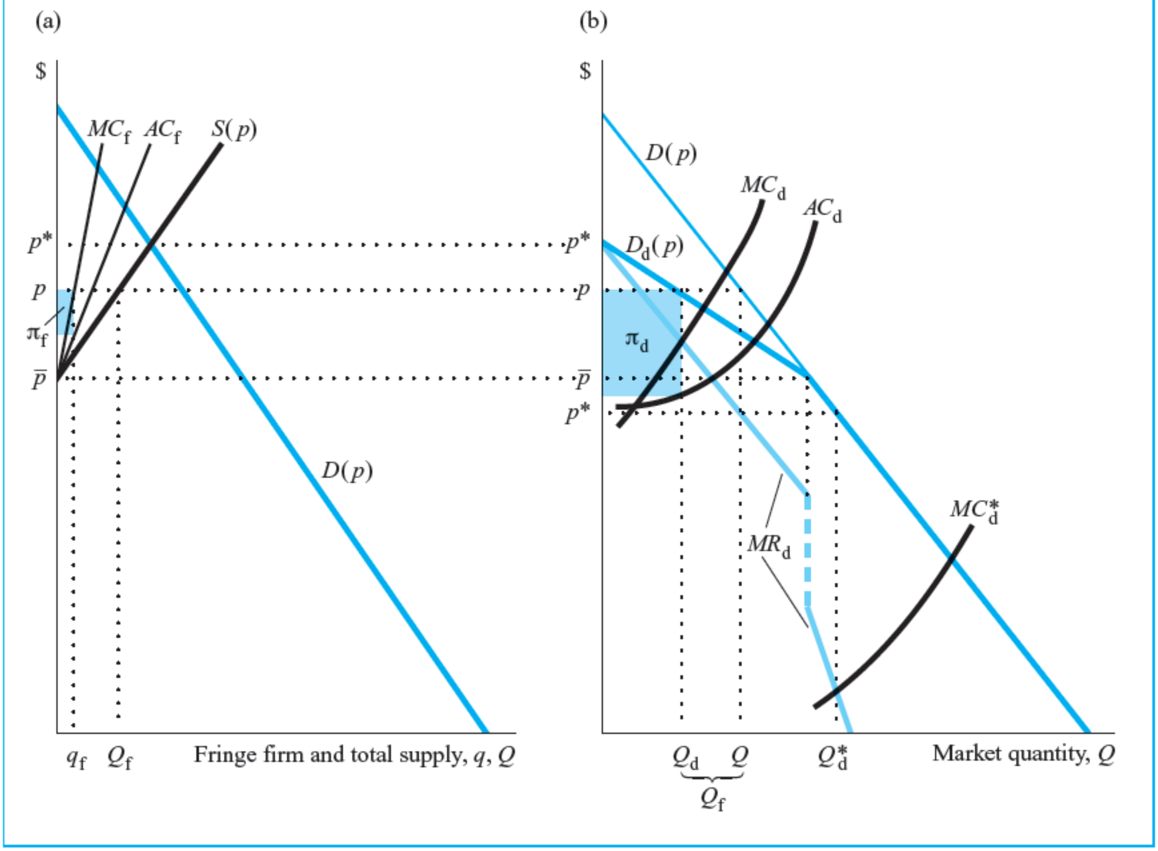
The Dominant Firm's Reasoning. Suppose you ran the dominant firm. How would you choose your output level? Given your firm's large size, you could drive up the market's price by restricting your output. Unfortunately for you, as your dominant firm lowers its output and price rises, the competitive fringe output increases because the fringe supply curve, $S(p)$, is increasing in p. As a result, market output falls less than you would like, and the market price does not rise as high as it would if your firm had a monopoly.

Thus, your dominant firm's problem is much more complex than that of a monopoly, which merely needs to consider the market demand curve (with its corresponding marginal revenue curve) and its marginal cost curve to determine its profit-maximizing output. Your dominant firm, in contrast, must consider not only those factors, but also how the competitive fringe responds to your actions.

To maximize your profits, you must take the competitive fringe's actions into account when setting your policy. A convenient way to calculate your optimal price level is to do the following thought experiment. For lack of an ability to stop them, let the fringe firms sell as much as they want at the market price: the price you set. Except at the very highest prices, the competitive fringe does not produce enough to meet all of the market's demand. Your dominant firm, then, is in a monopoly position with respect to this residual demand. Thus, you can determine your optimal output by a two-step procedure. First, determine your firm's residual demand curve; then, act like a monopoly with respect to the residual demand. This two-step procedure can be illustrated with the use of graphs.

A Graphic Analysis of Dominant-Firm Behavior. The first step is to determine the long-run residual demand curve facing the dominant firm. Figure 4.6 shows two graphs: (a) one for a representative competitive-fringe firm and for the entire competitive fringe, and (b) one for the dominant firm.

FIGURE 4.6 The Dominant Firm and the Competitive Fringe



The graph on the left, Figure 4.6a, shows the market demand curve, $D(p)$, and the supply curve of a typical, price-taking, competitive-fringe firm. The fringe firm's supply curve is its marginal cost curve above the minimum of its average cost curve $\underline{p}$. That is, the fringe firm's shutdown price is $\underline{p}$. Above $\underline{p}$, each fringe firm makes positive economic profits. At $\underline{p}$, each fringe firm makes zero profits and is indifferent between operating and shutting down.[16] Below $\underline{p}$, each firm shuts down, and the dominant firm is a monopoly.

The competitive fringe's supply curve, $S(p)$, is the horizontal summation of the individual fringe firm's supply curves, as Figure 4.6 shows. That is, $S(p) = nq_f(p)$, where n is the number of firms and q_f is the output of a typical fringe firm.

[16] As drawn, each fringe firm produces essentially no output at $\underline{p}$. If the firms had the usual U-shaped average cost curves, however, they would produce a positive amount of output at that price.

The dominant firm's residual demand curve is the horizontal difference between the market demand curve and the competitive fringe's supply curve:

$$D_d(p) = D(p) - S(p).$$

In Figure 4.6b, the market demand curve (thin blue line) is above the residual demand curve (heavy blue line) at prices above $\bar{p}$ and equal to it at prices below $\bar{p}$. That is, the fringe firms meet some or all of the market demand if price is above $\bar{p}$, but they drop out of the market and leave all of the demand to the dominant firm if price falls below $\bar{p}$. At p^*, the quantity that the fringe supplies equals the quantity that the market demands, so the dominant firm has no residual demand.

The dominant firm maximizes its profits by picking a price (or equivalent, an output level) so that its marginal cost equals its marginal revenue. The dominant firm's marginal revenue curve (MR_d) is derived from its residual demand curve and has two distinct sections. If the competitive fringe produces positive levels of output, the dominant firm's residual demand curve lies below (and is flatter than) the market demand curve. The dominant firm's marginal revenue curve, MR_d, in this region is flatter than the marginal revenue curve in the region where the dominant firm's residual demand curve and the market demand curve are coincident. There is a discrete jump between the two sections of the marginal revenue curve at the point where the residual demand curve and the market demand curve meet.

The dominant firm behaves as a monopoly would with respect to the residual demand; it sets its price (or output) so that its marginal cost equals marginal revenue. Because the marginal revenue curve has two sections, there are two possible types of equilibria; which one occurs depends on the dominant firm's cost curves.

We now consider two types of markets:

1. The dominant firm charges a high price, so that it makes economic profits and the fringe firms also make profits or break even.
2. The dominant firm sets a price so low that the fringe firms shut down to avoid making losses. The dominant firm is now a monopoly.

The Dominant Firm–Competitive Fringe Equilibrium

The first type of equilibrium occurs if the dominant firm's costs are not substantially less than those of the fringe firms.[17] The dominant firm's marginal cost curve, MC_d, crosses the first downward-sloping segment of the marginal revenue curve, MR_d, in Figure 4.6b.

The dominant firm chooses to produce Q_d level of output at price p (the height of the residual demand curve at the output level Q_d). At the price level p, the difference between the market demand, Q, and the dominant firm's output, Q_d, is the competitive fringe's supply, Q_f (which is shown in Figures 4.6a and 4.6b). If the dominant

[17] A mathematical analysis of this case is presented at **www.aw-bc.com/carlton_perloff** "Dominant Firm and Competitive Fringe Model."

firm's costs are this high, it does not drive the competitive fringe out of business. Its own profits are maximized at a price so high that the fringe firms make positive profits.

In most markets, positive economic profits would attract new entrants. In this market, however, no new firms can enter (by assumption), so both the dominant firm and the competitive fringe firms can make positive profits forever. In Figure 4.6b, the dominant firm's profits are labeled π_{d}. The profits of a typical fringe firm are positive as well (because $p > \bar{p}$), and a typical fringe firm's profits are shown as π_{f} in Figure 4.6a. Because the dominant firm's average cost is lower than that of the fringe firms (minimum $AC_{\text{d}} < \bar{p}$), the dominant firm makes more profits per unit (average profits), and it also sells more units than an individual fringe firm, so it must make more total profits as well.

Thus, the dominant firm maximizes its profits by charging a price so high that it loses some of its market share to the competitive fringe. It does not make sense for the dominant firm to set its price so low that it drives the fringe out of business, even though that would increase the number of units of output the dominant firm could sell. After all, few good business people accept the argument, "I lose a little on every sale, but make up for it in volume."

The dominant firm makes lower profits than it would if it were a monopoly and the fringe did not exist. The fringe can only hurt the dominant firm and benefit consumers. For example, in 1993, NEC Corporation, which then controlled half of all personal computer sales in Japan, had to cut its prices roughly in half due to increased competition from U.S. fringe firms.

The Dominant Firm as Monopoly. Now, suppose that the dominant firm has extremely low costs compared to the fringe firms, so that its marginal cost curve is MC^*_{d} in Figure 4.6b. Notice that MC^*_{d} crosses MR_{d} in the lower part of its two downward-sloping sections. The dominant firm chooses to produce Q^*_{d} level of output at price p^* (the height of the residual demand curve at output level Q^*_{d}). Because p^* is below the fringe firms' shutdown point ($\bar{p}$ = their minimum average cost), the fringe firms produce nothing ($Q^*_{\text{f}} = 0$). As a result, market output, Q^*, equals the dominant firm's output, Q^*_{d}.

The dominant firm sets a monopoly price, and no competitive-fringe firm enters. The dominant firm meets all the demand of the market, unchecked by the fringe, and is thus a monopoly. The reason it has a monopoly is that MC^*_{d} intersects MR_{d} along the segment of MR_{d} that is the same as the marginal revenue curve associated with the market demand curve. That is, the monopoly price is below $\bar{p}$, so no fringe firm wants to produce.

A Model with Free, Instantaneous Entry

If unlimited entry is possible, a dominant firm cannot set as high a price as it can if entry is limited or prevented. This section retains all the assumptions made in the preceding section except that now an unlimited number of competitive-fringe firms may enter the market. Firms enter if they can make positive profits.

In this situation, fringe firms cannot make profits in the long run; they either break even or are driven out of business. If identical fringe firms produce at all, the market

EXAMPLE 4.9 *China Tobacco Monopoly to Become a Dominant Firm*

Established in 1982, the Chinese government's tobacco monopoly, the China National Tobacco Corporation, has been the most profitable corporation in the world, accounting for 12% of the Chinese government's revenues. It sells to China's 310 million smokers, a quarter of the world's smoking population, who consume 1,700 billion cigarettes a year—about 30% of global consumption.

By imposing a 230% tax rate on foreign cigarettes, and by imposing import quotas and restrictions (such as designating only a few sales outlets for imported cigarettes), the government limited legal foreign cigarette sales to less than 2% of total Chinese sales in the late 1990s. By 2003, their share was only 10%.

To appease the World Trade Organization (WTO), China has agreed to lift restrictions on the retail sale of imported cigarettes by January 2004, to reduce the tariff on cigarettes from the current 65% to 24%, and to phase out the tariff over the next two years.

Thus, the state's monopoly will be turned into a dominant firm. Government officials expect that the price of imported cigarettes will drop in half, and that they will gain a major share of the market.

Sources: Glenn Collins, "U.S. Tobacco Industry Looks Longingly at the Chinese Market, but in Vain," *New York Times*, November 20, 1998:A10; "China to Lift Restrictions on Retail Sales of Imported Cigarettes Next Year," *AFX European Focus*, February 11, 2003; "Remove of Foreign Tobaccos Retailing Licenses to Cut Prices by Half" (sic), *China News*, February 14, 2003:1; "Chinese Tobacco Industry Facing Mergers and Recapitalizations," *China Business Times*, February 17, 2003:1.

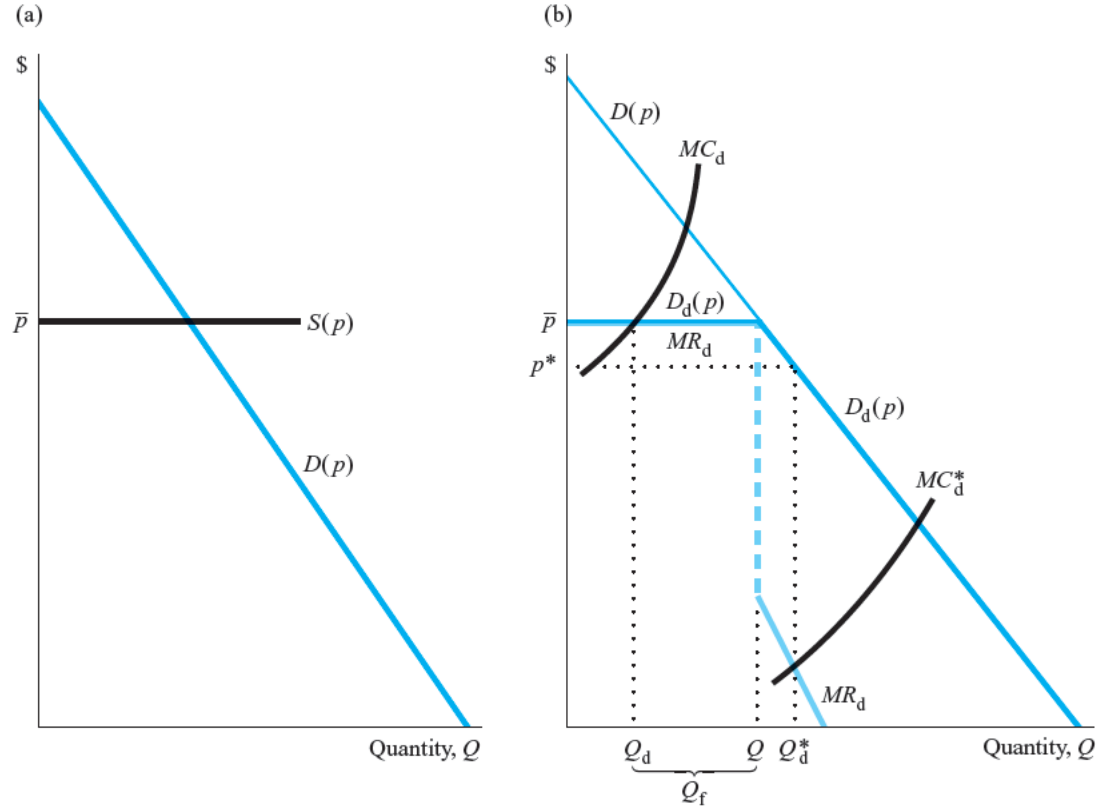
price ultimately can go no higher than a fringe firm's minimum average cost, so that fringe firms always just break even. After all, if they made positive profits, more firms would flood into the market and drive price down to the level where each earns zero economic profits. Because the dominant firm has lower costs than fringe firms, it makes positive profits, but its profits are lower than if entry did not occur.

Even with unlimited entry, the dominant firm can gain and hold indefinitely a large share of the market if it has some cost or other advantage (see Example 4.9). Another example is the Cheerleader Supply Co., which accounts for 60 percent of cheerleading uniforms and equipment sold in this country.[18] This is an industry with easy entry, and yet one firm has the lion's share of the market, presumably because it has superior products, a superior sales force, lower costs, or has generated goodwill with buyers.

The competitive-fringe firms' cost curves are the same as before. As more and more firms enter (n rises), the slope of the competitive-fringe supply curve becomes flatter

[18] According to its chief executive officer, Lawrence Herkimer, in Peter Applebome, "The World's Oldest and Fattest Cheerleader," *San Francisco Chronicle*, January 12, 1984:24.

FIGURE 4.7 Dominant Firm with Free, Instantaneous Entry by Fringe Firms



and flatter (it is n times the slope of a typical firm's supply, or MC, curve). As the number of firms grows large, the fringe's supply curve becomes essentially horizontal, as shown in Figure 4.7a. That is, as long as price is at least $\bar{p}$, the competitive fringe is capable of and is willing to supply any quantity that the market demands.

As shown in Figure 4.7b, the residual demand curve facing the dominant firm is horizontal at $\bar{p}$ so the corresponding marginal revenue curve is also flat (remember that in a competitive market a firm faces a horizontal demand curve, and hence its marginal revenue curve is identical to its demand curve at the market price). Below $\bar{p}$ the residual demand curve is the market demand, which slopes downward, so that the corresponding marginal revenue curve also slopes downward. Again, the marginal revenue curve corresponding to the residual demand curve jumps at the quantity where the kink in the residual demand curve occurs.

There are two possible equilibria. First, if the dominant firm's marginal cost is relatively high (MC_d in Figure 4.7b), so that it intersects the horizontal portion of the

MR_d curve, the price is $\bar{p}$, and the competitive fringe meets some of the market's demand. At this price, each fringe firm makes zero economic profits (because its average cost equals $\bar{p}$) and is indifferent between staying in business and leaving the market. How much is produced by the competitive fringe depends on the dominant firm's cost structure (that is, where MC_d intersects the horizontal marginal revenue curve), which determines the dominant firm's output, Q_d. Collectively, the fringe firms produce an output level $Q_f = Q - Q_d$, as Figure 4.7b shows.[19] It is possible that $Q_f = 0$ even though the presence of the fringe constrains price to equal $\bar{p}$.

Thus, if fringe firms flood into a market whenever positive profits can be made, the dominant firm cannot charge a price above the minimum average cost of a fringe firm. Although a dominant firm can make positive profits, competitive-fringe firms just break even. If the dominant firm's price would be above $\bar{p}$ in the absence of entry, consumers are better off if entry is possible because it results in lower prices.

The second type of equilibrium occurs if the dominant firm's marginal cost is lower (MC^*_d in Figure 4.7b), so that it hits the marginal revenue curve in the downward-sloping portion. Here, the price is so low that no fringe firm stays in the market when the dominant firm's costs are lower than the fringe firms' costs. This equilibrium (Q^*_d, p^*) is the same as discussed previously in the second no-entry equilibrium and is shown in Figures 4.6b and 4.7b. The dominant firm is a monopoly, and the potential supply of fringe firms is irrelevant.

SUMMARY

Monopoly or market power is the ability to price profitably above marginal cost. A single seller of a product, a monopoly, faces a downward-sloping demand curve and sets its price above marginal cost. As a result, less is purchased than if the market were perfectly competitive and society suffers a deadweight loss.

In some markets, however, there are benefits to monopoly. For example, the promise of future monopoly profits can spur a firm to develop new products or more efficient production techniques.

Not all firms that earn profits are monopolies, and not all monopolies earn profits. Just like a competitive firm, a monopoly can make either profits or losses in the short run. However, unlike a competitive firm, a monopoly can earn positive profits in the long run. A natural monopoly exists when it is efficient to have only one firm produce the market's output.

[19] Why don't fringe firms meet the entire demand at $\bar{p}$, instead of splitting it with the dominant firm? The answer is that the dominant firm has lower costs and can force some of the fringe firms out of the industry. Suppose that the dominant firm is producing its desired output of Q_d, and n fringe firms are producing $Q_f = Q - Q_d$. Now, if additional fringe firms enter this market, output exceeds market demand at $\bar{p}$. For the market to clear, the price must fall. Since the dominant firm is making positive profits, it stays in the industry. The fringe firms, however, start making losses (because they just break even at $\bar{p}$). Thus, some of the fringe firms must drop out of the industry until the price again rises to $\bar{p}$. Alternately stated, the dominant firm can always charge slightly below $\bar{p}$ to sell as much as it wants.

Monopsony is monopoly on the buying side. A firm with monopsony power sets lower prices and employs fewer resources than would prevail under competition. Like monopoly, monopsony imposes an efficiency cost on society. Monopsony power can persist only when resources are specialized in the long run.

A low-cost dominant firm has market power even though it competes with other firms. A profit-maximizing dominant firm does not attempt to drive out fringe firms at all costs. Its behavior depends on how great its cost advantage over fringe firms is and on how easily other firms can enter. If a large number of price-taking firms can enter the market whenever a profit opportunity occurs, and if they can produce at costs not much above those of the dominant firm, the dominant firm is unable to charge prices substantially above the competitive price. Even if fringe firms do not enter a market, the threat of their entry may cause a monopoly (in the sense that it is the only firm in the market) to set a lower price than it would in the absence of the fringe.

PROBLEMS

1. Does a monopoly's profit differ if it chooses price or quantity (assuming it chooses them optimally)? Why can't a monopoly choose both price and quantity?

2. After a shift in the demand curve, show that a monopoly's price may remain constant but its output may rise.

3. If the demand curve is $Q(p) = 5/p$, what is the elasticity of demand? What is total revenue when $p = \$1$ and when $p = \$30$? If production costs \$1 per unit, and the smallest production level is 1 unit, how much should the monopoly produce?

4. If the demand curve is $Q(p) = p^{\epsilon}$, what is the elasticity of demand? If marginal cost is \$1 and $\epsilon = -2$, what is the profit-maximizing price?

5. Suppose the demand curve for corn is $Q(p) = 10 - p$. Suppose that one firm owns all five units of corn in the world and has zero marginal cost. Does a monopoly sell less output than would be sold in a competitive market in which 100 firms each own 0.05 units?

6. Suppose the Environmental Protection Agency sets new requirements that raise the (fixed) costs of reporting compliance with pollution control rules (Pashigian 1984). How would this change affect (a) the market price, (b) the number of fringe firms, (c) total output, and (d) the dominant firm's share of the market? *Hint:* What does an increase in fixed costs do to the average cost curve of a fringe firm?

7. By showing the behavior of both a monopoly and a dominant firm in the same graph, show that monopoly profits are greater than the profit of a dominant firm in the no-entry equilibrium (MC_{d}). Show how much consumers benefit from buying from a dominant firm-competitive fringe rather than from a monopoly. *Hint:* A firm's variable costs are the area under its marginal cost curve up to the relevant output.

8. How would the no-entry model diagrams (Figure 4.6) change if fringe firms had the usual U-shaped average and marginal cost curves? Assume that because of a barrier to entry, there are only *n* fringe firms. Describe the types of possible equilibria.

9. Would a profit-maximizing dominant firm ever produce more than if it were a monopoly? *Hint:* Show the behavior of both a monopoly and a dominant firm (in the no-entry model) on the same graph and note where the marginal revenue curves cross.

10. What effect does a binding minimum wage have on a monopsony labor market?

Answers to odd-numbered problems are given at the back of the book.

SUGGESTED READINGS

Stigler (1965) provides a good, nontechnical introduction to the dominant firm-competitive fringe model. Fisher, McGowan, and Greenwood (1983) is a very readable, controversial discussion of the important IBM antitrust case.

CHAPTER 5

Cartels

People of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the public, or in some contrivance to raise prices. It is impossible indeed to prevent such meetings, by any law which either could be executed, or would be consistent with liberty and justice. But though the law cannot hinder people of the same trade from sometimes assembling together, it ought to do nothing to facilitate such assemblies; much less to render them necessary. —*Adam Smith*

In any market, firms have an incentive to coordinate their production and pricing activities to increase their collective and individual profits by restricting market output and raising the market price. An association of firms that explicitly coordinates its pricing or output activities is called a cartel. A cartel that includes all firms in a market is in effect a monopoly, and the member firms share the monopoly profits.

Cartels are more likely to occur in *oligopolistic markets*, where there are only a few firms, than in competitive markets. It is easier to reach and maintain an agreement on price or output when the number of firms is small. But even without an explicit agreement, firms may act so as to raise their collective profits. The factors that increase the likelihood that a cartel succeeds also affect whether oligopoly firms can elevate their prices above the competitive level. Thus, a study of cartels is also a study of oligopolies. In the next chapter, we use game theory to analyze oligopoly behavior.

Fortunately for consumers, although firms have an incentive to coordinate activities to restrict market output and raise prices, each member of the cartel has an incentive to "cheat" on the cartel agreement. Each cartel member wants to produce more output than is best for the cartel collectively. As a result, cartels tend to break apart even without government intervention.

When a cartel partially breaks apart so that some firms act outside of the cartel or when not all firms in the market join the cartel in the first place, the cartel may act like a dominant firm facing a competitive fringe of nonmember firms. As discussed in Chapter 4, entry of new fringe firms into a market can destroy the market power of a dominant firm or a cartel. Thus, only cartels that do not fall apart through lack of cooperation and that exist in markets in which entry is difficult can maintain market power for substantial lengths of time.

Four key questions are examined in this chapter:

1. Why do cartels form?
2. What factors cause some cartels to last and others to break up, even without government intervention?
3. How harmful are cartels?
4. What have governments done about cartels?

Why Cartels Form

United we stand, divided we fall.
Union gives strength. —*Aesop*

Why is Adam Smith correct that firms want to form cartels? The answer is that each individual firm wants to increase its own profit. But why should a firm's profit go up when the firms in a market form a cartel? After all, each competitive firm is maximizing its profit. How can the firms do better by forming a cartel if each is already maximizing its profit?

The answer involves a subtle argument. In a competitive market, each firm considers how much a reduction in its own output benefits it and ignores the gains to other firms, which benefit from a reduction in total market output to the extent that reduction raises the price. In contrast, a cartel takes into account the benefits to all its members of the reduction in each firm's output. Thus, a competitive market (in which each firm ignores the collective gain from its output reduction) produces more output than a cartel.

To illustrate the nature of this collective gain, consider two polar cases. First, suppose that a market is made up of many identical, competitive firms, each of which is a price taker. In contrast, suppose that all the firms join together to form a cartel and act as a monopoly. Figure 5.1a shows a typical firm's marginal cost curve. The sum of the individual firm's marginal cost curves is the market supply curve, which is shown in

FIGURE 5.1 Cartel

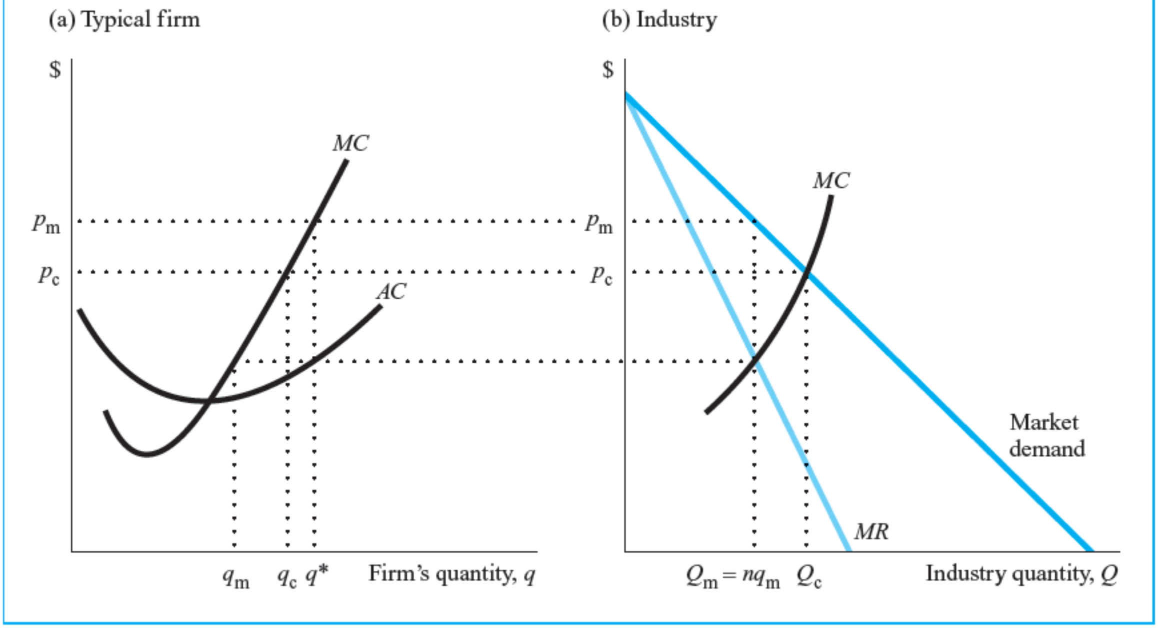


Figure 5.1b (labeled *MC*) along with the market demand curve. The competitive output, Q_c, is determined by the intersection of this supply curve with the market demand curve (Figure 5.1b), with each firm producing q_c units of output (Figure 5.1a) and the market price is p_c.

Why does it pay for the cartel to reduce output from the competitive level?[1] At the competitive output, the cartel's marginal cost is greater than its marginal revenue (Figure 5.1b), so it pays the cartel to reduce its output. Because the demand curve slopes downward, the marginal revenue curve lies below the demand curve, and marginal revenue is less than marginal cost at the competitive output Q_c. Thus, it pays for the cartel to reduce output from the competitive level—but by how much? It should lower output until its marginal revenue equals marginal cost, which guarantees that profits are maximized. The cartel increases its profits by lowering the aggregate cartel output to Q_m where *MR* equals *MC* (Figure 5.1b). The price rises to p_m. Because the cartel is made up of n identical firms, it requires each firm to reduce its output to $q_m = Q_m/n$. In this example, the identical firms share in the extra profits equally.

Why doesn't each competitive firm reduce its own output below the competitive level? At the competitive equilibrium, each competitive firm sets its marginal revenue

[1] As with a monopoly, the cartel can restrict output and let the demand curve determine price or raise price and let the demand curve determine output. The two approaches are equivalent.

equal to its marginal cost and has no incentive to further lower its output.[2] If it were to reduce its output by one unit, it would lose profits because the marginal revenue on the last unit produced (the price) would exceed its marginal cost. Thus, each competitive firm is maximizing its profits at the competitive output.

The gain in collective activity comes from the very slight slope to the competitive firm's demand curve. Although economists often say that each competitive firm acts as though it faces a horizontal demand curve—that it cannot raise its price by lowering its output—that is not absolutely correct. The demand curve does have a slight slope: A competitive firm that stops producing might raise the market price by a small amount.[3] That small slope can be ignored when talking about a single firm, but it cannot properly be ignored when talking about all firms collectively.

If all firms cut back by, say, 10 percent, the market price definitely rises; however, if only one firm cuts back by 10 percent, the effect on price is so small that it is hardly measurable. Each competitive firm decides that it doesn't pay to reduce its output significantly because its gain is less than its cost. If it reduces its output by one unit, its gain is the trivial amount by which price rises times the units it produces, whereas its loss is the price it would have received for this last unit.

A competitive firm ignores the good it does other firms by reducing its output and increasing the market price; it places no value on the gains of other firms. This gain by others is an *externality*.[4] Working cooperatively, the cartel members gain from the output reductions of each firm. When all firms belong to the cartel, all the gains from reducing output and raising price go to the cartel, which divides the gains among its members. Here, the externality created by each firm in reducing its output has been internalized by the cartel. As a result, it pays the cartel to reduce total output below the competitive level, even though it would not pay any competitive firm to reduce its output individually.

Creating and Enforcing the Cartel

> Socrates: *[Tell] me whether you think that a city, or an army, or a band of robbers or thieves, or any other company which pursue some unjust end in common, would be able to effect anything if they were unjust to one another?*
> Thrasymachus: *Of course not. . . .*
> Socrates: *[When] we say that any vigorous joint action is the work of unjust men, our language is not altogether accurate. If they had been thoroughly unjust, they*

[2] Because a price-taking competitive firm faces a horizontal demand curve, its marginal revenue curve is also horizontal and identical to the demand curve. Thus, the competitive firm's *MR* curve is horizontal at the competitive price p_c (where the market supply or *MC* curve hits the market demand curve in Figure 5.1b).

[3] See the discussion in Chapter 3 of the elasticity of demand facing a single competitive firm and Equation 3.1.

[4] An externality is a good (or bad) that is not priced by the market.

> *could not have kept their hands off one another. Clearly they must have possessed justice of a sort, enough to keep them from exercising their injustice on each other at the same time as on their victims. For the thorough villains who are perfectly unjust, are also perfectly incapable of action.*[5]

Would you join a cartel if all the other firms in your market were forming one? Such behavior is usually illegal in the United States and many other capitalist countries, so, no doubt, you would refuse on moral and legal grounds. Suppose it was not a moral person like you who was being asked this question—suppose it was your slightly shady cousin. Would your cousin join an illegal cartel conspiracy?

Well, it depends. Your cousin's first thought is likely to be: "What's in it for me?" It should be obvious to him that it is in his best interest to let all the other firms in the market form a cartel that does not include his firm. Then the cartel would restrict output, driving up the price, while his firm could produce as much as it wanted. Of course, every other firm in the market makes the same calculation. Now suppose the other firms tell him that, unless his firm agrees, none of the others will join the cartel and restrict output. Your cousin now realizes that he can't have his cartel and produce as much output as he wants too. He can only obtain the higher price if his firm agrees to a reduction in output.

Your cousin then thinks: "What do I have to lose? If the cartel is caught by the government and convicted, my firm will have to pay a fine. But if the chance of being caught is small or the fine is low, it may be worth it to me." That is, if the expected loss from such a fine is low enough, your cousin joins the cartel.

But your cousin is always looking for an edge. Once he's joined the cartel, he says to himself: "Why shouldn't I cheat and produce more output than the cartel's agreement permits? After all, the cartel probably won't know who's producing the extra output." Of course, if all firms in the cartel think this way, the cartel will fall apart. The success of the cartel, then, turns on its ability to enforce its agreement.

Figure 5.1 illustrates why a firm has an incentive to cheat on the cartel's agreement. As explained above, the cartel members agree to restrict output to Q_m, which drives the price to p_m, the monopoly price. Figure 5.1a shows the cost curves of your cousin's firm, which is one of n identical firms in the market (and in the cartel). The cartel wants your cousin to produce $q_m = Q_m/n$ output: the output corresponding to his firm's share of the cartel output. But at the cartel's price, p_m, your cousin's firm can maximize its profits by producing q^* units of output (where its marginal cost curve equals p_m). Thus, although it is in the cartel's best interest for every firm to restrict output, it is in your cousin's best interest for every firm except his own to restrict output.

Cartels have little effect on prices if members do not cooperate. For example, in Kuala Lumpur, representatives of four pepper-producing countries decided that they would set a minimum price for black pepper. Even though the pepper cartel (Brazil, India, Indonesia, and Malaysia) produces more than 95 percent of the world's pepper

[5] Based on Plato, 1957, 37–8. The names of the speakers have been added and some material has been dropped.

and could raise the price, it has never been able to do so because its members keep undercutting the cartel's minimum price.[6]

Factors That Facilitate the Formation of Cartels

Once a cartel forms, the firms must agree to fix price (or equivalently, reduce output) if it is to be successful.[7] Why are there successful cartels in some markets but not in others? Unfortunately, we know a great deal about cartels that get caught, but very little about those that escape detection. As a result, it is not known whether the cartels that find themselves in court are unsuccessful or merely unlucky. Some evidence suggests that cartels that end up in court are actually unprofitable and hence, perhaps, atypical (Asch and Seneca 1976). Other evidence (Suslow 1998) suggests that cartels tend to form in less profitable industries.

Many characteristics of markets and firms that contribute to successful price-fixing conspiracies have been identified using studies of cartels that have ended up in court (Stigler 1964a; Hay and Kelley 1974). These characteristics may be roughly divided into those that allow a cartel to raise the market price in the first place and those that prevent the cartel agreement from breaking apart due to cheating by members. The following sections describe some of the major factors that facilitate the formation of cartels. Factors that lead to their survival are discussed in the next section. See Example 5.1 for a description of one of the most important cartels in American history.

Large firms may decide independently to behave as though they had a cartel arrangement without a formal meeting; that is, each one can cut its output and hope that the others will do likewise. Inevitably, in oligopolies, firms take their rivals' actions into account (as discussed in more detail in Chapter 6). When firms in an oligopoly coordinate their actions despite the lack of an explicit cartel agreement, the resulting coordination is sometimes referred to as **tacit collusion** or conscious parallelism.[8] Stigler (1964a) explains that a theory of oligopoly could be based on cartel theory even in the absence of explicit agreements.

[6] "Chaos in the Cartel: Pepper Producers Pick a Purchasers' Price." *San Francisco Chronicle,* August 8, 1983:49.

[7] Generally, price fixing is discussed rather than output reductions because it is believed to be more common. Other firms' prices are sometimes easier to observe than their output levels. In a study of antitrust cases from 1890 through 1969, Posner (1970, especially p. 424) found that only 1.6 percent of the cases had only explicit production or sales quotas.

[8] The use of the terms *tacit collusion* and *conscious parallelism* has led to confusion and added fuel to legal disputes. One ambiguity in the terms arises because the oligopoly price lies between the competitive and monopoly prices. Although many economists use these terms to mean that the oligopoly price is at the monopoly or cartel level, others use them to refer to the case where the noncartel price is elevated above the competitive price. A similar ambiguity arises with the term *collusion*. Economists often use the word to refer either to nonconspiratorial behavior (lawful actions where firms do not engage in explicit price fixing) or to conspiratorial behavior (illegal, explicit price fixing), but lawyers commonly use the term to refer only to conspiratorial behavior. A separate ambiguity in the law exists around the meaning of the word *agreement* (see Carlton, Gertner, and Rosenfield 1997).

EXAMPLE 5.1 *An Electrifying Conspiracy*

On a steamy Saturday, May 9, 1959, at 2:30 PM, reporter Julian Granger of the Knoxville *News-Sentinel* sat at his desk reading routine handouts from local publicity sources. The usual weekly newsletter from the Tennessee Valley Authority (TVA) announced that several contracts had been awarded. Pretty dull stuff.

But then Granger read that one contract had been awarded to Westinghouse for transformers for $96,760. The newsletter went on to note that "Allis-Chalmers, General Electric, and Pennsylvania Transformer quoted identical prices of $112,712." How could three companies bid prices that were identical to the penny under a system of secret, sealed bids?

Later in the release, he read that two other companies had quoted identical prices on a $273,200 contract. On yet another contract for conductor cable, there were seven identical bids of $198,438.24. Identical down to the 24¢.

Granger's story, which appeared on page 25 of the second section of the Knoxville *News-Sentinel* on May 13, 1959, drew little reaction. The lead of his second story on May 17, 1959, stated, "TVA purchasing records revealed today that at least 47 large and small American manufacturers have taken part in identical bidding on a wide variety of items in the past three years." He noted that the TVA had no choice in many cases but to award the contract on a chance drawing from a hat.

Could these identical bids have happened by accident? Oh, come now. Granger showed that some identical bids quoted the same delivered prices even though distances from delivery points varied by hundreds or thousands of miles for equipment you couldn't exactly weigh on a bathroom scale.

General Electric and Westinghouse, the two biggest electrical equipment manufacturers, had equal bids more frequently than other firms. Between 1946 and 1957, they raised prices 10 times on switching gear in a parallel pattern: The announcements of these increases came within a few days of each other.

After the second story, Granger went to local suppliers and got nowhere—they seemed afraid to talk. Granger did learn, however, that the Knoxville Utility Board had received a long series of identical bids from the electrical equipment manufacturers: up to 11 at one time. Their purchasing agent, Karl Strange, told Granger that he had noticed an increase in the practice since the end of World War II.

The Scripps-Howard papers reprinted the first two articles nationwide. On May 19, Senator Estes Kefauver inserted Granger's second story verbatim in the *Congressional Record* of the hearings before the Subcommittee on Antitrust and Monopoly.

Eight days before Granger's first story broke, Ralph J. Cordiner, Chairman of the Board of General Electric (GE), testified before the subcommittee on a bill to promote more vigorous competition in a variety of industries. He asked, "Is it assumed that companies in industries affected by this bill have the ability to 'administer' prices in a manner not responsive to market supply and demand?" He responded to his own question: "If so, the assumption is false, because these companies are just as much subject to competitive market conditions as any others." He continued, "In all instances the prices are completely subject to the force of competition in the marketplace and the value the customer believes he is receiving." He concluded that the then-current antitrust laws were "well enforced."

Exactly six months before Cordiner testified so sanctimoniously about competition, seven of his top executives met with their competitors in the Hotel Traymore in Atlantic City, New Jersey, to jack up the price of power switchgear assemblies ($125 million in annual sales). Until that time, the firms had an agreement that their sealed government bids would be divided by each of the conspiring companies in the following ratios: General Electric, 42 percent of the market; Westinghouse, 38 percent; Allis-Chalmers, 11 percent; and I-T-E, 9 percent.

Apparently, the initial conspirators made room for a major new entrant at this meeting. Federal Pacific, the newcomer to the cartel, was given permission to quote prices slightly lower than the others for a brief period in order to establish its assigned share of the market. GE and Westinghouse agreed to lower their shares to 39 percent and 35 percent, respectively, giving Federal Pacific 7 percent of the market. Over the next 12 months, at least 35 such meetings were held, with GE playing a prominent and active leadership role.

Senator Kefauver, the chairman of the Subcommittee on Antitrust and Monopoly, moved the hearings to Knoxville in September 1959. The hearings demonstrated that the prices of heavy industrial electrical equipment had increased 50 percent since 1951.

At the hearings, many examples of identical bids were presented. Even when the bids were not identical, the companies followed a rotation pattern, where one bid was low, one high, and the other two identical. The next time around, the order of the firms might change, but there would be one low, one high, and two equal bids. Presumably, the firms took turns winning the bidding in the proportions agreed to earlier. Apparently they agreed to alternate winning the bidding according to a *phase of the moon* formula: The firm to give the low bid was determined by the fullness of the moon.

The Philadelphia Antitrust Office spent 18 months tracking down evidence used to indict 29 manufacturers (practically the entire heavy electric industry) and 44 of their top executives. Attorney General William P. Rogers issued the first official announcement of the indictments on February 16, 1960.

According to the first indictments returned by a federal grand jury sitting in Philadelphia, Westinghouse, Allis-Chalmers, Federal Pacific Electric Company, GE, I-T-E, and many of these firms' top executives had engaged in a conspiracy at least since 1956. The defendants were accused of fixing and maintaining high prices, allocating the business among themselves, submitting noncompetitive, rigged bids, refusing to sell equipment to other manufacturers of electrical equipment, or raising prices to them so that they could not effectively compete.

Conspiracy in this industry was facilitated by the relatively small number of firms and the large market shares of the largest firms. Electrical manufacturing had a four-firm concentration ratio (sum of sales by the four largest firms divided by total industry sales) of over 50 percent compared to about 25 percent in all manufacturing. The concentration ratios were above 75 percent in all specific product areas involved and over 95 percent for turbogenerators, power transformers, power switchgear assem-

Continued

blies, distribution transformers, low-voltage power circuit breakers, isolated phase buses, bushings, and lightning arresters.

Eventually, 45 executives and 29 corporations were indicted. Most of them made no defense in the face of this overwhelming evidence. Generally, vice presidents and division managers took the fall. The seven men with the highest positions received jail sentences, but those were typically only on the order of 30 days. In addition, 24 people received suspended sentences. Total fines to firms reached nearly $2 million, while individual fines were $137,500.

In addition to the government suits, nearly 2,000 private suits were filed. General Electric settled its lawsuits for over $200 million (including $6.74 million to the TVA and $1 million to other federal agencies), and Westinghouse for more than $100 million. Total damages paid by all companies topped $400 million.

Many articles and books presented this story as a triumph of the system over evil conspirators, but this conclusion is hard to understand. Many, if not most, of the top directors of these firms were not personally indicted or punished, and the total fines were a small fraction of the monopoly profits the cartel earned over its life span. (Sultan (1974, 1975) argues that the conspiracy did not significantly raise prices; however, Bane (1973) and Lean, Ogur, and Rogers (1982) find that prices did rise.) According to a U.S. Congress report, this long-lasting conspiracy may have raised prices by nearly 10 percent. Other estimates have been over twice that for specific products. Electrical manufacturing accounts for about one-twelfth of total manufacturing and about 3 percent of all economic activity. About 30 percent of this manufacturing was electrical apparatus, which had $5 billion in shipments in 1958. The indictments referred to only about $1.75 billion of these annual sales (about 10 percent of all electrical manufacturing sales). Even assuming that prices were only 10 percent too high on only $1.75 billion worth of annual sales, purchasers paid roughly $175 million too much during each year of the conspiracy, which apparently lasted for decades. From the viewpoint of the firms involved, this experiment in cartel behavior probably looked like a great success, even after they were caught and punished.

The threat of penalties for illegal price fixing apparently was not viewed as a sufficient deterrent by GE and Westinghouse. These same firms have been charged and punished repeatedly since the Sherman Antitrust Act first went into effect in 1890. There were 13 U.S. Department of Justice antitrust cases and 3 Federal Trade Commission cases against GE and Westinghouse between 1911 and 1952. The government "won" all of these cases, obtaining convictions, *nolo contendere* pleas, or consent decrees in each (Walton and Cleveland 1964, 16–20).

Presumably, the conspirators should have learned to be more careful as a byproduct of the publicity, if not the fines, in the 1960s cases. However, GE and Westinghouse were accused of conspiring to fix prices on turbogenerators, starting with a new pricing policy GE announced in May 1963, just two and one-half years after they were found guilty in this bid-rigging case.

Sources: Fuller 1962; Walton and Cleveland 1964; and U.S. Congress, Joint Committee on Internal Revenue Taxation, *Staff Study of Income Tax Treatment of Treble Damage Payments under the Antitrust Laws* (Washington, DC: Government Printing Office, 1965), 39 (cited by Posner 1975).

Three major factors are necessary to establish a cartel. First, a cartel must be able to raise price above the noncartel level without inducing substantial increased competition from nonmember firms. Second, the expected punishment for forming a cartel must be low relative to the expected gains. Third, the cost of establishing and enforcing an agreement must be low relative to the expected gains.

The Ability to Raise the Market Price. Only if a cartel is expected to raise the price above the noncartel price and keep it high do firms join.[9] The more inelastic the demand curve facing a cartel, the higher the price the cartel can set and the greater its profits. If the cartel's demand curve is inelastic (relatively vertical at the current price), raising price can significantly raise revenues (that is, quantity demanded falls by a smaller percentage than price rises) and profits. In contrast, if a potential cartel faces an elastic demand curve (relatively horizontal), raising price causes revenues to fall (because quantity would fall by more than price increases, and profits may rise only slightly). See Example 5.2.

Entry by nonmember firms or *close substitutes produced in other industries* prevents a cartel from raising price. If the cartel controls only a small share of the relevant market, which includes all close substitutes, firms not in the cartel undercut the cartel and prevent it from raising the market price; that is, the demand curve facing the cartel is relatively elastic. Even if all firms initially in a market form a cartel and raise the price, the higher price may induce enough new firms to enter that the cartel is unable to keep the price high in the long run. That is, the long-run elasticity of demand facing the cartel is very high (especially relative to the short-run elasticity). Obviously, the longer the cartel can expect to keep the price high, the greater the current value of creating a cartel.

Low Expectation of Severe Punishment. Cartels only form if members do not expect the government to catch and severely punish them. Large expected penalties reduce the expected value of forming a cartel in the first place. Before they were made illegal in the United States in 1890, explicit cartels were much more common. During periods when the Department of Justice has been relatively lax in enforcing the laws, price-fixing conspiracies have been more prevalent (Posner 1970). Internationally, where cartels are legal, they have been more common than in the United States.[10] Some governments have created cartels (as discussed below).

As long as coordinating firms did not use unlawful acts of violence, intimidation, or fraud, British courts did not stop price fixing in modern times until 1956.[11] A survey

[9] If the noncartel price is close to the cartel price, then firms may not believe that joining the cartel is profitable given the legal liability they potentially face from belonging to a cartel.

[10] See the extensive discussion of the Organization of Petroleum Exporting Countries (OPEC) cartel at our web site at **www.aw-bc.com/carlton_perloff** "OPEC."

[11] The Restrictive Trade Practices Act (passed by Parliament in 1956) required that all contracts or agreements among suppliers in restraint of trade be reported to the Registrar of Restrictive Practices. This law has been modified substantially since then. In 1973, the Office of Fair Trading took over this responsibility and agreements among service industries had to be reported as well. This agency was empowered to challenge agreements that were contrary to the public interest. A special Restrictive Practices Court decides whether such agreements are prohibited. In contrast to U.S. law, this court can accept the argument that benefits outweigh damages and allow price fixing. A new Competition Act was passed in 1980 to facilitate investigations by the Office of Fair Trading.

EXAMPLE 5.2 *The Viability of Commodity Cartels*

Attempts have been made to cartelize the market for many of the major internationally traded commodities. Most of these initiatives have failed, however, as the cartels fell apart quickly or were unable to raise prices substantially.

Eckbo (1976) studied 51 formal international cartel organizations in 18 industries, with the earliest agreement in 1918 and the latest in 1964. He defined a cartel as successful if it raised the price at least three times the marginal production cost of the member with the highest cost. Only 19 cartels (37 percent) were successful by this criterion. One of them, the iodine cartel, lasted 61 years. The remaining successful cartels had formal agreements that lasted from 2 to 18 years, with a median lifetime of 5 years and a mean of 6.6 years. Only 5 of the 19 lasted 10 years or longer.

Of these successful cartels, 3 (out of 9 for which there is information) broke down for nonmarket reasons such as government intervention or war. Of those that collapsed for market-related reasons, 7 out of 16 (44 percent) had internal conflicts among cartel members, whereas 9 (56 percent) ended because of external forces such as competition from nonmembers (the usual case) or reactions by buyers.

Two factors that allow a cartel to persist and raise prices are that (1) it can detect and prevent cheating by members and (2) it faces a relatively inelastic residual demand curve at noncartel prices. The cartel's residual demand curve is likely to be inelastic if it has a relatively large share of the market, the market demand is not very elastic, and noncartel members have inelastic supply curves.

The longest-lived cartel in Eckbo's survey, iodine (1878–1939), made all sales through a central cartel office in London, which prevented members from cheating. Maintaining a cartel is not sufficient for success, however, if it cannot raise prices.

The Organization of Petroleum Exporting Countries (OPEC), the International Bauxite Association (IBA), and the Conseil Intergouvernemental des Pays Exportateurs de Cuivre (International Council of Copper Exporting Countries, or CIPEC) differ in their market power because of their different market shares and the residual demand elasticities they face. OPEC quadrupled the world oil price initially; IBA tripled the price of bauxite; but CIPEC has been unable to raise copper prices significantly.

When OPEC was formed, it had approximately two-thirds of the world's oil reserves and a similar fraction of the noncommunist world's oil production. By 1975, IBA accounted for 85 percent of total noncommunist world bauxite production. In contrast, CIPEC accounts for only about one-third of the noncommunist world's copper production. Of Eckbo's successful cartels, 15 out of 19 (79 percent) had four-firm concentration ratios over 50 percent. In 14 of them (74 percent), the cartels' share of total production exceeded 75 percent.

Of the 9 successful cartels about which we have enough information, 7 faced inelastic demand curves (elasticities less than 1 in absolute value). In 8 of the 9 cases, no short-term substitutes for the commodity were available outside the cartel, although for 7 cartels there were long-term substitutes—which may be why they eventually ended.

Pindyck (1977, 1979) shows that dynamic, long-run adjustments in commodity markets are also important. OPEC faces a relatively inelastic fringe supply. Despite major price increases, non-OPEC petroleum producers have not substantially increased their supply in the short to medium run. Similarly, the world demand for bauxite is extremely inelastic (up to a limit price), even in the long run.

In contrast, in the short run and even more so in the long run, secondary copper, which is produced from scrap, is very responsive to price. As a result, CIPEC faces a much larger long-run elasticity than short-run elasticity. If CIPEC were to raise its price very much, others would increase their production from scrap.

Given these differences, it is little wonder that OPEC and IBA could raise prices while CIPEC could not. These factors may also explain why still other natural resources have not been successfully cartelized. Pindyck holds that other minerals, such as iron ore, manganese ore, lead, tin, zinc, and nickel, would also face high long-run residual demand elasticities due to secondary supplies from scrap. Recently, IBA has suffered setbacks because Brazil and other producers have not restricted output. It continues as a research group.

Indonesia and Grenada produce 98 percent of the world's nutmeg and agreed to form a nutmeg cartel in 1987. For the 15 months before the formal agreement, however, Grenada operated informally under an Indonesian guideline. The two Countries claimed they did not intend to force prices higher but merely wanted to ensure that there is "no price cutting"—presumably from the informal cartel level. They were not worried about the impact of the cartel on demand. Nutmeg has no close substitutes; it has a distinctive taste, so bakers are unlikely to change their recipes appreciably.

One long successful cartel is the De Beers diamond cartel, which, throughout the twentieth century, has been the largest-selling agent of most of the world's diamonds. Even the Soviet Union, which was the second-largest diamond exporter after South Africa, sold all its diamonds through the De Beers cartel for the last quarter-century. The breakup of the Soviet Union threatened the stability of De Beers, though Russia apparently has returned to the cartel.

Traditionally, as new mines were developed, De Beers gave them sufficient market share so that they agreed to sell through De Beers and accept its production control system. When Tanzania decided to act independently, De Beers depressed the price for the quality of stones sold by Tanzania, forcing it to rejoin the syndicate.

Sources: Fisher, Cootner, and Bailey (1972); Eckbo (1976); Pindyck (1977, 1979); Fisher (1981); Alan J. Wax, "Spicy New Cartel Sets Nutmeg Prices." *San Francisco Chronicle,* May 25, 1987:20; Clyde H. Farnsworth, "OPEC Isn't the Only Cartel That Couldn't," *New York Times,* April 24, 1988: 3; "Diamonds: Friends Again," *The Economist,* March 2, 1996, 338:59–60.

of industrial trade associations carried out by the Political and Economic Planning agency in 1953–56 found that 243 of the 1,300 associations (19 percent) attempted to fix prices (Phillips 1972).

Low Organizational Costs. Even if a potential cartel could raise prices in the long run and not be discovered, it will not form if the cost of initial organization is too high. The more complex the negotiations, the greater the cost of creating a cartel. Four factors keep the cost low, facilitating the creation of a cartel: Few firms are involved, the market is highly concentrated, all firms produce a nearly identical product, and a trade association exists.

Setting up a secret meeting without the government's knowledge is relatively easy when there are few firms involved. Even if there are many firms in a market, the largest firms may meet and establish a cartel (dominant firm) that does not explicitly include the smaller fringe firms. Of the 606 Department of Justice price-fixing cases (1910–72) examined by Fraas and Greer (1977), the average number of firms involved in each case was 16.7, whereas the median was 8, and the mode was 4.[12] That is, a few cases involving a large number of firms raised the average, but the most common type of case involved 4 firms, and half the cases involved 8 or fewer firms.[13]

Of the Department of Justice price-fixing cases (January 1963–December 1972) studied by Hay and Kelley (1974), only 6.5% involved 50 or more conspirators.[14] The average number of firms in the remaining cases was 7.25. Although only 26% of the cases involved 4 or fewer firms, nearly half (48%) involved 6 or fewer firms, and 79% involved 10 or fewer firms.

Of the global cartels from 1990 to 2003 studied by Connor (2003), the median number of corporate participants was five. More than half (77%) of these cartels had six or fewer firms. Only 13% had 10 or more participants, but most of those were organized by quasi-official European trade associations.

Even where cartels are legal, as are many international cartels not involving U.S. firms, the number of firms is crucial. For example, a long period (1928–72) of successful cartelization by two countries of the world mercury market was followed by years

[12] If the number of firms involved in each case is arranged in ascending order, then the middle number is the median number of firms. If there were 5 cases with 2, 4, 5, 8, and 9 firms involved, the median number of firms would be 5. Imagine a graph that plots the cases so that the horizontal axis shows the number of firms involved and the vertical axis shows the number of cases involving a given number of conspiring firms. The mode is the highest point on the plot. The mode is the most common number of conspirators.

[13] The median number of firms involved varied by industry. In natural resources markets, the median number of firms was 13. The corresponding numbers were 7 in manufacturing, 11 in distribution, 15 in construction, 4 in financial institutions, 4 in transportation, and 8 in services.

[14] Hay and Kelley (1974) studied horizontal price-fixing conspiracies that were prosecuted by the U.S. Department of Justice Antitrust Division. Their study excluded price fixing by various professional groups (because they were not covert), but included virtually all other cases that were filed and won in trial or settled by *nolo contendere* ("no contest") pleas. Pleading *nolo contendere* is equivalent to pleading guilty for the purposes of sentencing but is not an admission of guilt by the defendant. When such a plea is accepted by the court, a trial is not necessary. Occasionally, courts accept such pleas over the objection of the Department of Justice.

of unsuccessful attempts at price fixing by a larger group of countries (MacKie-Mason and Pindyck 1986).

If a few large firms make most of the sales in a market, and if they coordinate their activities, they can raise price without involving all the other (smaller) firms in the market. For example, Spain and Italy, which controlled 80% of the world's production of mercury, formed a successful cartel that did not formally involve five other producers (MacKie-Mason and Pindyck 1986).

Empirical evidence supports the view that cartels are more likely in concentrated industries.[15] In 42% of the Department of Justice price-fixing cases studied by Hay and Kelley (1974), the four-firm concentration ratio (the sum of the market shares of the four biggest firms) was over 75%; in another 34% of the cases, the ratios were between 51 and 75%. Thus, in 76% of the cases, the concentration ratio was greater than 50%. Only 6% of the cases had concentration ratios less than 25%. The overall average was 67.7%.[16] Of the global cartels studied by Connor (2003), cartel members usually controlled over 90% of the market's sales. Moreover, when entry caused the cartel's share to drop below 65%, cartel activity typically ceased.

Similarly, the existing evidence shows that cartels are often found in smaller geographic areas. In the U.S. Justice Department price-fixing and other antitrust cases from the passage of the Sherman Act (1890) through 1969 studied by Posner (1970), nearly half (47.4%) the conspiracies were in local or regional markets, 37.6% were nationwide, and 8.7% involved foreign trade. The smaller the geographical area of a market, the more likely it is that a few firms have a large share of the business.

Firms have more difficulty agreeing on relative prices when each firm's product has different qualities or properties. Each time a product is modified, a new relative price must be established. It is easier for a cartel to spot cheating when all it has to examine is a single price. It is relatively difficult to detect price cutting that is achieved by an increase in quality; a firm could increase its quality and hold its price constant if it wanted to increase sales without explicitly violating the pricing agreement.

In virtually all the price-fixing cases studied by Hay and Kelley (1974), the product was relatively homogeneous across firms. In the few exceptions, complicated products or services were allocated on a job-by-job basis that facilitated coordination, or a single issue was isolated for the agreement. For example, a group of swimsuit manufacturers agreed to delay end-of-season discounts. Similarly, virtually all the recent global conspiracies (Connor 2003) involved homogeneous products.

[15] Scott (1991a) shows that multimarket contact can be important. Large conglomerates may be potential rivals in a number of markets simultaneously. They can communicate about all these markets at once. It is also potentially more costly to deviate from a cartel agreement in one because it risks destroying all the cartel agreements. To the degree that multimarket contacts are important, the degree of concentration in a single market may be misleading.

[16] To minimize the systematic bias from excluding cases for which the concentration measure could not be determined directly, if the number of firms was known, Hay and Kelley (1974) calculated minimum concentration ratios by assuming each firm had an equal share.

Trade associations, by lowering the costs of meeting and coordinating activities among firms in a market, facilitate the establishment and enforcement of cartels. Most industries have trade associations that meet regularly. Not all industries with trade associations necessarily form cartels. However, as Adam Smith observed, such meetings are conducive to price-fixing agreements, and trade associations are often the mechanism by which large groups coordinate activities. In the Hay and Kelley (1974) study of Department of Justice price-fixing cases, trade associations were involved in 7 out of 8 cases in which more than 15 firms conspired, and in all cases involving more than 25 firms. Overall, 29 percent of the cases involved trade associations. Fraas and Greer (1977) found that 36 percent of all price-fixing cases involved trade associations. Moreover, the median number of firms involved was 16 when there was a trade association, compared to 8 for all cases. Posner (1970) found that 43.6 percent of all antitrust cases involved trade associations.

Enforcing a Cartel Agreement

Even if a market consists of a small number of firms producing a homogeneous good with no close substitutes, has an inelastic demand curve, and faces no threat of entry, a cartel cannot succeed if members can and want to cheat on the agreement. Some of the factors that lead to the formation of a cartel also help it to detect cheating and enforce its agreement.

Detecting Cheating. Cartel agreements are easier to enforce if detecting violations is easy. Four factors aid in the detection of cheating:

- There are few firms in the market.
- Prices do not fluctuate independently.
- Prices are widely known.
- All cartel members sell identical products at the same point in the distribution chain.

With relatively few firms, the cartel may more easily monitor each one, and increases in one firm's share of the market (an indication of price cutting) are easier to detect. Further, moral (or immoral) suasion may be easier when there are only a few conspirators (See **www.aw-bc.com/carlton_perloff** "Broker").

Hay and Kelley (1974) found that most of the price-fixing conspiracies lasting 10 or more years were in markets in which there were few firms and the largest firms made most of the sales. When a large number of firms was involved, conspiracies were generally discovered very quickly, especially because details about some of the large-group organizational meetings often were printed in local newspapers. In contrast, Posner (1970) found that, of the detected cartels, large ones lasted as long as smaller ones. He found that 52 percent of conspiracies involving 10 or fewer firms had lasted for 6 or more years, whereas 64 percent of larger conspiracies persisted that long. Presumably, the more firms involved in a conspiracy, the more likely is discovery by the government. In general, conspiracies are uncovered through information provided by private parties rather than by Department of Justice investigations.[17]

If a market has frequent shifts in demand, input costs, or other factors, prices in that market have to adjust often. In that case, cheating on a cartel arrangement may be difficult to detect, because it cannot be distinguished easily from other factors that cause price fluctuations. Cheating is easier to detect if prices are known. Some cartels have arranged for firms to inspect each other's books. In Posner's (1970) study of antitrust cases, at least 6.2 percent of the cases involved exchange of information, whereas 4.3 percent involved policing, fines, and audits. Of course, books can be faked, so such inspections cannot prevent all violations of the cartel agreement.

In some cases, governments help. For example, they often report the outcome of bidding on government contracts, so that cheating is instantly observable by the cartel (see Example 5.3). A quarter of the cases Hay and Kelley (1974) examined involved some form of bid rigging.[18]

Example 5.1 describes a *phases of the moon* scheme used by manufacturers of electrical products to rotate the winning of sealed bids. No firm could hope to win out of turn because its treachery against the cartel would be instantly exposed when the government announced the winner.

Vincent (The Fish) Cafaro, a former member of the Genovese organized crime family, told senators the mob rigged bids in New York City, controlling the concrete industry and construction unions.[19] He said the contractors and unions that won construction jobs through bid rigging were required to kick back 2 percent to the "2 Percent Club," an organization run by the Genovese, Gambino, Lucchese, and Colombo families of New York City. He estimated that at least 50 percent of the highrise construction in New York had a mob connection and added, "Legitimate guys ain't got a chance" to win contracts for those buildings. According to Mr. Cafaro, the 2 Percent Club split up all of the jobs worth over $2 million. Contracts worth over $5 million went to mob-run companies.

[17] Of the cases studied by Hay and Kelley (1974), detection was due to grand jury investigation of another case in 24%; to complaint by a competitor in 20%; to complaint by a customer in 14%; to complaint by local, state, or federal agencies in 12%; and to complaint by current or former employees in 6%. Each of the following methods was responsible for detection in 4% of the cases: complaint by a trade association official, investigation of conduct or of performance by the Antitrust Division, report of a newspaper, and referral to the Antitrust Division by the Federal Trade Commission. Each of the following methods was responsible for 2% of the cases: complaint by an anonymous informant, merger investigation, and private suit.

[18] Hay and Kelley (1974) found some cases in which sales to government agencies were explicitly excluded from the agreement. Apparently cartel members believed that price fixing was more likely to be detected and prosecuted if directed against the federal government. In other cases, some market segments were excluded from agreements in order to reduce potential friction among cartel members. Fraas and Greer (1977) found that 19 percent of all cases over a longer period involved bid rigging. Posner (1970) determined that 7.4 percent of all cases involved sales to the government, and 6.7 percent involved other bidding cases.

[19] "Witness Says Mob Is into Highrises," *San Francisco Chronicle,* April 30, 1988:A7.

EXAMPLE 5.3 *Concrete Example of Government-Aided Collusion*

Members of a cartel have an incentive to undercut an agreed collusive price. If cartel members cannot detect such "cheating," the cartel is likely to fail. If a government agency publishes the prices each firm charges, it may facilitate the maintenance of a cartel, as appears to have been the case recently in Denmark.

Stigler (1964a) explained how the same factors that lead to a successful cartel also can result in successful tacit collusion. Many of the conditions necessary for successful tacit collusion in the ready-mixed concrete market were already in place prior to the government actions. The ready-mixed concrete market in Denmark consists of a relatively small number of firms. The two largest firms have plants throughout the country and compete with a number of smaller firms. Because ready-mixed concrete can be kept in a mixer truck for only about two hours after the mixing, it is typically shipped no more than 20 miles from the production site. As a consequence, relatively few firms compete in a given area: Usually fewer than 5 of the 115 plants in Denmark can serve a particular customer. Although the national four-firm concentration ratio is 57%, many local markets are monopolistic and the average market share of the largest firm in an area is 70%.

In 1993, the Danish antitrust authority started gathering and publishing firm-specific transaction prices for two grades of ready-mixed concrete. The government hoped that by providing buyers with price information, seller competition would be stimulated and price would fall.

According to Albæk, Møllgaard, and Overgaard (1997), the government's actions led to successful tacit collusive behavior: Prices rose and the variance in prices across firms was nearly eliminated. The average prices of the reported grades increased by 15–20% within less than a year (compared to an annual inflation rate of no more than 2%). No dramatic changes in costs or other factors explain this large increase (indeed wages, a major component of costs, fell during this period). Moreover, the variance in firms' prices dropped from about 30% around the average to only about 2–4%. Thus, it appears that the firms were able to set a higher price as a consequence of the government's information program.

When smaller contractors complained about the arrangement, they were given the right to split all jobs worth over $3 million. He said the Genovese family was "a very disciplined organization" with strict rules and capital punishment for serious violations.

Public availability of information can greatly simplify cartel enforcement. Publicly announcing price increases and decreases well in advance is one method of making price information available to all interested parties. An extreme, special case of sharing information occurs when a single sales agent or pool is used by all firms for all their sales, as was the case in 3 percent of the cases Fraas and Greer (1977) examined and in 6 percent of the cases studied by Posner (1970). Sales agents are commonly used in European cartels.

If some firms are *vertically integrated* (the same firm produces inputs, manufactures the product, and sells at the retail level), it may be difficult for the cartel to determine at what point in the distribution chain cheating occurs. In contrast, if all firms sell to the same type of customer (for example, at the retail level), cheating is easier to detect.

Cartels with Little Incentive to Cheat. A cartel may find enforcement easy under certain circumstances. Members have no incentive to cheat on the cartel agreement if their marginal cost curves are relatively inelastic, their fixed costs are low relative to total costs, their customers place small, frequent orders, or they have a single sales agent.

If a firm's marginal cost curve is nearly vertical, it has little to gain by cheating on the cartel agreement because it costs too much to substantially increase its output. In Figure 5.1a, if the marginal cost curve were nearly vertical, q^* would be close to q_c. Marginal cost curves are likely to be nearly vertical if firms are operating near their full capacities. Indeed, cartels may force their marginal cost curves to be more vertical by signing union contracts that require double wages for overtime work or using similar techniques (Maloney, McCormick, and Tollison 1979).[20]

Suppose a firm incurs a large fixed cost to build a plant in which it can produce at a constant marginal cost at any output level up to the plant's capacity. Such a firm has substantial unutilized capacity when demand falls (such as during a recession). It has an incentive to lower its price below the cartel level to stimulate its sales.

If there are many customers in a market who make small purchases, no firm has an incentive to lower prices below the cartel level. If it does so without announcing the price cut, other customers are unlikely to learn of the price cut; hence its sales will not rise. If the firm advertises its price reduction, the other cartel members will learn of the cut and retaliate. In contrast, when only a few customers place large, infrequent orders, a cartel has trouble detecting and preventing cheating.[21] Firms have an incentive to grant price reductions to large buyers to keep them as customers.

Legal cartels can try to prevent cheating by requiring that a single agent or organization sell output from all firms. For example, the iodine cartel, one of the longest-lived international cartels (61 years: 1878–1939), made all its sales through a central office in London (Eckbo 1976). See Example 5.4.

Methods of Preventing Cheating. Unless a cartel can detect violations of its price-fixing agreement and prevent reoccurrences, member firms engage in secret price cutting (or output expansions) that destroys the cartel. Although economists and lawyers understand a number of mechanisms that aid cartels in enforcing their agreements, the most successful cartel agreements and their enforcement mechanisms may be unknown. Here, we concentrate on six methods: fix more than just price, divide the market, fix market shares, use most-favored-nation clauses, use meeting-competition clauses, and establish trigger prices.

[20] So long as a cartel raises its marginal cost curve by more than its average cost curve, such actions increase profits (Salop, Scheffman, and Schwartz, 1984).

[21] Hay and Kelley (1974) argued that bid rigging and allocation of jobs among cartel members occur in industries in which orders are relatively large ("lumpy") compared to total sales.

EXAMPLE 5.4

Relieving the Headache of Running a Cartel

The major use for bromine in the late 1800s was as a headache remedy and sedative. The bromine industry was not concentrated and consisted of many small producers. Ordinarily, such an industry would not be a good candidate for a cartel because of the large number of producers.

In the early 1880s, the price of bromine fell by 40 percent. In 1885, the National Bromine Company ("bromine pool") was formed. It purchased the bromine produced by all manufacturers. The bromine pool then sold bromine to two independent distributors of chemicals, Powers and Weightman of Philadelphia and Malinckrodt Chemical Works of St. Louis, which were obligated collectively to buy the entire output of the pool. In turn, these two distributors sold to customers in their territories. The bromine pool required that manufacturers sell exclusively to it and that, if any manufacturer violated this provision, contracts with all other manufacturers could be terminated. Moreover, if a manufacturer entered the industry and did not contract with the bromine pool, contracts with other manufacturers could be terminated. The two distributors who were obligated by contract to purchase all the pool's output accumulated large inventories of bromine, which they threatened to dump on the market if any manufacturer failed to cooperate with the pool. Indeed, inventories were sold during price wars in 1886 and 1888 in order to punish competitors and to restore pricing discipline.

The distributors were much better able than individual manufacturers to monitor sales to customers and thus to detect whether any manufacturers were making secret sales. For their role in the cartel, the independent distributors took a significant share of the cartel profits. In 1892, this share was renegotiated to give a larger portion to manufacturers. (The original pool was replaced by W. R. Shields, which performed similar functions.)

The very successful bromine cartel lasted from 1885 to 1902. During its reign, the average price of bromine was about 25 percent higher than the average in the years before the cartel's formation.

There were only three periods of extended price wars over the cartel's roughly 20-year life span. The cartel ended because Dow Chemical Company developed a low-cost method of processing bromine in the 1890s. Dow initially signed contracts with the two bromine distributors that the bromine cartel had used as its exclusive distributors. But in 1902, Dow had grown so large that it decided not to rely exclusively on distributors to sell its products, and began selling directly to customers. The pool fell apart and the price of potassium bromide (the major bromine product) plunged 45 percent in two months. Undoubtedly, the demand by owners of small bromine-processing firms for potassium bromide as a headache remedy increased.

Source: Levenstein (1993).

To prevent cheating, successful cartels must do more than just set a price. Posner (1970, 400) finds that at least 14% of all Department of Justice antitrust cases involved explicit collusion on terms besides basic price (and this figure apparently does not include explicit rules on dividing the market, exchanging information, or sales quotas).[22]

Some cartels succeed in preventing cheating by assigning each firm certain buyers or geographic areas, which allows cheating to be detected easily. Fraas and Greer (1977) found that 26% of price-fixing cases involved market allocation schemes. Posner (1970) found that 7.8% of the antitrust cases involved an allocation of customers, 14.6% involved a division of territories, and 1.8% involved a division of product markets (or 24% overall). The two-country mercury cartel used a geographic division of markets: Spain supplied the United States, and Italy supplied Europe.

Another effective technique is for members of a cartel to agree to fix market shares (say, at their precartel levels). (See Example 5.5.) As long as market shares are easily observable, no firm has an incentive to cut its price. If it lowered its price, its share would increase, and other firms would retaliate. For example, cartel members who detect changes in the output levels of other firms could adjust their own output to maintain their proportionate shares of market output (Osborne 1976; Spence 1978a, 1978b). All firms expect this reaction, so no firm has an incentive to increase its own output only to earn lower profits after retaliation. As **www.aw-bc.com/carlton_perloff** “Conjectural Variations” discusses, fixing market shares can result in the cartel price.

A **most-favored-nation clause** in a sales contract guarantees the buyer that the seller is not selling at a lower price to another buyer (Salop 1986). A variant of such clauses was used in sales of large steam-turbine generators. The two major sellers, General Electric and Westinghouse (see Example 5.1), each used clauses in their contracts stating that the seller would not offer a lower price to any other buyer, current or future, without offering the same price decrease to the initial buyer. This rebate mechanism created a penalty for cheating on the cartel: If either company deviated from the agreement by cutting its price, it would have to cut prices to all previous buyers as well.

A **meeting-competition clause** in a long-term supply contract or in an advertisement guarantees the buyer that if another firm offers a lower price, the seller will match it or release the buyer from the contract (Salop 1986). Such a clause makes it difficult for a firm to cheat, because buyers will bring news of lower prices to the cartel. Thus, surprisingly, these clauses could be associated with high cartel prices rather than the low ones they seem to guarantee.

[22] Posner (2003, 51) notes, “The machinery [of cartelization] may include sales quotas, exclusive sales agencies, industry-wide price-fixing committees, the levying of penalties for infractions, provisions for the arbitration of disputes, the establishment of an investigative apparatus product standardization, allocation of customers, and the division of geographical markets.”

EXAMPLE 5.5 *Vitamins Cartel*

In the 1990s, there was a massive worldwide cartel involving many different vitamins, including biotin, folic acid, and vitamins A, B1, B2, B5, B6, C, and E, among others. Vitamins have a wide variety of uses as additives to human and animal diets and in skin and healthcare products. The various vitamins are not substitutes for each other.

Vitamin production is highly concentrated among a few firms. At the time of the cartel, the three largest producers were Hoffman-LaRoche (which since has sold its vitamins business), which produced 40 to 50 percent of all vitamins; BASF, with a 20 to 30 percent share; and Aventis (formerly Rhone-Poulenc), with a 5 to 15 percent share. These major manufacturers produced many of the same vitamins. Over half of all vitamin sales were for vitamins A and E, which were sold by all three major producers.

Allegedly beginning in 1989, Hoffman-LaRoche, BASF, and Rhone-Poulenc held meetings to discuss allocations of market sales around the world so as to reduce competition, and soon thereafter other firms became involved in a worldwide cartel.

The cartel fixed market shares for each vitamin by country, agreed on price increases, specified target prices and minimum prices, and shared information to ensure that each firm was abiding by its allocation. Sometimes the firms explicitly discussed large individual customers and agreed on those customers' prices and how much of the customers' needs each manufacturer would supply.

The firms met regularly. There were four levels of meetings: the highest level involving senior executives who determined overall strategy and adherence to the agreements; the next level involving marketing executives about two or three times a year; another meeting (usually quarterly) involving marketing managers of individual products to monitor the implementation of the allocations; and finally, quarterly meetings of regional marketing managers to discuss pricing, implement price increases, and adjust allocations. The "budget meetings" in August were used to outline allocations for the coming year, together with price increases.

All cartel members could agree that if the market price drops below a certain level (called a **trigger price**), each firm will expand its output to the precartel level (Friedman 1971); that is, all firms will abandon the cartel agreement. In this case, a firm that cuts its price might gain in the extremely short run, but would lose in the end due to the destruction of the cartel by this predetermined punishment mechanism.

One reason to use trigger prices is that, in some markets, firms have difficulty distinguishing between cheating by other firms and random fluctuations in price due to fluctuations in demand or supply costs. It is possible, however, for cartels to modify their punishment methods to prevent cheating even when random shocks occur (Green and Porter 1984). If firms were to permanently revert to competitive behavior whenever they detected a fall in price, the cartel could be destroyed by a random fluctuation in

The price was usually raised in increments of 5 percent, and the effective date of a price increase was often April 1 of the next year. Hoffman-LaRoche typically initiated the price increases, with the other firms following suit. The detailed exchanges of sales information allowed the firms to monitor adherence to their sales allocation. If a firm had sold too much in a given year, it could be required to buy supplies from the other firms to restore the original sales allocation.

Although the exact amounts by which the cartel raised prices are subject to dispute, the increases during the cartel period were sizable. For example, the average price for vitamin A rose by 40 percent and that of vitamin E increased over 60 percent from 1990 to 1998. The price of vitamin C rose by about 30 percent during the identified cartel period and fell by 50 percent thereafter.

Starting in the late 1990s, Rhone-Poulenc participated in the U.S. Justice Department's corporate leniency program. The first cartel member to confess to the DOJ—if it is not a ringleader or enforcer in the conspiracy and if the DOJ is not aware of the illegal activity—is granted automatic amnesty. Rhone-Poulenc revealed the existence of the cartel and details about its operations so as to avoid antitrust fines. Subsequently other cartel members agreed to pay multimillion-dollar fines. These same firms had to pay fines to Canadian and European Union competition authorities as well.

The fines were substantially larger than had been collected in previous antitrust cases. The largest U.S. fines were $500 million for Hoffman-LaRoche, $225 million for BASF, and $72 million for the Japanese company Takeda. Two Hoffman-LaRoche executives went to jail. The biggest Canadian fines were the $48 million (Canadian) collected from Hoffman-LaRoche, the $18 million from BASF, and the $14 million from Rhone-Poulenc. The major European fines were €462 million for Hoffman-LaRoche, about €300 million for BASF, and €37 million for Takeda. Private antitrust settlements collected additional amounts from the cartel members.

Source: Official Journal of the European Communities, Commission Decision of 21 November 2001. (Case Comp/E-1/37.512—Vitamins.)

price (rather than price cutting by one firm). Instead, if the firms agreed to behave competitively only for a predetermined length of time and then to revert to the cartel behavior, a random fluctuation in price would not destroy the cartel permanently.[23]

One attraction of this scheme is that, even if the agreement temporarily breaks down, it can be reestablished without further meetings. In a market in which random

[23] If the firms revert to their precartel output level, the price falls to the precartel level as well. A more severe punishment, a price below the precartel level, may be used instead: With a lower price, it may be possible to shorten the punishment period. For an illustration of how cartel prices might be set to minimize cheating, see Davidson and Martin (1985). Where members of a cartel disagree on how to behave, some kind of voting mechanism may be used (Cave and Salant 1987).

price fluctuations can mask cheating on the cartel agreement by firms, such an agreement could lead to recurrent sharp declines in price and cartel profit levels. When a random drop in price occurs, cartel members punish themselves unnecessarily.

Nonetheless, this mechanism may be attractive to the cartel because, if the punishment period (when all firms produce large levels of output) is long enough, it is never in a firm's best long-run interest to cheat on the cartel. Thus, cartel members realize that the price only falls below the trigger price because of random fluctuations (because no firm ever engages in price cutting). The cartel must keep punishing itself, however; if it stopped, price cutting would occur.[24] Example 5.6 provides an example of the American rail-freight industry in the 1880s that may illustrate such behavior.

Cartels and Price Wars

Many observers, seeing large price fluctuations in a market, argue that the firms in that market are trying to form a cartel that keeps breaking apart. They conclude that government intervention is not required because competitive forces keep destroying the cartel. Yet, these fluctuations could be part of a rational, long-run cartel policy involving trigger prices, as discussed in the preceding section. This trigger-price argument holds that price wars occur more often during unexpected business cycle downturns (recessions and depressions) when price is likely to decline in response to lowered demand (Green and Porter 1984; Staiger and Wolak 1992). We expect then that cartels are more likely to terminate during a price war. Other economists argue that price wars should occur in periods of high demand (Rotemberg and Saloner 1986). They reason that the benefit from undercutting the cartel price is greatest during booms.

To see whether either or both theories are realistic, Valerie Y. Suslow (1998) investigates the stability of cartels over the business cycle by examining 72 international cartel agreements covering 47 industries during the period 1920–39.

Because major European countries had no systematic antitrust legislation prior to World War II, these cartels were legal and had formal written contracts. As of 1927, cartels were legal in Switzerland, whereas Belgium, France, Spain, Italy, and the Netherlands did not explicitly prohibit them. Under German law, cartels were legal; however, Germany passed antitrust legislation in 1923 that was designed to guard against abuses of economic power. In 1930, Great Britain adopted a resolution recognizing cartels as a fact of economic life, but calling for the *principle of publicity,* which required compulsory notification, registration, and publication of the cartel agreements. Other European countries followed Great Britain's policy in the mid-1930s. It was not until after World War II that France passed legislation to control cartel activity.

[24] Bernheim and Ray (1989), Evans and Maskin (1989), Farrell and Maskin (1989), and others point out that instead of going into a punishment phase, cartel members may renegotiate their cartel agreement. These papers indicate, however, that it may be possible to form agreements that avoid this renegotiation problem.

EXAMPLE 5.6

How Consumers Were Railroaded

During the 1880s, a cartel of U.S. railroads openly operated as the Joint Executive Committee (JEC). Prior to the Sherman Act of 1890, no law prohibited such a cartel. As Porter explains, the JEC appears to have used a trigger-price strategy (Green and Porter 1984).

The JEC agreement allocated market shares rather than the absolute quantities shipped. Each railroad set its rates individually, and the JEC office reported weekly accounts, so that each railroad could see the total amount transported. Because total demand was quite variable, each firm's market share depended on both the prices charged by all firms and unpredictable market fluctuations.

Entry occurred twice between 1880 and 1886 [the period Porter (1983a) studies]. In each case, the cartel passively accepted the entrants, allocated them market shares, and thereby allowed the cartel agreement to persist.

On a number of occasions, however, when the cartel thought that cheating had occurred, it cut prices for a time, and then returned to the cartel price. Porter finds that noncooperative periods averaged about 10 weeks in duration and occurred in 1881, 1884, and 1885. The 1881 and 1884 incidents each occurred about 40 weeks after a new firm entered. He notes, however, that these price wars were not triggered by an unexpected tapering off of demand.

Porter also finds that price was 66 percent higher and quantity was 33 percent lower in co-operative periods. As a result, the cartel as a whole earned about 11 percent more revenues in cooperative periods.

Sources: MacAvoy (1965); Ulen (1980); and Porter (1983a). See also Ellison (1994).

It should have been easier for these cartels to survive than for illegal ones in the United States. German, French, or British firms were participants in roughly half the cartels, and U.S. firms were involved in one-third of them. In the 1940s, U.S. firms were indicted for their participation in 10 of these international cartels.

According to Suslow, the median cartel lasted slightly more than 5 years; 75 percent lasted more than 2 years, and 20 percent lasted more than 10 years. There was an industry pattern. Of the single-episode cartels, 40 percent involved chemicals, with only 6 percent in metals. In contrast, 46 percent of the multiple-episode cartels involved metals, with only 17 percent in chemicals.

In the 42 cartel episodes in which the number of firms is known, 83 percent had 10 or fewer firms, 64 percent had 5 or fewer, and 39 percent had 3 or fewer. Of the 74 percent of the 39 cartels for which there is market-share information, each had a world market share of over 50 percent. Thus, as with U.S. cartels, these international cartels involved relatively few firms with large collective market shares.

TABLE 5.1 Market Conditions Facilitating Global Price Fixing in Lysine, Citric Acid, and Vitamins A and E in the Early 1990s

Market Conditions	Lysine	Citric Acid	Synthetic Vitamins A & E
High seller concentration (CR4[a])			
Global market	> 95%	> 80%	> 95%
U.S. market	> 97%	= 90%	100%
Few cartel participants	4 or 5	4 or 5	3
High cartel supply control	95–99%	65–70%	95–100%
Low buyer concentration (CR4)	< 30%	< 40%	< 20%
Homogeneous product	Perfect	High	High
High barriers to market entry:			
Large plant scales	$150 million+	$150 million	Probably
Sunk investment costs	Yes	Yes	Yes
Technology secret	Yes	Yes	Yes
Slow building of new plants	3 years+	3 years+	3 years+
Buyers' observation of market prices	None	Some	Little
Annual market growth	10%, steady	8%, steady	2–3%, steady

[a] CR4 is the share of sales of the four largest firms in the industry.

Source: Connor (2003).

Suslow estimates the probability that a cartel will fall apart at a specific time, given that it survives until that time. Controlling for other factors, she found that cartels are more likely to fail during business-cycle downturns (recessions and depressions).[25] Moreover, cartels that were alive during periods of growth were less likely to end than others. In general, greater volatility in aggregate economic activity over the lifetime of the cartel (frequent upswings and downturns) increases the probability of cartel breakdowns.

We have discussed several factors that help a cartel to form and to prevent cheating. Many large, successful cartels possess these properties. Table 5.1 (from Connor 2003) shows the presence of these market conditions in the lysine, citric acid, and vitamins A and E industries in the early 1990s, when each had an operating cartel. Entry plays a particularly important role (see de Roos 1999 on lysine).

[25] Hajivassiliou (1989) reaches similar findings based on his study of the 1880–86 U.S. railroad cartel. Porter (1983a) and Lee and Porter (1984) examine this cartel's behavior under the maintained assumption of the Green and Porter model. Town (1991) rejects both the Green and Porter and the Rotemberg and Saloner hypotheses for this cartel and concludes that price wars were not related to demand fluctuations.

Consumers Gain as Cartels Fail

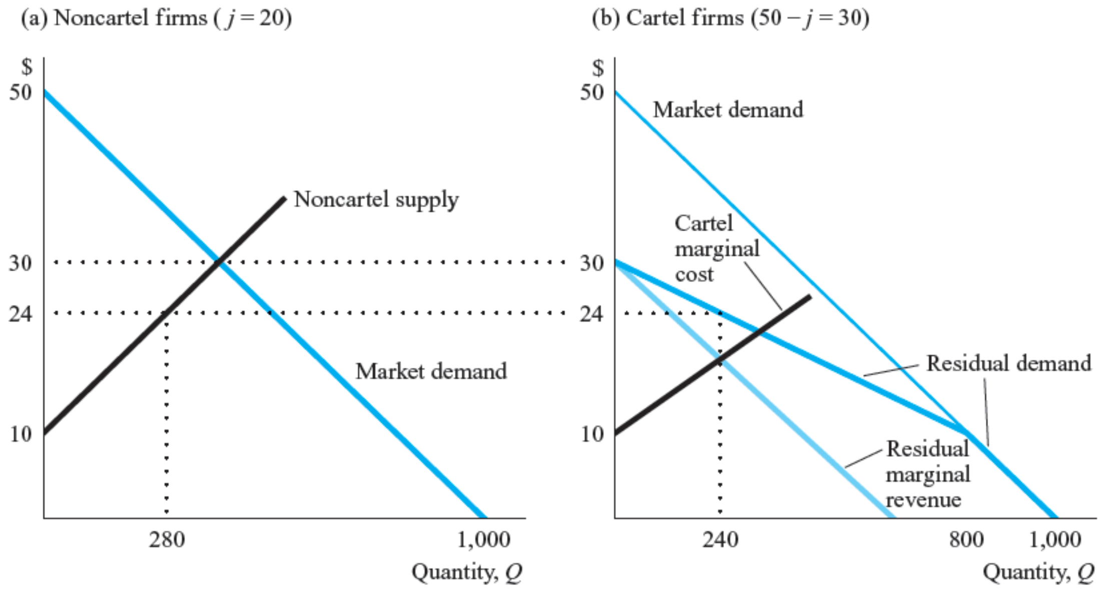
Firms that follow a cartel's rules look with disfavor on firms that produce more than the cartel says they should. Violators of the cartel's rules may be called "cheaters" or worse by the firms that obey them. Consumers, however, benefit from such noncartel behavior. The violators of the cartel agreement produce more than the cartel wants, which lowers the market price.

A numerical example illustrates the effects of noncompliance by some firms (see Appendix 5A for details). The market in this example includes 50 identical firms. We assume that no more firms can enter this market.

Of the 50 firms, j firms do not follow the cartel's agreement to restrict output; they sell as much as they want. These firms are price takers. *The cartel is a dominant firm facing a competitive fringe,* as we studied in Chapter 4.

The residual demand facing the cartel is obtained by subtracting the fringe supply from the market demand. Figure 5.2b shows the residual demand curve (thick dark blue line) that lies below the market demand curve (thin blue line) at prices above the competitive firms' shutdown level ($p = 10$).[26] The residual demand curve has a kink

FIGURE 5.2 Imperfect Cartel



[26] Because each competitive firm's supply curve is $q = 10 + p$, it would lose money if it produced positive levels of output at prices below 10.

TABLE 5.2 **Market Variables Under Various Degrees of Cartelization (50 Firms)**

	Number of Noncartel Firms	Price (p)	Market Elasticity	Market Output	Industry Profits (π)	Consumer Surplus (CS)	Welfare ($CS + \pi$)
Monopoly	0	33.33	−2.00	333	6,667	2,778	9,445
	1	32.41	−1.84	352	6,524	3,094	9,618
	10	26.97	−1.17	461	5,318	5,304	10,622
	20	24.00	−0.92	520	4,360	6,760	11,120
	30	22.44	−0.81	551	3,743	7,591	11,337
	40	21.67	−0.76	567	3,391	8,027	11,418
	49	21.431	−0.75	571	3,267	8,162	11,428
Competition	50	21.429	−0.75	571	3,265	8,163	11,429

CS = Consumer Surplus is the triangle with area $(1{,}000 - 20p)^2/40$.
Cartel's Market Share (as a percentage) = 100 times the cartel's sales divided by total sales.
DWL = Deadweight Loss (competitive welfare − actual welfare).
Price Markup (as a percentage) = $100(p - MC)/p$.

in it at $p = 10$. Because cartel firms have the same cost functions as noncartel firms, the cartel cannot afford to produce at prices below $p = 10$ either, so the lower portion of the residual demand curve is not of interest. The profit-maximizing cartel chooses its output, 240, by setting its marginal revenue (the curve marginal to its residual demand curve) equal to its marginal cost, as shown in Figure 5.2b. This output determines the cartel price, 24. At that price, the noncartel firms' output is 280, as Figure 5.2a shows.

Table 5.2 shows what happens as the number of firms belonging to the cartel changes. The market is in competitive equilibrium when all 50 firms act independently and do not belong to the cartel ($j = 50$). The competitive market price is $21.43, and consumer surplus and total welfare are maximized.

At the other extreme, if all the firms join the cartel ($j = 0$), the cartel is a monopoly. The monopoly price is $33.33, or 56 percent higher than the competitive price. Only 333 units of output are produced by the market, or 58 percent as much as the competitive quantity, 571. However, each firm's profits of $133.33 is more than double the competitive level, $65.31. Consumer surplus is only about one-third as great, and total welfare is only 83 percent as great as under competition; that is, consumer losses are greater than the cartel gains. This loss to society, the *deadweight loss* to monopoly (see Chapter 4), is 18 percent of sales and 71 percent of consumer surplus (at the monopoly price).

			Cartel Firm		Noncartel Firm	
DWL as % of Sales	Cartel's Market Share (%)	Price Markup (%)	Output	Profits	Output	Profits
17.9	100	50	6.66	133.33	—	—
15.9	94	48	6.72	128.02	22.41	251.10
6.5	63	36	7.27	96.95	16.97	143.99
2.5	46	25	8.00	80.00	14.00	98.00
0.7	32	16	8.89	71.08	12.44	77.38
0.1	18	8	10.00	66.70	11.67	68.09
0.0	2	1	11.27	65.32	11.431	65.33
—	0	—	—	—	11.429	65.31

As the cartel gains members, the incentive of a cartel firm to cheat also grows, because the discrepancy between a nonmember's profit and a cartel firm's profit increases. At every price, nonmembers earn more than cartel members, because nonmembers produce more yet sell at the same price as cartel members.

Consumers benefit if firms refuse to join the cartel. If only one firm refuses to abide by the cartel rules and is a price taker, market price is about 3 percent lower than the monopoly price, deadweight loss is 10 percent lower, and consumer surplus is 10 percent higher.

Table 5.2 also shows that it hardly pays for one firm to try to act as a price setter by itself. If one firm forms a "cartel" consisting only of itself (so that there are 49 noncartel firms), it faces a residual demand curve with only a slight slope. It can reduce its output from the competitive level of 11.429 to 11.27 units, thereby maximizing its profits, which rise by 1¢ from $65.31 to $65.32. The 49 noncartel firms, seeing the higher price and acting as price takers, increase their output to 11.431 units, causing profits to rise by a phenomenal 2¢ to $65.33—each noncartel firm's profits rise by more than those of the single-firm cartel. All firms' profits rise in this case because the expansion of output by the 49 noncartel firms does not completely offset the reduction in output by the single firm. Total market output falls from 571.43 to 571.38 units, causing the price to rise from $21.429 to $21.431. The welfare losses from such a limited cartel are small. The larger the market share of the cartel, the greater the efficiency cost. See Example 5.7.

EXAMPLE 5.7

The Social Costs of Cartelization

Posner (2003) estimates the social cost of several (mainly international) well-organized, overt cartels using his theory, discussed in Chapter 4, that all cartel profits are dissipated in rent-seeking activity. His results are shown in the table.

Industry	Cartel Price Increase (%)	Elasticity	Social Costs (as % of Industry's Sales)
Nitrogen	75	2.33	64
Sugar	30	4.33	35
Aluminum	100	2.00	75
Aluminum	59	3.63	56
Rubber	100	2.00	75
Electric bulbs	37	3.70	44
Copper	31	4.22	36
Cast-iron pipe	39	3.56	42

Note: These figures are based on the Appendix in Posner (2003). Apparently, two figures are given for aluminum because he has two estimates of the cartel price increase.

The elasticities are based on the cartels' price-increase data, on the assumptions that the industry is charging the profit-maximizing monopoly price and that the demand curve is linear. Because of these restrictive assumptions, Posner warns that these results should be viewed with some caution. Nonetheless, if these figures are anywhere near accurate, the social costs of these cartels are large.

Source: R. A. Posner, *Antitrust Law,* ©2003 by The University of Chicago. All Rights Reserved.

Price-Fixing Laws

> *Nothing is illegal if a hundred businessmen decide to do it.*
>
> *—Andrew Young*

In the late nineteenth century, cartels were legal and operated in several U.S. industries, including oil, powder, railroads, sugar, and tobacco. The Sherman Antitrust Act of 1890 was passed in response to this activity "to protect trade and commerce against unlawful restraints and monopolies" (see Chapter 19). In 1914, the Federal Trade Commission Act established the Federal Trade Commission (FTC), and its Section 5 holds that "unfair methods of competition are hereby declared illegal." This act is still used by the FTC in prosecuting antitrust violations, as is the Sherman Act by the U.S. Department of Justice.

The Sherman Act makes illegal conspiracies whose sole purpose is to raise price. For example, in the *Addyston Pipe and Steel* case in 1899, outright bid rigging and the dividing of the market into regional monopolies were found illegal.[27]

The *Trenton Potteries* case of 1927 and *Socony-Vacuum Oil Co.* case of 1940 determined that price fixing was a per se violation (the action by itself is illegal), regardless of whether the price set was above the competitive price.[28] These cases established that the violation is the *attempt* to charge the monopoly price; the government does not have to show that the defendants succeeded in their attempt (Posner 2003). Cartels formed solely to raise price are strictly prohibited.

Table 5.3 (based on Connor 2003) shows the number of price-fixing cases the DOJ filed over the period 1970–99, how many fines were collected, and the number in which individuals received prison sentences. Example 5.8 shows that U.S., Canadian, EC, and other antitrust authorities have heavily fined corporations engaged in global conspiracies since the early 1990s.

This approach to preventing price fixing is based on evidence of conspiracy rather than the economic effects of the conspiracy. The government seeks evidence of conspiracies (such as secret meetings in smoke-filled rooms) rather than economic evidence (such as price increases). Cases involving only tacit collusion (that is, without explicit communications between the parties) are not actionable under antitrust laws.

The current laws have been successful in eliminating overt (but not tacit) collusion. Posner (2003, 52) observes that the "elimination [of the cartel] is an impressive and remains the major, achievement of American antitrust." Increasingly, many other countries, especially those in Europe, actively try to prevent cartels.

TABLE 5.3 Fines and Sentences in U.S. Department of Justice Price-Fixing Cases, 1970–1999

Years	Number of Criminal Cases Filed	Cases in Which Fines Imposed	Cases in Which Prison Sentences Imposed
1970–1979	176	156	25
1980–1989	623	513	196
1990–1999	416	324	61

Source: Connor (2003).

[27] *Addyston Pipe and Steel Co. v. United States,* 175 U.S. 211 (1899). This citation is from the U.S. Reporter volume 175 and starts on page 235. The case was decided by the U.S. Supreme Court in 1899.

[28] *United States v. Trenton Potteries Co.,* 273 U.S. 392 (1927) established a per se rule. A later case, *Appalachian Coals, Inc. v. United States,* 288 U.S. 344 (1933), however, appeared to deviate from this per se rule. *United States v. Socony-Vacuum Oil Co.,* 310 U.S. 150 (1940) firmly established the per se rule.

EXAMPLE 5.8 *Prosecuting Global Cartels*

Between World War II and the 1990s, few private (non–government-run) global cartels were observed, although there were a number of government-organized cartels, as discussed in Example 5.2. The period since 1990 has seen an explosion of private international cartels that have been identified and prosecuted by antitrust agencies in North America and Europe. From 1993 through July 2003, antitrust authorities discovered (engaged in public investigation, handed down indictments, or imposed fines) 167 private cartels with corporate or individual participants from at least two countries. By July 2003, 128 were at least partially prosecuted, and public investigations were initiated for another 39. Moreover, approximately 35 secret U.S. grand-jury investigations of international cartels were ongoing.

Since the early 1990s, U.S., Canadian, EC and other antitrust authorities have heavily fined corporations engaged in global conspiracies. The prosecutions of 20 global (multicontinent) cartels after 1995 resulted in heavy fines in all cases and prison sentences in half. From 1996 through mid-2003, U.S., Canadian, EC and other antitrust authorities have imposed fines of over $5.3 billion on corporations engaged in international conspiracies.

Of the cartels, 31% operated in two or more continents, with most of them in North America, Europe, and Asia (covering 51% of affected sales); 20% had sales within more than one EU country (25% of sales); 26% operated in a single European country (13% of sales); 19% were only in the United States or Canada (11% of sales); and only six international cartels were discovered elsewhere.

Compared to previous international cartels, their markets encompassed more of the world, and hence the damages they inflicted were accordingly greater. Firms engaged in international cartels that were successfully prosecuted by the U.S. Department of Justice (DOJ) had sales that exceeded $55 billion. The DOJ collected over $900 million from international price fixers in 1999 alone—far more than it collected in the previous 108 years of U.S. antitrust enforcement.

For the 60 cartels for which he had adequate information, Connor (2003) reported that the cartels raised prices by an average of 28% (25% in organic chemicals and 35% in other industries). International cartel sales were concentrated in a few industries: 39% involved intermediate organic chemicals, 52% were other manufacturers (mostly metal, cement, plastics, graphite products) and the remaining 9% were in

SUMMARY

Firms have incentives both to form cartels and to cheat on the cartel arrangement. Firms want to join cartels if the cartel is capable of raising prices for sustained periods of time. Prices are more likely to be significantly elevated above the competitive level when the cartel controls a substantial share of the market's output, when it faces a relatively inelasconstruction, transport, finance, and other services. However, over 80% of sales in cartels discovered before 2000 involved food and agriculture.

The world's major antitrust agencies have responded to these threats to worldwide economic well-being by imposing unprecedented sanctions (1990–2003: $2.3 billion by the United States, €3.6 billion by the European Union). Of those corporations accused by the DOJ of criminal price fixing, fewer than 1% were foreign-based firms prior to 1995, whereas more than 50% were non-U.S. corporations after 1997. The DOJ has convicted cartel executives from 12 foreign countries, sending many to prison. Similarly, between 2000 and 2002, the EC fined 42 companies that were guilty of global price fixing, of which 55% were non-EU firms.

Executives were fined in 50% of cases and received prison sentences in 33% of U.S. cases. Sixty-two executives were fined, and 43 fugitives were indicted. The individual fines added up to $24.4 million, and four fines were greater than $350,000, but the median fine was only $50,000. Of the 62 executives fined, 30 received prison sentences averaging 11.1 months. Contrary to DOJ claims for all criminal price-fixing cases, Connor failed to find an upward trend in the length of prison sentences for participants in international cartels.

Canada prosecuted 18 cases between 1991 and 2003. Most of these cases followed U.S. actions. The initial Canadian prosecution started an average of eight months after a U.S. conviction. Canada fined 68 corporations $133 million (U.S.), almost all of which were non-Canadian. These fines were about 6% of the corresponding U.S. fines. Canada also fined four individuals a total of $600,000.

From 1990 to 2003, the EU prosecuted 35 cases, of which 16 were global cartels. The EU fined or granted amnesty to 259 corporations, of which 30% were located outside the EU. On average, EU decisions lagged U.S. prosecutions by 34 months. The EU fines were about 72% of comparable U.S. fines on the same cartel cases.

Although the antitrust agencies have acted more aggressively than in the past and are imposing larger fines, cartels continue to thrive. For the period 2000 through 2003, 23 international cartels were discovered on average per year, six times faster than a decade earlier. Apparently firms believe that the fines are merely a cost of doing business, as more than 50 corporations were members of multiple cartels (up to 13).

Source: Connor (2003).

tic demand curve, and when entry is limited. The expected rewards of forming illegal cartels are greater when detection by the government is unlikely and the fines are low.

Cartels fail due to cheating by member firms or by competition from firms outside the cartel. Individual firms have an incentive to cheat on a cartel agreement because they can make higher profits by increasing output or undercutting the cartel's price. A cartel can maintain its agreement only if cheating can be detected and adequately punished. Cartels have developed a number of techniques, including division of the market and complex contract clauses, in order to enforce their agreements.

When cartels succeed in raising prices, there is a loss of consumer surplus. The gain to the cartel is less than the loss to consumers: The difference is a deadweight (efficiency) loss. The fewer the firms that go along with the cartel agreement, the less market power the cartel has, and hence the less it harms consumers and society.

The U.S. government and many others have antitrust laws that penalize firms that form cartels. At least in the United States, price-fixing cartels have been vigorously prosecuted.

PROBLEMS

1. Historically, at each Organization of Petroleum Exporting Countries (OPEC) meeting, Saudi Arabia (the largest oil producer) argued that the cartel should cut production. The Saudis complained that most OPEC member countries, including Saudi Arabia, produced more oil than their cartel agreement allotted them. Illustrate using a graph and explain why cartel members would produce more than the allotted amount given that they know that overproduction will drive down the price of their product.
2. What are the main factors that increase the likelihood of a cartel being successful?
3. Use a graph to show why an increase in the market demand elasticity reduces a cartel's monopoly power. Show how an increase in the market demand elasticity affects the elasticity of the residual demand curve.
4. (Problem based on Appendix 5A.) Show that the sum of a cartel's output plus the output of noncartel firms is less than the competitive output and that the corresponding price is higher than the competitive price.
5. (Problem based on Appendix 5A.) Show that a cartel's price falls as the number of noncartel firms (j) increases.

Answers to odd-numbered problems are given at the back of the book.

SUGGESTED READINGS

A good survey of modern thought on cartels is Jacquemin and Slade (1989). If you want to see how an actual cartel operates, go to **www.aw-bc.com/carlton_perloff**, Chapter 5, "A Cartel at Work." There, you will find a link (**www.usdoj.gov/atr/public/speeches/4489.htm**) to a speech given April 6, 2000 by James M. Griffin, Deputy Assistant Attorney General, Antitrust Division, U.S. Department of Justice, entitled "An Inside Look at a Cartel at Work: Common Characteristics of International Cartels," in which Griffin outlines how the international lysine cartel worked. The Department of Justice can supply you with an actual tape of the lysine cartel meetings. Contact the United States Department of Justice, Antitrust Division, Freedom of Information Act Unit, 325 Seventh Street, N.W., Suite 200, Washington, DC, 20530. Connor (2001) extensively analyzes global cartels since 1990.

APPENDIX 5A

The Effects of Cartel Size

This appendix derives the equations used in the example reported in Table 5.2, which shows how price and output vary with the number of cartel members. The total number of firms is assumed to be fixed at n—no further entry is possible.

The market demand curve is linear:

$$Q = a - bp, \tag{5A.1}$$

where a and b are positive constants, Q is market output, and p is the price. The elasticity of demand is

$$\epsilon = \frac{dQ}{dp}\frac{p}{Q} = 1 - \frac{a}{Q} = \frac{-bp}{a - bp}. \tag{5A.2}$$

Each firm has a linear marginal cost (MC) of

$$MC = d + eq, \tag{5A.3}$$

where q is the output of one of the n firms and d and e are positive constants. As a result, the competitive supply (the output produced at the point where marginal cost equals price) is

$$Q = nq = \frac{n(p - d)}{e}. \tag{5A.4}$$

Competitive equilibrium is determined by setting the right-hand sides of the quantity-demanded equation (5A.1) and the quantity-supplied equation (5A.4) equal, and solving for p_c (the equilibrium price). The equilibrium quantity, Q_c, can be found by substituting p_c into Equation 5A.1 or 5A.4. The equilibrium values are

$$p_c = \frac{ae + nd}{be + n}, \tag{5A.5}$$

$$Q_c = n\left(\frac{a - bd}{be + n}\right). \tag{5A.6}$$

Now suppose that $n - j$ firms in the market form a cartel and the remaining j firms ($j < n$) do not. As shown in Figure 5.2b, the residual demand, Q_r, is the market demand minus the noncartel supply, $Q_{nc} = jq$:

$$Q_r = Q - jq = a - bp - \frac{j(p - d)}{e}. \tag{5A.7}$$

The cartel acts as a monopoly with respect to its residual demand and sets its marginal revenue, MR_m, equal to its marginal cost. The cartel's revenues, R_m, may be found by solving Equation 5A.7 for p as a function of Q_r and multiplying that by Q_r to obtain

$$R_m = pQ_r = \left(\frac{ae + jd - eQ_r}{be + j}\right)Q_r. \tag{5A.8}$$

By differentiating R_m with respect to Q_r, we obtain the cartel's marginal revenue:

$$MR_m = \frac{ae + jd}{be + j} - \left(\frac{2e}{be + j}\right)Q_r. \tag{5A.9}$$

The cartel's marginal cost is

$$MC_m = d + \left(\frac{e}{n - j}\right)Q_m. \tag{5A.10}$$

The quantity the cartel chooses to produce, Q_m ($= Q_r$), is determined by equating the cartel's marginal revenue, Equation 5A.9, and marginal cost, Equation 5A.10:

$$Q_m = \frac{(n - j)(a - bd)}{be + 2n - j}. \tag{5A.11}$$

By differentiating Q_m with respect to j, it can be shown that the cartel's output falls as the number of nonmember firms rises.

CHAPTER 6

Oligopoly

The Puritan's idea of hell is a place where everybody has to mind his own business. —*Wendell Phillips (attributed)*

Although there is only one model of competition and one model of monopoly, there are many models of oligopoly: a small number of firms acting independently but aware of one another's existence. Unlike monopolistic and competitive firms, noncooperative oligopolists cannot blithely ignore other firms' actions.

As the only firm in an industry, a monopolist has no rivals. At the other extreme, individual competitive firms are too small to affect the industry's price, so each firm reasonably ignores the actions of any other; only the industry's collective actions matter to a competitive firm. In contrast, because there are only a few firms in an oligopoly, each firm knows that it can affect market price and hence its rivals' profits: Ford cannot and does not ignore Honda when making decisions. Thus, oligopoly differs from competition and monopoly in that a firm *must* consider rival firms' behavior to determine its own best policy. This interrelationship between firms is the key issue examined in this chapter. The factors discussed in Chapter 5 that influence a cartel's success also affect how oligopolistic rivals interact. For that reason, Stigler (1964a) regarded cartel theory as the basis for understanding the forces at work in any oligopoly.

Many industries are highly *concentrated* (see Chapter 8): A few firms make virtually all the sales. For example, the top four cereal manufacturers sell 90 percent of all breakfast cereals, and the top eight sell 98 percent. Only a handful of manufacturers produce many common consumer durables.

Where transportation costs or tariffs are so high that it is not cost-effective to ship a product outside of a small geographic region or local market, oligopolies are common. In countries with smaller markets (fewer consumers), many industries are oligopolistic.[1]

This chapter presents the best-known noncooperative oligopoly models. To keep the discussion as simple as possible, five strong assumptions are made:

1. Consumers are price takers.
2. All firms produce *homogeneous* (identical) products: Consumers perceive no differences among them.
3. There is *no entry* into the industry, so the number of firms remains constant over time.
4. Firms collectively have market power: They can set price above marginal cost.
5. Each firm only sets its price or output (not advertising or other variables).

The next chapter extends these models to consider *heterogeneous* (differentiated) products and entry of new firms. Chapter 14 discusses advertising, and Chapter 11 discusses other strategic actions beyond setting price or output.

The equilibrium price in an oligopoly market lies between that of competition and monopoly. In all of the oligopoly models, each firm maximizes its profits given its beliefs about how other firms behave: Each firm's expected profits are maximized when its *expected marginal revenue equals its marginal cost.* As in earlier chapters, a firm's marginal revenue depends on the *residual demand curve* facing that firm (the market demand minus the output supplied by its rivals). Indeed, the differences in the various oligopoly models are reflected in terms of differences in the residual demand curves facing firms.

All the oligopoly models may be seen as examples of noncooperative *game theory* (von Neuman and Morgenstern 1944), which uses formal models to analyze conflict and cooperation between **players** (strategic decision makers—firms in this chapter).[2] A **game** is any competition in which strategic behavior is important. Each firm forms a *strategy* or battle plan of the actions it will take (such as the prices it will set) to compete with other firms. Each firm's **payoff** (the reward received at the end of a game, profits) depends on the actions of all the firms.

The various oligopoly models differ in the type of actions firms may use (such as set prices or set outputs), the order in which they may take actions (such as which firm sets its price first), the length of the game (one-period model or many periods), and in other ways. Although there is extensive agreement among economists on the model of competition and the model of monopoly, there is no consensus on a single noncooperative oligopoly model. One reason for the lack of agreement is that market characteristics of real-world oligopolies differ substantially, so that the appropriate model varies by market.

Three of the best-known oligopoly models are the Cournot, Bertrand, and Stackelberg models. Firms set output levels in the Cournot and Stackelberg models, whereas

[1] Similarly, *oligopsonies* (few buyers) are frequently observed. For years, only two firms purchased most of the mussels caught in New England, and four firms bought most of the Pacific tuna.

[2] Another branch of game theory, cooperative game theory, is rarely used in modeling oligopolies. A notable exception is Telser (1972, 1978).

they set prices in the Bertrand model. All the firms act at the same time in the Cournot and Bertrand models, whereas one firm sets its output level before the others in the Stackelberg model. These differences in the actions firms use and the order in which they act result in different equilibria.

Similarly, some markets last for only one period and others last for many. For example, a *static* or *one-period* game model is appropriate when firms from all over the country meet only once at a one-day crafts fair. Such firms set their price or quantity that day and do not have an opportunity to observe how their rivals behave and then change their own behavior in the future.

A *multiperiod* game model should be used to analyze how two crafts shops that are located next to each other compete day after day for years. Where firms compete repeatedly over time, firms may adjust their beliefs about rivals' behavior over time and may use more complex strategies than in single-period models. For example, a firm's strategy might require it to set different output levels depending on how its rivals behaved in previous periods. One possible outcome of such a model is that firms may restrict output in early periods and then produce larger quantities in the last period.

Only sets of actions by a rival that are in the rival's best interests are considered credible strategies by a firm. For example, if one firm threatens to price below cost forever unless another rival leaves the market, that strategy is not credible because it would lead to the bankruptcy of the price cutter (see Chapter 11). In multiperiod games, more complex credible strategies are possible than in a single-period game.

After discussing the basic concepts of noncooperative game theory, this chapter presents the three best-known oligopoly models: Cournot, Bertrand, and Stackelberg. The chapter concludes by presenting experimental and empirical tests of various oligopolistic models that support the predictions of some of these models.

The key questions examined in this chapter are

1. What factors determine the oligopoly equilibrium? How does the equilibrium vary with the number of firms, the types of actions firms may take, and the order in which firms act?
2. Is the equilibrium price more likely to be closer to the monopoly price when markets last for more than one period?
3. Are the best-known oligopoly models consistent with experimental evidence?

Game Theory

When the One Great Scorer comes to write against your name—He marks—not that you won or lost—but how you played the game.

—Grantland Rice

Game theory analyzes the interactions between rational, decision-making individuals who may not be able to predict fully the outcomes of their decisions. Models of oligopoly can be viewed as games of strategies or actions (such as setting output, price, or advertising levels). Oligopolistic games have three common elements:

1. There are two or more firms (players).
2. Each firm attempts to maximize its profit (payoff).
3. Each firm is aware that other firms' actions can affect its profit.

The third element is the crucial one. Oligopolistic markets differ from competitive and monopolistic markets because each firm's actions significantly affect its rivals. For example, oligopolists may form cartels for the purpose of mutually beneficial actions; yet, because each firm's interests are different from those of other firms, the best outcome for a particular firm is not always in the collective best interest.

In competitive and monopolistic markets, firms do not act as if the actions of rivals affect their payoffs or as if other firms may have different objectives. Indeed, the competitive model may be viewed as a game against an impersonal mechanism (the market) rather than against other players with strategies.

The equilibrium payoffs are dictated by the number of firms, the rules of the game, and the length of the game. The major single-period oligopoly models differ as to the rules of the game. After showing how the best-known single-period models vary depending on the rules of the game and the number of firms, the chapter illustrates the effect of the length of a game by contrasting single-period games to multiperiod games of various lengths.

Single-Period Oligopoly Models

Early work on oligopoly theory concentrated on single-period, or static games. Such models are appropriate for markets that last for only brief periods of time, so that rival firms compete once, but never again. In such models, complex, long-run strategies and reputations for hard-nosed competition are irrelevant.

The three best-known single-period oligopoly models date from long before the introduction of game theory. These models can be interpreted as game theoretic models, and that is how this chapter presents them.[3] All well-known single-period oligopoly models use the concept of a Nash equilibrium, which we discuss before turning to the Cournot, Bertrand, and Stackelberg models.

[3] Earlier versions of these models were called *conjectural variation models* (see **www.aw-bc.com/carlton_perloff** "Conjectural Variation"). Game theorists view these conjectural variation models as unsatisfactory because they use dynamic stories to explain behavior in a single period. In the conjectural variations models, each firm chooses its price or output to maximize its profit based on its conjecture (hypothesis or expectation) about how each rival firm will respond to its actions (*variations*). For example, a firm may believe its rival will do nothing in response if the firm raises its price. A belief about a rival's reaction is called a *conjectural variation*. The conjectural variation approach has been used in empirical work (see **www.aw-bc.com/carlton_perloff** "Conjectural Variation" and Chapter 8). Another early static model, sometimes described as a conjectural variations model, that has an underlying, implicitly dynamic story is the kinked demand curve: see **www.aw-bc.com/carlton_perloff** "Kinked Demand."

Nash Equilibrium

John F. Nash (1951) defined the most widely used equilibrium concept. A set of strategies is called a **Nash equilibrium** if, holding the strategies of all other firms constant, no firm can obtain a higher payoff (profit) by choosing a different strategy. Thus, in a Nash equilibrium, no firm *wants* to change its strategy.

In the Cournot and Stackelberg models, firms' strategies concern setting quantities. In the Bertrand model, firms set prices. The Nash equilibrium concept is also useful when strategies include setting advertising or other variables in addition to output or price (see Chapter 11).

The Cournot Model

The French mathematician Augustin Cournot presented the first—and probably still the most widely used—model of noncooperative oligopoly in 1838. Cournot (1963) assumed that each firm acts independently and attempts to maximize its profits by choosing its output. The discussion starts with the *duopoly*, or two-firm, case and then considers what happens as the number of firms increases.

A Cournot Duopoly. Consider a market of melons in an isolated town:

- *No entry:* There are two firms and no entry by other firms is possible (these firms own the only good farm land anywhere in the area).
- *Homogeneity:* The firms produce identical (homogenous) melons, so the sum of their outputs equals industry output: $Q = q_1 + q_2$, where Firm 1 produces q_1 and Firm 2 produces q_2.
- *Single period:* This market and the two firms only exist for one period. The melons cannot be stored: They must be sold as soon as produced or they spoil.
- *Demand:* The market demand curve (Figure 6.1) is a linear function of price:

$$Q = 1{,}000 - 1{,}000p. \tag{6.1}$$

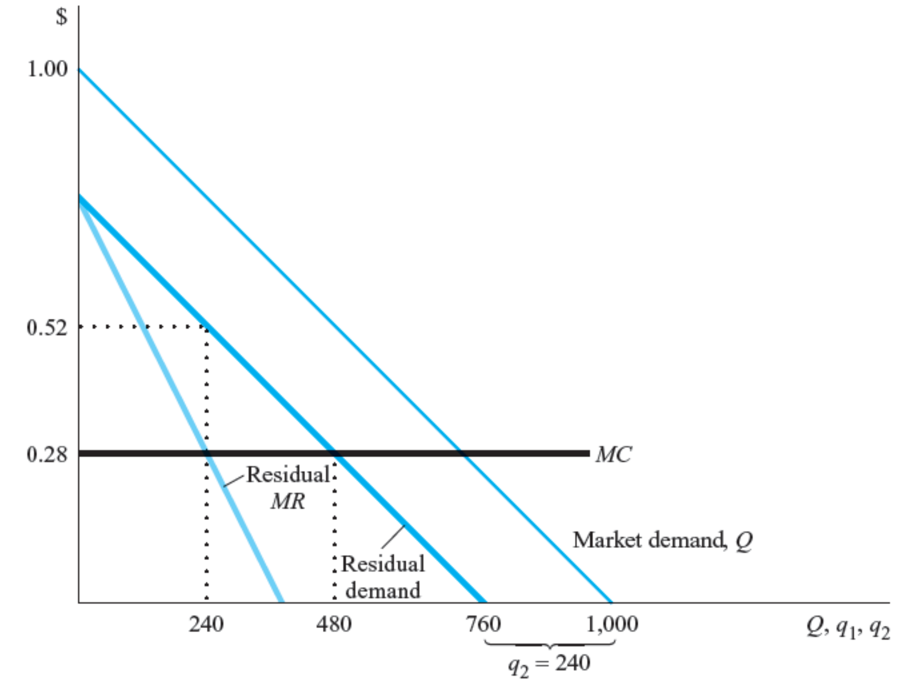
  For example, $Q = 0$ melons when $p = \$1.00$, $Q = 500$ when $p = \$0.50$, and $Q = 1{,}000$ when $p = 0$.
- *Costs:* Each firm has a constant marginal cost, *MC*, of production of 28¢ per melon and no fixed costs. Thus, its average cost is also 28¢. Each firm can produce enough output to meet the entire market's demand, as Figure 6.1 shows.

What strategy should Firm 1 use to choose its output level? The answer depends on its belief about Firm 2's behavior. If Firm 1 believes that Firm 2 will sell q_2 melons, it can determine the q_1 that will maximize its profit. Firm 1 can sell all but q_2 units of the amount demanded by the market; that is, it faces the *residual demand curve,*

$$q_1 = Q(p) = q_2, \tag{6.2}$$

which is the market demand curve from Equation 6.1, minus the expected output of Firm 2, q_2. As Figure 6.1 shows, the residual demand curve is obtained by shifting the market demand curve q_2 units to the left. Thus, because the market demand curve hits

FIGURE 6.1 Residual Demand Facing a Cournot Firm



the horizontal axis at 1,000 melons, the residual demand curve hits the horizontal axis at $1{,}000 - q_2$. In the figure, q_2 is assumed to be 240 units.

Firm 1 has a monopoly over those consumers whose demands are not met by Firm 2. To maximize its profit, it sets q_1 where its marginal revenue curve based on the derived residual demand curve (residual *MR* in Figure 6.1) intersects its marginal cost curve. The profit-maximizing q_1 for various beliefs about q_2 are as follows:

If Firm 1 believes Firm 2 will sell q_2	Firm 1's profit-maximizing q_1 is
0	360
100	310
200	260
240	240
300	210
360	180
400	160
720	0

Rather than use a table, we can summarize the relationship between Firm 1's profit-maximizing quantity and Firm 2's quantity in an equation,

$$q_1 = R_1(q_2), \tag{6.3}$$

which is called a *best-response function* (or *reaction function*), which shows the best (highest profit) action (output) by a firm given its beliefs about the action its rival takes. To derive the best-response function, it is necessary to express the intersection between the marginal revenue curve and the marginal cost curve algebraically (see Appendix 6A for a mathematical derivation).

Firm 1's residual demand curve is linear, so its marginal revenue curve is also linear and has twice the slope of the residual demand curve: The *MR* curve hits the quantity axis at half the quantity of the demand curve (see Chapter 4). In Figure 6.1, where q_2 equals 240, the residual demand curve intersects the horizontal *MC* curve at $q_1 = 480$. In general, the residual demand curve intersects the marginal cost curve at $720 - q_2$. The marginal revenue curve corresponding to the residual demand curve crosses the marginal cost curve at half that value, or where $q_1 = 240$.[4] More generally, Firm 1's best-response function is

$$q_1 = R_1(q_2) = 360 - \frac{q_2}{2}, \tag{6.4}$$

as Figure 6.2 shows. If $q_2 = 0$, Firm 1 produces $q_1 = R_1(0) = 360$, the monopoly output level. The residual demand curve of a Cournot firm facing no competition is the market demand curve. Because the market demand curve intersects the marginal cost curve at 720, a monopoly's marginal revenue curve intersects the marginal cost curve at half that quantity, or 360. At the other extreme, Firm 1 does not cease production until $q_2 = 720$.

Firm 2's best-response function is derived in a similar way. The firms are identical (same costs, identical products), so Firm 2's best-response function is the mirror image of Firm 1's:

$$q_2 = R_2(q_1) = 360 - \frac{q_1}{2}. \tag{6.5}$$

Firm 2's choice of output depends on the output it expects Firm 1 to produce.

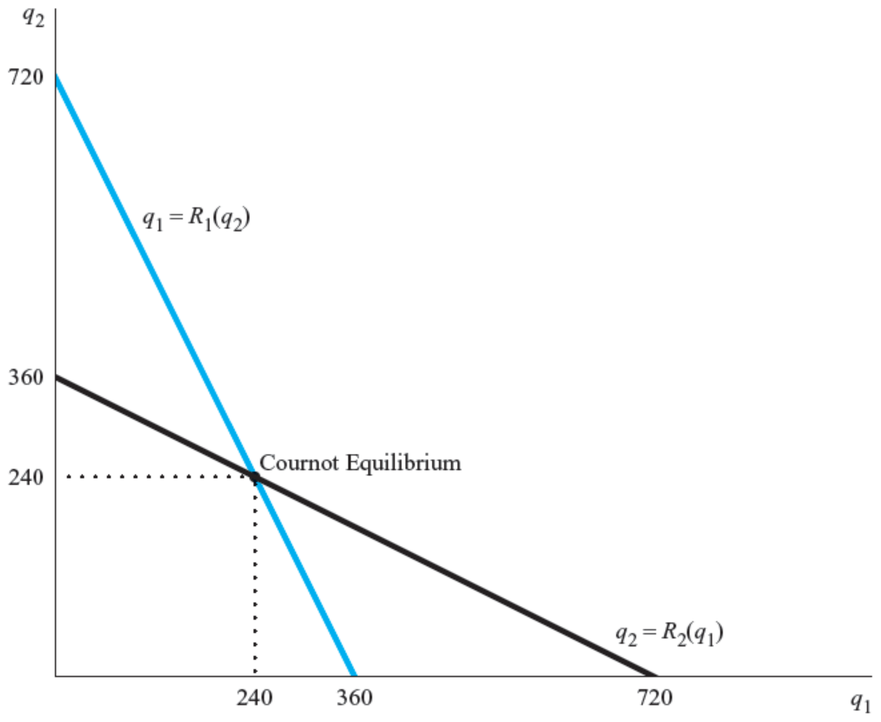
As Figure 6.2 and Table 6.1 illustrate, the two firms' best-response functions cross once at $q_1 = q_2 = 240$.[5] At the intersection of the best-response functions, if each firm believes that the other firm will sell 240 units, it wants to sell 240 units too.

Equilibrium. This point of intersection (240, 240) of the best-response functions is called a Cournot equilibrium. In the Cournot equilibrium, each firm sells the quantity

[4] If Firm 2 produces $q_2 = 240$, the residual demand curve facing the first firm is $q_1 = Q(p) - q_2 = (1{,}000 - 1{,}000p) - 240 = 760 - 1{,}000p$, or $p = 0.76 - 0.001q_1$. Thus, the first firm's revenue is R $= pq_1 = 0.76q_1 - 0.001q_1^2$, so its residual marginal revenue function is $dR/dq_1 = 0.76 - 0.002q_1$. Residual marginal revenue equals marginal cost where $0.76 - 0.002q_1 = 0.28$, or $q_1 = 240$.

[5] The intersection can be determined algebraically by simultaneously solving the two best-response function equations. Substituting Firm 1's best-response function $q_1 = 360 - q_2/2$ for q_1 in Firm 2's best-response function, $q_2 = 360 - q_1/2$, Firm 2's output is $q_2 = 360 - 1/2(360 - q_2/2)$. Simplifying, $q_2 = 240$. Substituting 240 for q_2 in Firm 1's best-response function, we learn that q_1 also equals 240.

FIGURE 6.2 Cournot Best-Response Functions



that maximizes its profits given its (correct) beliefs about the other firm's choice of output—the *best response* to the other firm's output level. Moreover, in equilibrium, each firm's beliefs about its rival's output is confirmed.

If each firm believes the other will produce 240 units and if each firm produces 240 units, neither firm wants to change its output. A firm is unwilling to produce at a point not on its best-response function because doing so would result in a lower profit. The only point where both firms are on their best-response functions is at the intersection of the best-response functions. A nonintersection point cannot be an equilibrium; an equilibrium point is one in which neither firm wants to change its behavior. In the Cournot equilibrium, total market output is 240 + 240 = 480 melons and the price is 52¢ per melon (Table 6.1).

In a single-period model in which firms only choose output levels, any output levels at which no firm believes it can increase its profits by increasing or decreasing its output are, by definition, a Cournot equilibrium, and no combination of outputs could be an equilibrium except a Cournot equilibrium. Thus, in a single-period model in which firms choose output levels independently, the Cournot equilibrium is not only realistic; it is the only plausible equilibrium (Friedman 1983, 32–3).

TABLE 6.1 A Comparison of Oligopoly Equilibria: A Linear Demand and Constant Marginal Cost Example

	Output			Profits ($)		
	Firm	Industry	Price (¢)	Firm	Industry	Consumer Surplus
Monopoly	360	360	64	129.60	129.60	64.8
Cournot Duopoly	240	480	52	57.60	115.20	115.2
Stackelberg Duopoly		540	46		97.20	145.8
Leader	360			64.8		
Follower	180			32.4		
Competition*		720	28	0	0	259.2
Cournot: *n* firms	$\frac{720}{n+1}$	$\frac{720n}{n+1}$	$\frac{100+28n}{n}$	$\frac{518.4}{(n+1)^2}$	$\frac{518.4n}{(n+1)^2}$	$\frac{259.2n^2}{(n+1)^2}$
Stackelberg: *n* firms		$\frac{360\,(2n-1)}{n}$	$\frac{28n+36}{n}$		$\frac{129.6\,(2n-1)}{n^2}$	$\frac{64.8\,(2n-1)^2}{n^2}$
Leader	360			$\frac{129.6}{n}$		
Followers	$\frac{360\,(n-1)}{n}$			$\frac{129.6\,(n-1)}{n^2}$		

Market demand: $Q = 1{,}000 - 1{,}000\,p$
$MC = 28$¢
*Efficient point, Bertrand equilibrium, Cournot equilibrium with unlimited number of firms.

The remaining question is how do firms form their beliefs? Is it reasonable that they each choose to believe that the other firm will produce 240 units? One practical answer is that experience often influences beliefs, but if we introduce experience, we are bringing a dynamic element into a supposedly static model. The failure to provide a theoretical basis for underlying beliefs is a criticism of the Cournot model and other static models.[6] This criticism led Stigler (1964a) to develop his analysis based on cartel theory (Chapter 5) and game theorists to develop multiperiod game models (discussed later in this chapter).

[6] One interesting response is provided by Daughety (1985), who describes the type of *infinite regress* reasoning firms might use. Firm 1's model of Firm 2 is subject to Firm 2's model of Firm 1's model and vice versa. That is, Firm 1's manager thinks about what Firm 2's manager is thinking: "I think that Firm 2's manager thinks that I think that Firm 2's manager thinks that I think . . ." Firm 2's manager thinks about Firm 1's manager's model similarly. Based on this type of reasoning, each firm chooses an output level. Daughety shows that the Cournot equilibrium is the only possible result of this type of reasoning.

EXAMPLE 6.1 *Do Birds of a Feather Cournot-Flock Together?*

Flocking birds must frequently look up (or *scan*) while feeding to see an approaching predator in time. Frequent scanning, however, decreases the feeding rate of individual birds. Because any bird can give the warning of impending doom (which greatly reduces the probability of death), it is in an individual bird's best interest for another member of the flock to scan (incur costs) while it eats constantly (benefits). Not surprisingly, as the size of the flock increases, each bird spends less time scanning and more time feeding.

Using high-speed cameras with telephoto lenses and trained predators that fly over the flock from time to time, scientists determined how members of a flock of yellow-eyed juncos behave (Pulliam, Pyke, and Caraco 1982). The scientists compared two game-theoretic models as explanations for this behavior. In the cooperative model, the birds work together, whereas in the selfish solution (analogous to the Cournot-Nash model), they operate independently. The following table compares the observed and predicted scanning rates (as percentages of time) under the cooperative and selfish models (using the most likely parameter estimates of each model—which explains why the numbers are different in the two models when there is only one bird in the "flock"):

		Predicted	
Number in Flock	Observed	Cooperative Model	Selfish Model
1	13.9%	15.9%	18.6%
2	7.85	6.2	3.4
3	6.22	5.9	0.6
4	6.02	5.5	0.0
5	5.87	5.2	0.0
6	5.66	4.9	0.0
7	5.58	4.7	0.0
8	5.59	4.5	0.0
9	4.88	4.4	0.0
10	4.65	4.0	0.0

In Cournot's equilibrium concept, no firm wants to change its output level given that the other firms produce at the equilibrium quantities. Because the Cournot equilibrium is a special case of the Nash equilibrium where firms have strategies over quantities, it is often referred to as a *Cournot-Nash equilibrium* or a *Nash-in-quantities equilibrium*. Example 6.1 describes a nonbusiness game in which the Nash generalization is apt.

A Comparison of the Cournot and Cartel Equilibria. The firms are worse off and the consumers are better off at the Cournot equilibrium than if firms act collusively as a cartel (monopoly). The cartel output is 360 and the cartel price is 64¢ (Table 6.1). The Cournot industry's output (480) is a third larger, and the price (52¢) is 19 percent lower than in the cartel equilibrium.

The predictions of the cooperative model do not differ in a statistically significant way from the observed values. A statistical test shows that the probability of the selfish (Nash) model is less than 0.005: It is virtually impossible that the observed scanning rates are the outcome of selfish behavior.

The scientists who conducted this study were surprised that birds behaved cooperatively. After all, a "selfish" bird—one that did not cooperate by sharing scanning duty—should have had an advantage over a "cooperative" bird. Thus, they expected selfishness to be an "evolutionary stable strategy."

After a little more thought, however, they concluded that the results made sense when the same players in a "game" meet again and again: "The only strategy evolutionarily stable to invasion may be to reciprocate in kind." That is, it is optimal to cooperate only as long as other birds cooperate. They called a bird that follows this conditional cooperative strategy a *judge.* A judge behaves like a cooperative bird if others are cooperative and like a selfish bird if others are not cooperative. They calculated the flocking "payoff" matrix for a flock of two birds with these three types of strategies, where the payoff (the numbers in the following table) is the probability that Bird 1 survives predator attacks for a day under adverse conditions.

		Bird 2		
		Cooperative	Selfish	Judge
Bird 1	*Cooperative*	0.513	0.492	0.513
	Selfish	0.528	0.503	0.503
	Judge	0.513	0.503	0.513

Notice that a judge does as well as a cooperative bird with another cooperative bird and better than a cooperative bird when paired with a selfish bird. As long as the birds are all cooperative or judges, we would expect to see cooperative behavior.

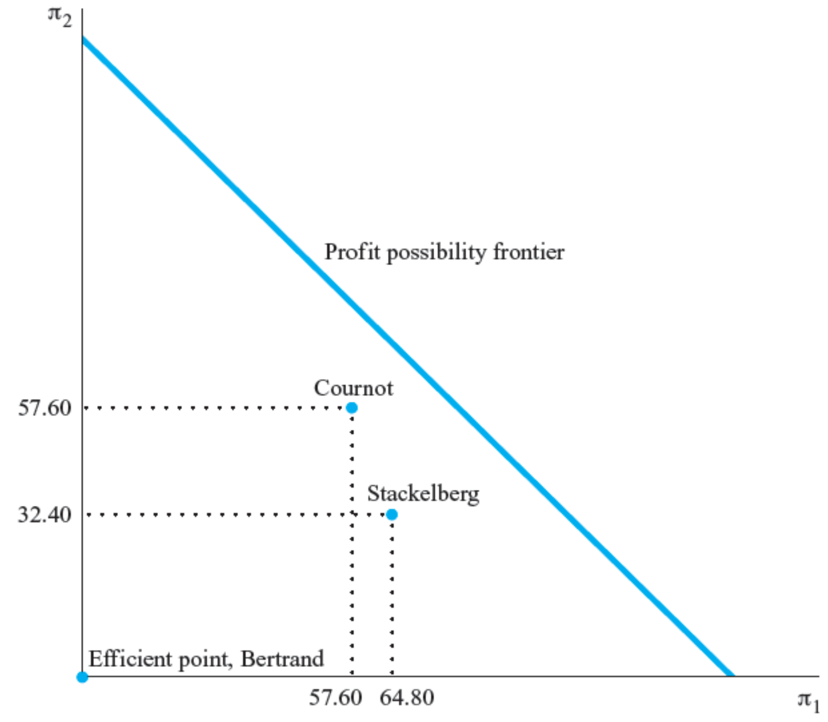
Consumers benefit from lower prices. The consumer surplus under the cartel is 64.8; the Cournot consumer surplus is 115.2 (Table 6.1). Thus, consumer surplus falls 44 percent if the Cournot firms form a cartel.

Lerner's price-cost margin, $(p - MC)/p$, is lower for a Cournot oligopoly than for a cartel. The cartel's price-cost margin is 56 percent, whereas the Cournot industry's margin is 46 percent—only 82 percent as large as the cartel's.

The Cournot firms have an incentive to form a cartel. The profit of Cournot Firm i is $(p - AC_i)q_i = (52¢ - 28¢)240 = \$57.60$. The sum of the profits of the Cournot firms, $\pi_1 + \pi_2$, is \$115.20, but the cartel's combined profits are \$129.60 (Table 6.1). Thus, if the firms form a cartel, their profits rise by 12.5 percent.

The maximum combined profits that two collusive firms can earn is \$129.60. There are many ways to divide these monopoly profits: Firm 1 could earn \$0 and Firm

FIGURE 6.3 The Profit Possibility Frontier



2 could earn \$129.60; each could earn half, \$64.80; Firm 1 could earn \$129.60 and Firm 2 could earn \$0; and so forth for any combination in which the sum of profits is \$129.60. The *profit possibility frontier,* $\pi_1 + \pi_2 = \$129.60$, in Figure 6.3 shows the highest profit one firm could earn, holding the profit of the other firm constant.[7]

Figure 6.3 also shows the Cournot equilibrium profit level, where each firm earns \$57.60, so $\pi_1 + \pi_2 = \$115.20$. The Cournot equilibrium lies well inside the profit possibility frontier, giving the firms an incentive to collude and increase their profits to the levels on the profit possibility frontier.

Comparison of the Cournot Equilibrium and the Social Optimum. How does the Cournot equilibrium compare to the social optimum where price equals marginal

[7] A profit possibility frontier is derived by holding Firm 1's profit constant at some level and then maximizing the profit of Firm 2 with respect to q_1 and q_2. In our example, the solution is

$$(q_1 - 1{,}080)q_1 + (q_2 - 1{,}080)q_2 - 2q_1q_2 + 259{,}200 = 0.$$

For any particular q_1, this equation simplifies into a quadratic in q_2. The relevant root is the smaller one. See Friedman (1983:22–7).

EXAMPLE 6.2

Oligopoly Welfare Losses

By exercising market power, oligopoly firms create deadweight loss, *DWL*. Bhuyan (2000) estimated the standard *DWL* triangle for 35 U.S. food manufacturing industries.* For all food industries, the *DWL* as a percentage of total sales was 5.5%. The percentage loss was as high as 33.4% for cereal preparation; 31.8% for soybean oil; 26.2% in flour and grain products; 17.5% in creamery butter; 10.4% for pickles, sauces, and salads; 10.2% in bottled and canned soft drinks; and 7.2% for malt beverages. The percentage losses were less than half a percent for canned specialties; dehydrated fruits, vegetables, and soup; rice milling; pet food; candy and confections; salted and roasted nuts and seeds; cottonseed oil; fresh or frozen seafood; and macaroni and spaghetti.

* Bhuyan (2000) also reports other measures of social loss. The higher profits from differentiated product oligopoly may induce firms to introduce new products that consumers value (as we discuss in the next chapter) and that may offset the loss from high prices.

cost (as in the competitive equilibrium)?[8] If both firms act as price takers at a price equal to the marginal cost of 28¢, they make zero profits per melon, so they are indifferent as to how many melons they produce. At a price of 28¢, 720 melons are demanded by the market (as determined by the intersection of the *MC* and market demand curves in Figure 6.1). If the firms split the total sales, each firm produces 360 melons. Consumer surplus is 259.2 (Table 6.1).

Thus, in our linear example, at the social optimum, twice as much output is produced as by a cartel and one-and-a-half times as much as by a Cournot duopoly. The competitive price is only 44 percent of the monopoly price and 54 percent of the Cournot price. At the social optimum output, consumer surplus is four times greater than under a cartel and two-and-a-quarter times greater than under a Cournot duopoly. The Cournot duopoly equilibrium, then, lies between the competitive and monopolistic equilibria. In linear examples such as this one, it is closer to the monopolistic equilibrium. See Example 6.2 for estimates of the welfare losses due to oligopoly in actual markets.

Three or More Cournot Firms. If there are n (≥ 2) identical Cournot firms, the same type of analysis can be used to derive the Cournot equilibrium, as Appendix 6A

[8] We compare Cournot to the social optimum rather than the competitive equilibrium because the competitive equilibrium requires a potentially unlimited number of firms, and there are only two firms in this market. In the socially optimal equilibrium, firms maximize profits subject to the constraint that price equals marginal cost (as in a competitive equilibrium). This equilibrium concept was introduced by Shubik (1959), who called it the *efficient point*. See Shubik with Levitan (1980) and Friedman (1983) for a discussion of the fixed costs consistent with the *efficient point*. Henceforth, we use the terms *social optimum* and *competitive equilibrium* interchangeably.

TABLE 6.2 Cournot Equilibrium with Few and Many Firms

			Firm		Industry	
	Number of Firms	Price (¢)	Output	Profit ($)	Output	Profits ($)
Monopoly	1	64	360	129.60	360	129.60
	2	52	240	57.60	480	115.20
	3	46	180	32.40	540	97.20
	4	42.4	144	20.74	576	82.94
	5	40	120	14.40	600	72.00
	6	38.3	102.9	10.58	617.1	63.48
	7	37	90	8.10	630	56.70
	8	36	80	6.40	640	51.20
	9	35.2	72	5.18	648	46.66
	10	34.5	65.5	4.28	654.5	42.84
	15	32.5	48	2.30	675	32.26
	20	31.4	34.3	1.18	685.7	23.51
	50	29.4	14.1	0.20	705.9	9.97
	100	28.7	7.1	0.05	712.9	5.08
	500	28.1	1.4	0.002	718.6	1.03
	1000	28.1	0.7	0.001	719.3	0.52
Competition	∞	28	~0	0.00	720	0.00

shows. Firm 1's best-response function is $q_1 = R_1(q_2, \ldots, q_n)$. If the other $n - 1$ firms produce an identical amount of output, q, then Firm 1's best-response function is $q_1 = R_1(q_2, \ldots, q_n) = 360 - q(n - 1)/2$. The other firms have similar best-response functions. As a result, the Cournot equilibrium quantity is $q = 720/(n + 1)$, and the equilibrium price is $p = (1 + 0.28n)/(n + 1)$.

Table 6.2 shows that the larger is n, the smaller is output per firm, whereas the larger is industry output, and the lower is price. The effect of additional rivals on quantity and price is initially very strong, but tapers off as the number of firms increases. If there are only 2 firms, the price is 86 percent above the competitive price. However with 10 firms, the price is only 23 percent above the competitive price, and with 50 firms it is only 5 percent above the competitive price. If the number of firms is extremely large, the output per firm, industry price, and industry output approach the socially optimal levels. Consumers are better off (lower prices, higher consumer surplus) and firms are worse off (lower profits) as the number of firms increases.[9]

In summary, the Cournot model includes monopoly and competition as extreme cases, and the Cournot equilibrium approaches the competitive one as the number of firms increases. See Example 6.3 on mergers.

[9] Ruffin (1971) discusses the conditions that must hold for the Cournot equilibrium price to converge to the competitive price as the number of firms grows large.

EXAMPLE 6.3

Mergers in a Cournot Economy

If all Cournot firms join together and behave as a monopoly, collective profits increase. Suppose only some of the Cournot firms merge (or coordinate actions as a cartel). The Cournot model has disturbing implications about the oligopolists' profits (given linear demand functions and identical constant marginal and average costs):

- In an industry with at least three firms (before mergers), if only two firms merge, their collective profits will fall.
- A merger of a larger number of firms *may* increase the size of the collective losses due to merger.
- For any given number of (premerger) firms, if a merger of firms causes collective losses, a merger by a smaller number of firms also causes losses. Similarly, if a merger of *k* firms causes gains, a merger by a larger number of firms also causes gains.
- If less than 80 percent of the firms merge, mergers will be collectively unprofitable.
- If any given share (less than 100 percent) of the firms in an industry merge, there is an initial industry size (number of firms) such that the merger causes losses.

These results imply that if the equilibrium is determined according to the Cournot model and firms maintain Cournot beliefs, then they will merge or form a cartel only if virtually all other firms in the industry join them. Alternatively, firms merge only if the firms do not maintain Cournot beliefs or the merger generates efficiencies.

Sources: Salant, Switzer, and Reynolds (1983) and Patinkin (1947). Compare Aumann (1973), Okuno, Postlewaite, and Roberts (1980), Farrell and Shapiro (1990), McAfee and Williams (1992), and Rothschild, Heywood, and Monaco (2000).

The Bertrand Model

Cournot's work was well ahead of its time. The first major challenge to his book came in 1883, 45 years after it was published. In this critique, Joseph Bertrand argued that it is hard to see who sets prices in oligopolistic markets if the firms do not set them. Cournot, by having firms choose output rather than price, fails to state explicitly the mechanism by which prices are determined (but, for that matter, so does a competitive model).

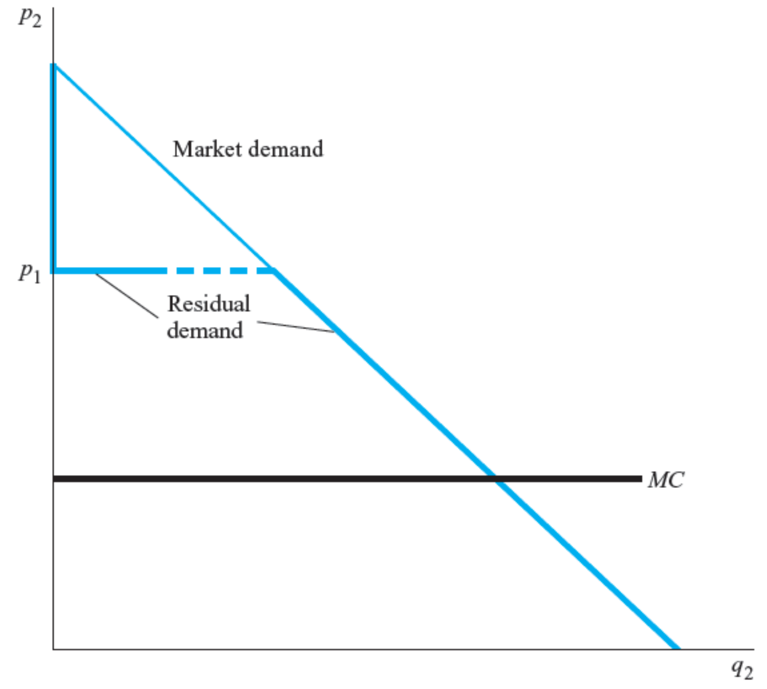
In Bertrand's model, firms set prices rather than output. If consumers have complete information and realize that firms produce identical products, they buy the one with the lowest price. In a Bertrand model, each firm believes its rival's price is fixed;

by a slight price cut, the firm is able to capture all its rival's business. In the Bertrand equilibrium discussed below, firms make zero profits and no firm can increase its profits by raising or lowering its price, which, when it exists, is equivalent to the social optimum (competitive equilibrium) discussed above.

An Example. To illustrate the Bertrand equilibrium, let us make the same assumptions as in the Cournot example: no entry, homogeneous products, single period, the same demand curve, Equation 6.1 (which we can rewrite as $p = 1 - 0.001Q$), and the same constant marginal cost of 28¢. The only important change is that firms now set prices rather than quantities. Each firm is willing to sell as much quantity as is demanded at the price it sets.

Suppose that Firm 1 charges a price p_1, which is greater than its marginal cost of 28¢. If Firm 1 makes any sales at all, it earns a positive profit. Because both firms produce identical products, however, all consumers buy from Firm 2 if p_2 is even slightly below p_1; none buy from Firm 2 if p_2 is above p_1; and consumers are indifferent between the two firms when $p_2 = p_1$. Thus, as Figure 6.4 shows, the residual demand curve (thick blue line) facing Firm 2 is zero when p_2 is above p_1, equals the market demand when p_2 is below p_1, and is horizontal at p_1. If both firms charge the same price,

FIGURE 6.4 A Bertrand Firm's Residual Demand Curve

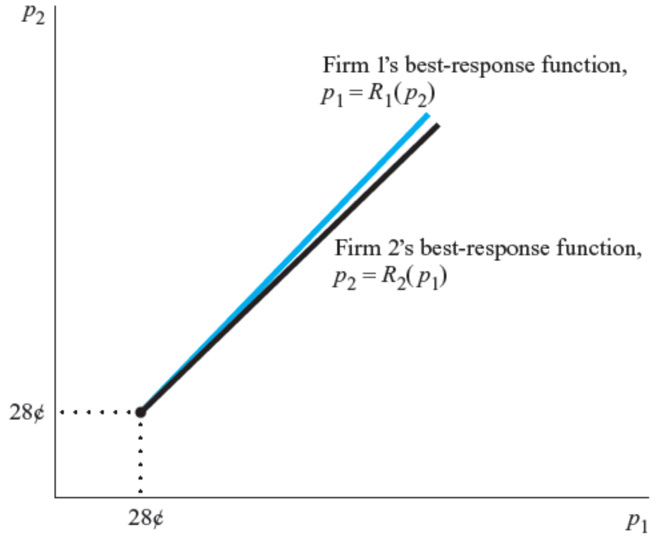


we assume that they split the total market demand. In Figure 6.4, where the demand facing Firm 1 is horizontal (at $p_2 = p_1$), half the horizontal line is dashed to indicate that Firm 1 sells only half the total amount demanded.

When both firms charge 28¢, neither firm profits by changing its price. If a firm lowers its price, it loses money (because price is then below marginal and average cost). If either firm raises its price, it makes no sales at all.

The only possible *Bertrand equilibrium* or *Nash-in-prices equilibrium* is $p = MC =$ 28¢.[10] Figure 6.5 illustrates this result using best-response functions in price space (the firms' prices are on the axes). Given whatever price, p_1, Firm 2 believes that Firm 1 will set, Firm 2 wants to set a price, p_2, that is slightly below p_1, as long as p_2 is greater than 28¢. That is, Firm 2's best-response function lies slightly below the 45° line (where the two prices are identical) through the point (28¢, 28¢). If Firm 1 sets p_1 below 28¢, Firm 2 does not respond because it cannot make a profit at any price. Similarly, Firm 1's best-response function lies slightly above the 45° line and above 28¢. The only intersection of these best-response functions (and hence the only equilibrium) is where price equals marginal cost.

FIGURE 6.5 Bertrand Best-Response Functions



[10] This equilibrium can be obtained from other game-theoretic models. For example, Grossman (1981a) shows that if a firm's strategy is a supply function (the quantity supplied as a function of price rather than just the price itself), under weak conditions the competitive, or Bertrand, equilibrium is a Nash equilibrium. This outcome is the result of the threat of entry, which we consider in Chapter 11.

If both firms charge a price equal to marginal cost, they earn zero profits. Thus, the Bertrand equilibrium for homogenous goods is the same as the social optimum (competitive equilibrium), as shown in Figure 6.3. Consumers prefer the Bertrand equilibrium to the Cournot or cartel equilibria.

A Comparison of the Bertrand and Cournot Equilibria. In the absence of an auctioneer, it is difficult to imagine how prices are determined if firms set output (Cournot) rather than prices (Bertrand). As a result, some economists find Bertrand's model more attractive than Cournot's, because it explains how prices are set.[11]

Because price rather than output is the decision variable, the Bertrand firm's residual demand curve differs substantially from that of a Cournot firm. When goods are homogeneous and all firms charge the same price, a Bertrand firm's residual demand curve is kinked (Figure 6.4). By slightly lowering its price, a firm may increase sales from none of the market to all of the market. Such sudden shifts in sales are rarely observed in most industries. Rather, demand curves facing individual firms appear to be smooth (nonkinked), as in the Cournot model, so that each firm's output shifts slightly for small price changes. Thus, Bertrand's model of a homogeneous good may be more realistic in explaining who sets prices, but the nonkinked Cournot demand curve facing individual firms is more realistic.

The Cournot equilibrium is intuitively appealing: With a small number of firms, output and price lie between the competitive and monopolistic equilibria. In contrast, the Bertrand equilibrium is counterintuitive: So long as there are at least two firms, the Bertrand price is the competitive price (marginal cost).

This last result, however, depends on a number of strong assumptions: The output is homogeneous, the market lasts for only one period, and any firm can produce as much as it wants at constant marginal cost. If any of these assumptions is relaxed, the Bertrand price does not equal marginal cost. The next chapter shows that if firms differentiate their products, the Bertrand price is above marginal cost. Later in this chapter, we show that, if markets last for many periods, the equilibrium price is likely to be closer to the monopoly price (even if firms set prices rather than quantities). The next section shows that a price equal to marginal cost is not a Bertrand equilibrium if firms have limited production capacity.

Capacity Constraints in Bertrand's Model: Edgeworth's Model. In 1897, Francis Edgeworth showed that, if firms have limited capacity to produce, there is no single-price, static Bertrand equilibrium. To illustrate Edgeworth's point, suppose the previous Bertrand example is modified so that each firm's maximum output capacity is 360, which is half the amount demanded at a price equal to marginal cost. That is, each firm's

[11] A firm that must sell its product immediately and cannot store it, as in the melon example, must adjust its price rapidly, as necessary, or it will be stuck with useless output. On the other hand, a firm that cannot change prices quickly or can do so only at great cost—say, because it prints elaborate catalogs—may meet fluctuations in demand by varying output (see Chapter 17).

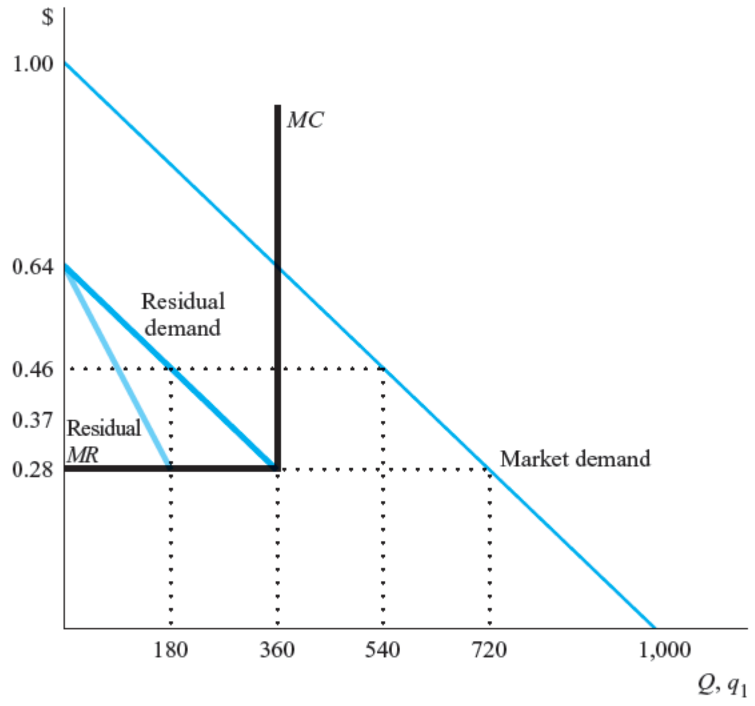
average and marginal cost curves are horizontal at 28¢ up to 360 units, and then the average and marginal cost curves are vertical (the cost of the next unit of output is infinite).

With limited capacities, is the original Bertrand equilibrium ($p_1 = p_2 = 28$¢, $Q = 720$) still an equilibrium? That solution is feasible given our assumptions, because, at those prices, the firms' combined output can just satisfy the market's demand of 720 units. That solution, however, is not an equilibrium.

For it to be an equilibrium, neither firm should want to change its behavior. At that proposed equilibrium, however, each firm wants to raise its price. In particular, suppose that Firm 1 believed that Firm 2 will charge $p_2 = 28$¢. What price should it set to maximize its profit?

As before, Firm 1 does not want to lower its price because it would suffer a loss if it charged less than marginal cost. If Firm 1 raises its price, all consumers want to buy from the second firm. Half the market, however, is unable to buy at that price because of the second firm's limited capacity. The first firm faces a positive residual demand from frustrated consumers who are unable to buy from Firm 2, as shown in Figure 6.6. The residual demand facing Firm 1 is the market demand minus the 360 units sold by Firm 2 (where only the portion above its marginal costs is of interest to Firm 1).

FIGURE 6.6 Bertrand Residual Demand When Firms Have Limited Capacity



Firm 1 can maximize its profits by acting like a monopoly with respect to its residual demand. Its marginal revenue equals its marginal cost at a price of 46¢, and it makes positive profits (whereas Firm 2 makes no profits on its sales). Thus, the original Bertrand equilibrium is *not* an equilibrium if firms have limited capacity.

Is there an equilibrium price? Suppose that Firm 1 sets a price of 46¢. If Firm 2 sets a price slightly below 46¢, all consumers want to buy from it. Given its limited capacity, however, Firm 2 meets only two-thirds of the market demand. Firm 2 sells twice as much as Firm 1 at almost the same price, so its profits are double those of Firm 1.

By similar reasoning, if Firm 1 sets any price below 46¢ and above 37¢, Firm 2 will want to set a slightly lower one. If, however, Firm 1 sets a price at (or below) 37¢, Firm 2 would earn more by charging 46¢.[12] Thus, there is no single-price, static equilibrium.[13] See Example 6.4.

More generally, one can show that there is no static equilibrium if firms have extremely limited capacity, or, equivalently, their average costs rise rapidly at some relatively low output level (Shubik with Levitan 1980). If any firm can meet the entire market's demand, however, an equilibrium exists that is identical to the efficient solution.

The Stackelberg Leader-Follower Model

In many ways the saying "Know thyself" is not well said. It were more practical to say "Know other people." —*Menander*

Heinrich von Stackelberg (1952) presented the third important oligopoly model in 1934. In the Stackelberg model, firms set output, and one firm acts before the others.

The *leader* firm picks its output level and then the other firms are free to choose their optimal quantities given their knowledge of the leader's output. In some industries, historical, institutional, or legal factors determine which firm is the first mover. For example, the firm that discovers and develops a new product has a natural first-mover advantage.

[12] Each firm's owner makes the following calculation: If I drop my price to p, which is slightly below my rival's price, I can sell my maximum output, 360 melons. On the other hand, if I raise my price to 46¢, I can only sell 180 melons, but I make more per melon. At what price, p, are my profits the same as if I set my price at 46¢? To answer that question, I equate my profits at 46¢ to those at p and solve for p: (46¢ − 28¢)180 = (p − 28¢)360. That is, my profits are equal if I raise my price to 46¢ or lower it to 37¢.

[13] Technically, Edgeworth showed that there is no equilibrium in "pure strategies." There is no simple rule of the sort discussed in the Cournot and Bertrand models for firms to follow at all times that leads to equilibrium. Kreps and Scheinkman (1983), Dasgupta and Maskin (1986), and others show that mixed-strategy (see Appendix 6B) equilibria exist. Kreps and Scheinkman (1983) and Davidson and Deneckere (1986) also show that, if firms play a two-stage game in which they choose their capacity levels and then price according to Bertrand, the Bertrand equilibrium is the same as the Cournot equilibrium under certain circumstances. See also Allen and Hellwig (1986). Maggi (1996) presents a more realistic and elegant two-stage game that has the property that a solution in pure strategies always exists and varies, depending on the circumstances, between Bertrand and Cournot.

EXAMPLE 6.4 *Roller Coaster Gasoline Pricing*

In dynamic games, many outcomes are possible, including an "Edgeworth cycle" in which prices rise, then fall, and then rise again. Using the theoretical work of Maskin and Tirole (1988b), Noel (2001) examined pricing in Canadian retail gasoline markets. Noel discovered three patterns of pricing, the most prevalent being an Edgeworth cycle.

In about 40 percent of the cities he examined, Noel observed a pricing pattern in which the retail price spikes quickly above the wholesale (called the "rack") price and then slowly drifts down for the next 2–3 weeks; then, just before it reaches the wholesale price, the retail price quickly spikes up again. The steepness of the upward spike depends on the prevalence of lots of small firms.

The second most prevalent pattern is sticky prices, where retail prices change infrequently (every two months) even though the underlying wholesale price changes. This pattern was observed in markets with only a few retail firms. The third pattern is "normal" pricing where the retail price tracks movements in the wholesale price.

The average difference between the retail price and the wholesale price varies with the retail pricing pattern. The average difference was highest for normal pricing, next highest for sticky pricing, and lowest for Edgeworth pricing.

Source: Noel (2001).

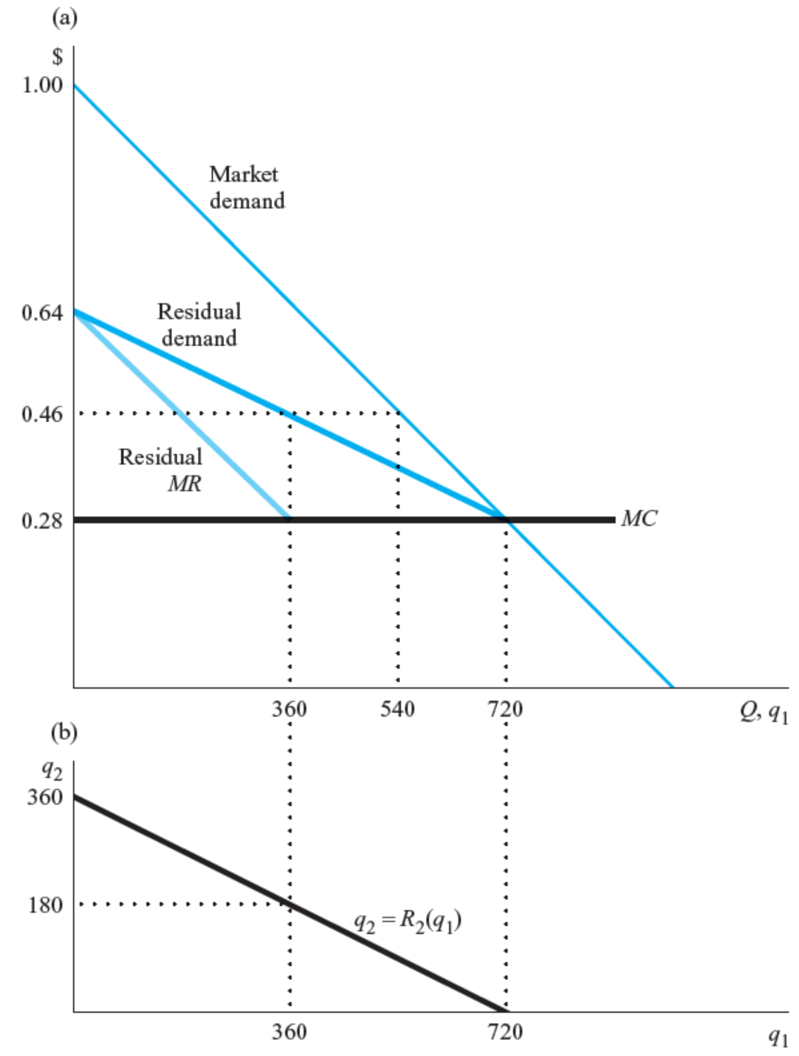
An Example. Suppose one of the melon-producing firms from before is a follower firm (Firm 2) and the other is the leader (Firm 1). Firm 1 realizes that once it sets its output, q_1, the follower firm will use its Cournot best-response function to pick its optimal $q_2 = R_2(q_1)$.

The leader, therefore, picks q_1 to maximize its profit subject to the constraint that the follower firm chooses its corresponding output using its Cournot best-response function. In the Stackelberg equilibrium, the leader is better off and the follower is worse off than in a Cournot equilibrium. In short, *knowing how its rival will behave allows a leader to profit at the follower's expense.*

Because the firms have identical costs, Firm 1 knows the Cournot best-response function of Firm 2, $R_2(q_1)$, which Figure 6.7b shows. Consequently, the leader knows how much the follower will produce at any level of output the leader chooses. Thus, the leader can calculate the *total* production corresponding to any output level it chooses, and it chooses the level that maximizes its profits.

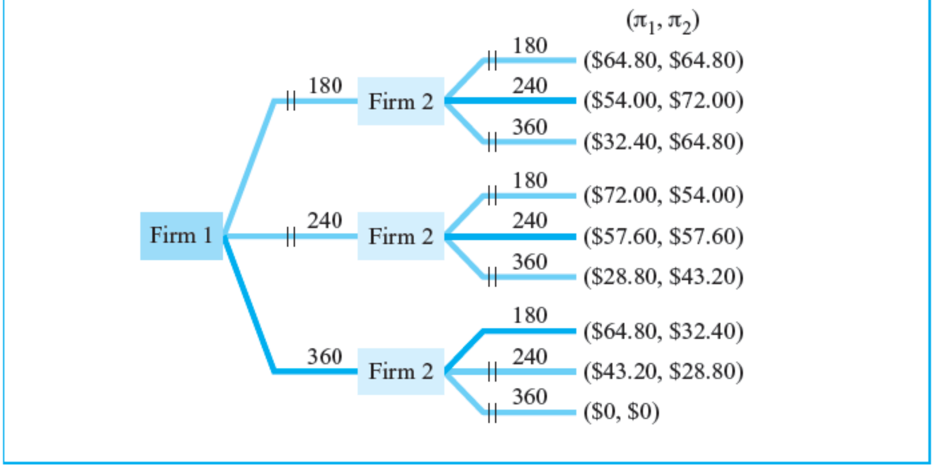
By subtracting the follower's output (as summarized by the follower's best-response function in Figure 6.7b) from total demand, the leader calculates its residual demand curve (Figure 6.7a). The leader picks its output, q_1, where its marginal revenue based on its residual demand curve equals its marginal cost. Firm 1 maximizes its profits by producing 360 melons (Figure 6.7a and Table 6.1). Firm 2 produces only 180 melons, which is determined by substituting 360 into Firm 2's best-response function (Figure 6.7b).

FIGURE 6.7 The Stackelberg Equilibrium



The Stackelberg game can be analyzed using the extensive-form representation of the game (or *decision tree*), which shows the order in which firms make their moves, each firm's strategy at the time of its move, and the payoffs. There are an infinite number of combinations of outputs the two firms can produce. Figure 6.8 shows only some of the output levels they can choose: the Stackelberg follower quantity (180), the Cournot quantity (240), and the Stackelberg leader quantity (360).

FIGURE 6.8 Extensive-Form Representation of Stackelberg Game



Each line represents an action, and each box is a point of decision by one of the players. Starting from the left of the diagram, Firm 1 picks its output level, then Firm 2 picks its output level, and the payoffs (Firm 1's profit first) are shown on the right. If Firm 1 chooses the Cournot quantity, 240, Firm 2's profit is maximized at $57.60 at the Cournot quantity, 240. A pair of straight lines are drawn through each of Firm 2's other two action lines, showing that Firm 2 does not want to choose those actions.

If Firm 1 chooses the Stackelberg leader quantity, 360, Firm 2's profit is maximized at $32.40 at the Stackelberg follower quantity, 180. Similarly, if Firm 1 produces 180 melons, of the three choices it has, Firm 2 produces 240.[14] Given the way Firm 2 will respond, Firm 1 realizes that its expected profit will be highest if it produces 360 melons, where it earns $64.80.

Stackelberg Equilibrium Compared to Other Equilibria. The Stackelberg leader produces more output (360 melons) and the follower less output (180) than would a Cournot firm (240), as Table 6.1 shows.[15] Total Stackelberg output (540 melons) is greater than the Cournot output (480), but less than the social optimum (competitive equilibrium) output (720). The Stackelberg price, 46¢, is higher than the competitive price, 28¢, but lower than the Cournot price, 52¢. As a result, consumer surplus is

[14] Given Firm 1 produces 180 units, if Firm 2 can pick any quantity, it produces 270 for a profit of $72.90.

[15] The Stackelberg equilibrium differs from the Cournot equilibrium because the leader acts first and the follower knows for certain its rival's output. That is, the leader's claim that it will produce a large quantity is *credible* because it has already done so. If the two firms were to act simultaneously (the Cournot game), one firm's claim that it will produce a large quantity may not be viewed as credible by a rival firm.

higher with a Stackelberg duopoly, 145.8, than with a Cournot duopoly, 115.2, but lower than the social optimum, 259.2.

The Stackelberg equilibrium lies inside the profit possibility frontier (Figure 6.3). The leader, Firm 1, makes a profit of $64.80 and the follower, Firm 2, only makes half that, or $32.40. Thus, total industry profits ($97.20) are less than the combined profits in the Cournot ($115.20) or collusive ($129.60) equilibria.

A Comparison of the Major Oligopoly Models

The three major noncooperative oligopoly models make different assumptions about whether firms choose output or price and whether the firms choose simultaneously or sequentially. As a result, they predict very different firm and industry outputs, prices, profits, and consumer surpluses in equilibrium (Table 6.1).

If there is only one firm, all three models predict monopoly behavior. The more firms, *n,* in the industry, the closer the Cournot (Tables 6.1 and 6.2) and Stackelberg (Table 6.1) equilibria to the social optimum.

However, the Bertrand equilibrium with homogeneous goods is unaffected by the number of firms in the industry. As long as the market has at least two firms with unlimited capacity, the Bertrand oligopoly equilibrium is the same as the social optimum.[16]

Multiperiod Games

The most important recent development in game theory is the analysis of repeated, or multiperiod, games. This analysis shows that Stigler's (1964a) cartel theory analysis of oligopoly (Chapter 5) is closely related to multiperiod oligopoly models based on game theory.

In a multiperiod game, firms may use complex strategies in which they change behavior in one period depending on the outcome in previous periods. Repeated games where players know their rivals' previous actions and condition their actions in this period on those previous actions are often referred to as **supergames**.

The chief advantage of a multiperiod model is that it allows for more complex and realistic interactions between firms than a single-period model. For example, a firm can signal to another firm that it wants to avoid vigorous competition by lowering its output for a few periods. If the other firm responds by lowering its output, both firms can charge a higher price. If either firm increases its output, the other can retaliate for a while by raising its output (and lowering price) to punish the transgressor. Because of this ability to send signals and punish in multiperiod markets, firms that would produce at the Cournot-Nash level in a single-period model may further restrict output and make larger profits in a multiperiod model.

[16] However, with heterogeneous goods (discussed in the next chapter), the Bertrand equilibrium differs from the competitive equilibrium, and the number of firms in the industry affects prices.

This result can be illustrated using a specific game, known as the prisoners' dilemma game, that is repeated an infinite number of times. The results of repeating the game only a finite number of times are also considered, and some other recent work on multiperiod games is discussed.

Single-Period Prisoners' Dilemma Game

The game is up. —*Shakespeare*

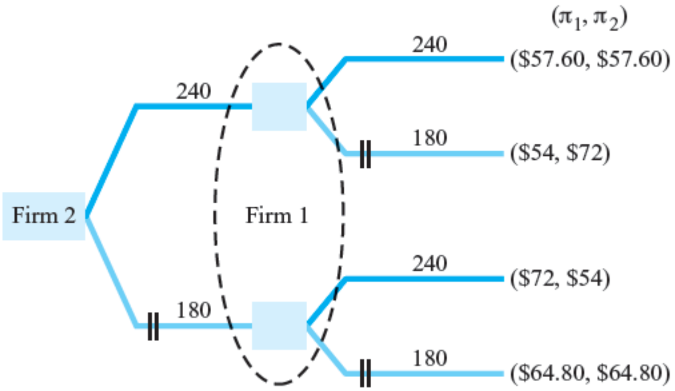
Suppose in our Cournot example that firms were restricted to choose one of only two possible output levels: The firms can only produce the cartel output level (each produces 180 units) or the Cournot level of output (each produces 240). The two firms must act simultaneously. Their actions and their payoffs, which depend on the strategies both choose, are summarized in Figure 6.9a. The first firm's profit is shown in the upper right and the second firm's profit is shown in the lower left of each cell. If both firms choose to produce 240 units, each earns a profit of $57.60; if both produce 180 units, each earns a profit of $64.80. If the first firm produces 240 units and

FIGURE 6.9 Prisoners' Dilemma Game

(a)

		Firm 1	
		240	180
Firm 2	240	$57.60 $57.60	$54.00 $72.00
	180	$72.00 $54.00	$64.80 $64.80

(b)



the second firm produces 180, however, the first firm earns $72, and the other only earns $54. A **normal-form representation** (or *strategic form*) of a game is a matrix, such as in Figure 6.9, that shows all the strategies available to each player (who must choose actions simultaneously) and the payoffs to each player for each combination of strategies.

Each firm must choose its action or strategy without knowing what the other firm will do. That is, the firms are engaged in a **game of imperfect information**, in which a firm must choose an action without observing the simultaneous (or earlier) move of its rival.

Figure 6.9b shows the options facing Firm 1. It is an extensive-form representation of this game. In this particular game tree, Firm 2 does not literally move before Firm 1: They move at the same time. As a result, Firm 1 is uncertain about Firm 2's (simultaneous) move. The dotted ellipse around Firm 1's two decision nodes (junction points or boxes in the game tree) indicates that Firm 1 cannot distinguish between the two nodes at the time it must decide on its strategy; that is, Firm 1 does not know which strategy Firm 2 will use. In the figure, Firm 1's payoff is listed first.

How should Firm 1 choose its strategy? The firm should reject any strategy that is *strictly dominated* by any other strategy. One strategy strictly dominates another if it produces as high or higher payoff than the other regardless of the action chosen by a rival firm. If one strategy dominates all other strategies regardless of the actions chosen by rival firms, the firm should choose this **dominant strategy**.

Although not all games have dominant strategies, this game does (see Appendix 6B for a game where players use random strategies). What strategy should Firm 1 choose? To answer that question, the manager of Firm 1 could use the following reasoning:

- If *Firm 2 chooses the high-output strategy* (240), and I choose it also, my profit is $57.60 (the first number in the payoff at the upper right of Figure 6.9b and the upper right number in the top left square in Figure 6.9a); if I use my low-output strategy (180), I only earn $54. I prefer $57.60 to $54, so *I'm better off with my high-output strategy.*
- If *Firm 2 chooses the low-output strategy* (180), then if I use my high-output strategy (240), my profit is $72; if I use my low-output strategy (180), my profit is only $64.80. Again, *I'm better off with my high-output strategy.*
- Therefore, *whichever strategy Firm 2 uses,* I'm better off using my high-output strategy. *The high-output strategy is a dominant strategy.*

The payoff table is symmetric, so the high-output strategy is also dominant for Firm 2. The double lines in Figure 6.9b drawn through each of the low-output action lines show that those actions are not chosen.

Both firms use the high-output strategy, and this strategy is a Nash equilibrium in strategies. Given the strategy of Firm 2, Firm 1 has no incentive to change its strategy and vice versa. Suppose in Figure 6.9a that Firm 2 produces 240 units. If Firm 1 changes from producing 240 units to 180 units, its profit falls from $57.60 to $54, so it does not want to change. Similarly, if Firm 1 produces 180 units, Firm 2 does not want to change its strategy. Because neither firm wants to change its strategy of producing 240 units if its rival also produces 240 units, both firms producing the large output is a Nash equilibrium.

This Nash equilibrium does not maximize the players' collective payoff. The two firms would be better off if they could cooperate and both use the high-price strategy. If both produce 180 units, combined profits are \$129.60; whereas, if both produce 240, combined profits are only \$115.20. If the game is only played once, its outcome is nonoptimal from the players' collective viewpoint. This game is called a **prisoners' dilemma** because both firms have dominant strategies that lead to a payoff that is inferior to what they could achieve if they cooperated.[17]

Infinitely Repeated Prisoners' Dilemma Game

If the single-period prisoners' dilemma game is repeated forever, the price in a given period is more likely to be higher than the price in a single-period game. For example, at a special event such as the Super Bowl, a souvenir firm may compete with others for a short period and then never see its rivals again. Such firms are relatively unlikely to succeed in charging a cartel price because each one knows it can cheat on any agreement with no fear of reprisal. In contrast, souvenir stands at popular tourist attractions that face each other over long periods are more likely to charge a relatively high price.

In the single-period prisoners' dilemma game, each firm took its rival's strategy as given and assumed it could not influence that strategy. If this game is repeated, however, each firm can influence its rival's behavior by *signaling* and *threatening to punish.* See Example 6.5.

Because both firms gain by a reducing output, they have an incentive to communicate to avoid the prisoners' dilemma problem, which stems from a lack of trust. Because antitrust laws make direct communications illegal, firms may try to communicate indirectly through their choice of strategy if (and only if) the game is repeated. For example, a firm can use a multiperiod strategy of setting a low quantity (or high price) and taking losses for several periods to signal its willingness to collude.

Similarly, a firm can threaten to punish its rival if it does not collude. (See Example 6.6.) To illustrate how penalties can be used to insure collusion, we use the quantity setting, single-period prisoners' dilemma game such as in Figure 6.9a. Each of the two firms in the industry can produce at different output levels in different periods. One possible strategy for a firm is to produce the Cournot-Nash level of output, q_n, each period ($q_n = 240$ in our example). If the other firm does the same, each earns the Cournot-Nash profits, π_n, each period (\$57.60). Alternatively, the firms can restrict output with each firm producing q_m, which is half the monopoly output, and earning a profit of π_m ($> \pi_n$) each period ($q_m = 180$ and $\pi_m = \$64.80$).

[17] Luce and Raiffa (1957) attribute the prisoners' dilemma game to A. W. Tucker. In the original version, two prisoners are accused of committing a crime. They are held in different rooms, so they are unable to communicate. Each prisoner has two choices of strategy: to talk or not to talk. If neither talks, each gets a one-year sentence on a minor charge. If both talks, each gets a five-year sentence. If one talks, but the other does not, the prisoner who confesses goes free, while the other gets a 10-year sentence. By the same reasoning as in the Figure 6.9a, both talk even though they are obviously better off if both keep quiet.

EXAMPLE 6.5

Copying Pricing

As every student knows, the way to learn something is to copy the relevant article and then absorb it through osmosis. As a result, copy shops often spring up near colleges. In the early 1970s, four firms in the Harvard Square area of Cambridge, Massachusetts, satisfied a large portion of the copy business of students from Harvard, MIT, Tufts, and other colleges in the area.

Initially, the smallest of the "big four" firms, Copy Cat Educational Services, Inc. (located in the J. August Clothing Store), charged much higher prices than its larger competitors. Then Jimmy Jacobs, the owner of the clothing store and copying service, lowered his prices to a level that the other firms contended was too low to make profits.

One competitor, Gnomon Copy, posted a sign in its window on the "Xerox Price Story." Gnomon charged that Jacobs "had sent word to the other Xerox services in Harvard Square that he was going to drive them all out of business . . . if they did not raise their prices to match his, which were then substantially higher than the going rate. . . . Now, Jacobs has carried out his threat."

Gnomon said it was meeting Copy Cat's price in Harvard Square in order not to lose customers, but that it would keep its prices at its other stores at their previous levels, which Gnomon considered "to be fair and reasonable." They urged customers to boycott Copy Cat, claiming, "You may pay a higher price today, but you will insure a viable competitive situation for the future."

Within hours of the posting of Gnomon's sign, according to Gnomon employees, Jacobs barged into their shop and said, "You call your boss and tell him he's got five minutes to take that down or I'll photograph it and use it in a libel suit." Upon reflection, Gnomon management decided it had not gone far enough and assigned an employee to hand out leaflets in front of Jacobs's copy center/clothing store. In turn, Jacobs filmed the leafleteers in the presence of a reporter, while a Gnomon salesperson was frantically searching for her camera to photograph Jacobs photographing everyone else. According to the reporter, "When the Gnomon salesperson ran out of notices, he began modeling his clothes. Mr. Jacobs laughed and kept filming."

The other firms did not lower their prices at first, and their business suffered. These other firms supported Gnomon's charge that Jacobs tried to get them to fix their prices at a higher level and had threatened to punish them by undercutting their prices if they did not cooperate. Jacobs said that he believed that all the firms had been charging too much and that he had decided to lower his prices, but only to a level where he could still make profits. He vigorously denied any attempt to fix prices.

Eventually prices in Harvard Square settled at a lower level, though cost changes may have been partially responsible. This example illustrates how firms try to influence their rivals' actions through their own price or output decisions and through communications when the game is played repeatedly.

Source: Vin McLellan, "Harvard Square: War of the Xerox Machines," *The Phoenix*, February 9, 1971.

EXAMPLE 6.6

Car Wars

In 1955, American passenger automobile production was 45 percent greater than it was in 1954 or 1956. Why?

Based on sophisticated econometric tests, Bresnahan (1987) contends that American automobile manufacturers' successful tacit collusion fell apart in 1955, but was reestablished in 1956. During the 1950s, major entry by foreign manufacturers had not yet occurred; thus, American manufacturers could collectively reduce output and increase price.

A casual perusal of automobile output for this period certainly indicates that 1955 was an unusual year. American automobile production from 1953 through 1959 was 6.13, 5.51, 7.94, 5.80, 6.12, 4.24, and 5.60 million cars. Thus, the 7.94 million cars produced in 1955 not only were substantially more than were produced in the adjacent years, 1954 and 1956, but also were substantially more than were produced in any year for the rest of that decade.

Not surprisingly, the large output in 1955 drove down the price of cars. Adjusting for quality, the price in 1955 was approximately 6 percent lower than in the adjacent years. The price fell by a smaller percentage than quantity increased, so that total expenditures rose. Automobile expenditures were (in billions of 1957 dollars) $13.9 in 1954, $18.4 in 1955, and $15.7 in 1956. In other words, consumers spent 32 percent more in 1955 than in 1954 and 17 percent more in 1955 than in 1956.

Firm 1 considers using the following two-part strategy:

- Firm 1 produces q_m output each period so long as Firm 2 does the same.
- If Firm 2 produces a different level of output in any period, t, then in period $t + 1$ and thereafter, Firm 1 produces q_n.

If Firm 2 believes that Firm 1 will follow this strategy, Firm 2 should produce q_m.[18] Firm 2 knows that it can make greater profits in period t by producing more than q_m in that period. If it does so, however, in the $t + 1$ period and every period thereafter, Firm 1 would produce q_n. As already demonstrated, when Firm 1 produces q_n, Firm 2 maximizes its profits by producing q_n also.

Thus, Firm 2 can earn unusually high profits in period t, but then it earns relatively low profits for the rest of the time. Unless Firm 2 puts very little value on future profits, it is in Firm 2's best interest to tacitly collude and produce q_m in each period.

In short, if the future matters significantly, the one-period gains from deviating from the monopoly output cannot compensate for the losses from getting π_n forever instead of π_m. Indeed, Firm 1 need not punish Firm 2 forever to induce it to cooperate; all it needs to do is produce q_n for enough periods so that it does not pay for Firm 2 ever to

[18] As in the single-period models, the question remains as to how firms form these beliefs. Unlike the single-period models, however, firms may form beliefs based on the history of the game in multi-period models.

deviate. Thus, because strategies can involve signals and threats of punishment, firms are more likely to charge the monopoly price in multiperiod than in single-period games.

Types of Equilibria in Multiperiod Games

All repeated games do not result in high prices, however. The type of equilibrium in a repeated game depends on a player's ability to effectively threaten other players who are not cooperative. The effectiveness of a threat depends on the interest rate, the length of the game, and the credibility of the threat.

Credibility. At the beginning of a game, each firm chooses a strategy to maximize its present discounted profits. If interest rates are so high that profits in future periods are worth substantially less than profits in the current period, future punishment is inconsequential and hence has no effect on current behavior.[19] Lower interest rates, therefore, make the threat of punishment more effective. The more periods left in the game, the larger the total punishment that can be inflicted on a transgressor, because the punishment can be inflicted for more periods. However, if the threat is not credible, in the sense that Firm 2 does not believe that Firm 1 will actually inflict the punishment in future periods, then Firm 2 ignores the threat altogether.

The importance of credibility is illustrated by a two-period prisoners' dilemma example (where, now, firms can choose any output level—not just two possible levels). Suppose Firm 1 is a cartel member and announces that it will produce the collusive quantity, q_m, in Period 1 and the Cournot-Nash quantity, q_n, in the second period if Firm 2 will produce the collusive quantity in the first period. Firm 1 also announces (or signals somehow) that if Firm 2 produces more than q_m in the first period, Firm 1 will punish it by producing a very large quantity (greater than q_n), in the second period. If Firm 2 believes that Firm 1 will carry out its threat, and the potential losses in the second period are large enough, it produces q_m in the first period.

Firm 2, however, does not view Firm 1's threat as credible. Suppose Firm 2 does not produce q_m in the first period. There is now only one period left in the game. Firm 1 *can* punish Firm 2 in Period 2 by producing more than q_n, thereby lowering Firm 2's profits below the Cournot-Nash level. But will Firm 1 do that? Probably not, because it is not in Firm 1's best interests to do so in Period 2. Firm 1 can only harm Firm 2 by harming itself and lowering *each* firm's profits below π_n. In the second period, however, Firm 1 does not benefit from doing so. It is too late to affect Firm 2's behavior in the first period, and there are no future periods. Indeed, in Period 2, Firm 1 should act as though it is participating in a one-period game and produce q_n. Thus, Firm 2 does not view Firm 1's threat as credible; to carry it out would be like locking the barn after the horse is stolen.

A monopoly price in the first period is possible, however, if Firm 1 can make its threat credible by precommitting itself to punish Firm 2 in the second period. Ignor-

[19] If the interest rate is 10 percent, $1 of profit in the first period is worth $1, but $1 of profit guaranteed in the second period is only worth 1/(1.1) ≈ 91¢ in the first period. The *present discounted* value of $1 of profits in both the first and second years is $1.91. With a high enough interest rate, profits in the future are essentially irrelevant to current decisions.

ing the legal issues, if Firm 1 writes a binding and enforceable contract in Period 1 that says it will forfeit an enormous sum of money if it does not punish Firm 2 in the second period if necessary, then its threat is credible.

Research in multiperiod games concentrates on equilibria that result from credible strategies and rules out other equilibria. Restrictions on possible equilibria are called **refinements**. One widely used refinement is to consider only **perfect Nash equilibria**: those Nash equilibria in which strategies (threats) are credible (Selten 1975). For example, in the two-period game in which Firm 1 threatens to punish Firm 2 in the second period if Firm 2 produces too much in the first period, the threat is credible only if the punishment is in Firm 1's best interest *in the second period.*

More generally, a strategy or threat is credible only if the firm will stick to that strategy in any **subgame**: a new game that starts in any period t and lasts to the end of the game. If the proposed strategies are best responses (Nash equilibria) in any subgame, these strategies are called a **subgame perfect Nash equilibrium** (or *perfect Nash equilibrium*).

One way to obtain a subgame perfect Nash equilibrium is to solve the game backward. We illustrate this technique for a two-period game. In the last period (the only interesting subgame), the strategy of each firm must be based on its one–period best-response function. That is, there is a Nash equilibrium in the second period in which the strategies are optimal in the sense that the players would have chosen them if the game were beginning in Period 2. In the second period, the Nash, or best-response, strategy for both firms is to produce q_n. Thus, Firm 1's only credible claim is that it will produce q_n in the second period. Because Firm 1's threat to punish in Period 2 is not credible, both firms also produce q_n in Period 1.

Now consider a game that lasts for a finite number of periods, T, greater than 2. To solve for a perfect Nash equilibrium requires working backward from the last period. In the last period, the firms produce q_n, by the preceding reasoning. Thus, Firm 1 cannot credibly threaten to punish Firm 2 for producing a large quantity in period $T - 1$. What happens in period $T - 1$? Effectively, it is now the last period. By the same reasoning, both firms produce q_n in that period. This reasoning can be repeated for $T - 2$, $T - 3$, and other earlier periods, with the conclusion that the firms produce q_n in each period. That is, the T-period game equilibrium simply repeats the single-period equilibrium T times (Selten 1978).

The intuition behind this argument is that firms cheat (produce more then q_m) in earlier periods because it is in their best interests to cheat in later periods—hence they do not have credible threats to produce q_m. As a result, any attempt to produce q_m in earlier periods unravels. The entire argument, then, depends crucially on the firms' cheating in the last period. The argument implicitly assumes that there is a *known,* fixed number of periods, T. All firms cheat in the last period, *if they know it is the last period.* If the period in which the game will end is not known until that period is over, a player is less likely to deviate from the cartel output level in that period. A game with a finite but unknown number of periods, so that players do not know which period is the final one, is therefore similar to a game with an infinite number of periods, and hence an enforceable cartel agreement is feasible.

To summarize, the subgame perfect Nash equilibrium depends on the number of periods in a multiperiod game and whether that number is known. First, we argued that producing the cartel output, q_m, each period is a subgame perfect Nash equilibrium in a game with an infinite number of periods. Then we showed that in a game with a known, finite number of periods, producing the Cournot-Nash output, q_n, in each period is a subgame perfect equilibrium. Finally, we contended that if the number of periods is finite but firms do not know which period is last until after it is over, then, again, the cartel equilibrium is a subgame perfect Nash equilibrium. Indeed, even without an explicit cartel agreement, the cartel equilibrium is a possible equilibrium as long as firms have the appropriate beliefs about each other's (credible) threats. These beliefs may be acquired by observing the history of a rival's behavior.

In a multiperiod game, a firm may have many strategies that involve different actions over time. In the previous example, Firm 1 could produce q_m in Period 1 and q_n thereafter if Firm 2 fails to produce q_m in Period 2. In general, Firm 1's output in any period can be a complicated function of its rival's output in previous periods. This multiplicity of strategies raises two problems for firms. First, it is unclear how firms know or form beliefs about their rivals' strategies if they are so complicated. That is why explicit communications in a cartel can be effective (see Farrell 1987). Second, with many possible strategies, there can be many possible Nash equilibria in strategies.

An infinite number of other subgame perfect Nash equilibria are possible in games with an infinite number of periods and little or no time discounting. The *folk theorem* (Friedman 1971, 1977; Fudenberg and Maskin 1986), which describes this set of subgame perfect Nash equilibria in infinitely long games, says, loosely, that any combination of output levels could be infinitely repeated so long as each firm's profits at those levels are at least as great as the minimum each firm could earn in a one-period game. As a result, in addition to the cartel solution, another perfect equilibrium in the infinitely repeated game is for each firm to produce the Cournot-Nash output, q_n, each period. Much current research is directed at further refining these results to provide better explanations of which equilibria occur. Without further refinements, almost any output level is a sustainable equilibrium, which makes this theory difficult to apply to actual industries.[20]

The folk theorem shows that in a dynamic setting almost any outcome can be sustained as an equilibrium as players follow strategies with punishment. Some of these strategies can be very simple. For example, in a dynamic "tit for tat" strategy, one player cooperates provided the other just did, but does not cooperate if the other player failed to cooperate on his last move. In experimental settings, this simple tit for tat strategy is frequently observed and leads to cooperation in the prisoners' dilemma game.

This chapter concentrated on games that assume no uncertainty about underlying economic conditions. Games in which firms are uncertain about rivals' actions or eco-

[20] However, with additional restrictions, one may be able to estimate a subgame perfect multiperiod model. For example, see Karp and Perloff (1989a, 1993a).

nomic conditions are even more complex.[21] Because such games have so many possible outcomes, economists studying them often place further restrictions or refinements on the possible equilibria to eliminate some possibilities.[22] Much of the current research in game theory is focused on games with uncertainty.

Experimental Evidence on Oligopoly Models

The various oligopoly models predict different equilibrium outcomes because they make alternative assumptions about how firms behave, the number of firms, the rules of the game (nature of the market), and the length of the game. Because all these models are logically consistent, one cannot choose between them on purely theoretical grounds. One can ask, however, whether their assumptions are reasonable or whether their predicted outcomes are consistent with actual market outcomes.

Chapter 8 discusses statistical studies of particular industries. Here, we discuss a number of experiments. Some economists conduct laboratory experiments to determine how college students behave under controlled conditions. Students play a game in which they set output or price for the firms each manages. Because the students keep their profits, they have an incentive to maximize profits in the experimental market. Postexperiment interviews indicate that a few students try to "win" the game by maximizing the difference in their profits relative to other players, rather than maximizing their own profits. The vast majority, however, do try to maximize their own earnings.

The experimental equilibria are compared to the various theoretically predicted equilibria. A survey of these experiments (Plott 1982, 1523) concludes that "three models do well in predicting market prices and quantity: the competitive equilibrium, the Cournot model, and the monopoly (joint maximization) model. Experiments help define the conditions under which each of these alternative models apply."

To give some idea of the results obtained, we discuss four representative, multiperiod game experiments. Virtually all simulation experiments use linear demand curves and a constant marginal cost.

[21] For example, in a multiperiod model in which Firm 1 does not know Firm 2's costs, Firm 1's beliefs about Firm 2's costs may affect how it behaves. Firm 1 may drop out of the market if it believes Firm 2's costs are much lower than its own. As a result, Firm 2 may attempt to convince Firm 1 that its costs are very low, perhaps by setting a very low price for several periods. Firm 1 uses the history of Firm 2's behavior in forming its beliefs about Firm 2's costs, taking into account Firm 2's attempts to mislead. Using the additional information about Firm 2's behavior that becomes available each successive period, Firm 1 updates its beliefs about Firm 2's costs. Firm 1 can combine the information about Firm 2's actual behavior with its prior beliefs to form a new estimate of the probability that Firm 2's cost is low by using Bayes's law from probability theory. An equilibrium in which firms form their beliefs using Bayes's law and in which each strategy is subgame perfect is called a *Bayesian perfect equilibrium.*

[22] For more detail on this and related topics, such as refinements and sequential equilibrium, see Harsanyi (1967–68), Kreps and Wilson (1982a, 1982b), Kreps and Spence (1984), Bernheim (1984), Pearce (1984), Tirole (1988), Shapiro (1989), Myerson (1991), Fudenberg and Tirole (1991), Binmore (1992), and Gibbons (1992). For theoretical and empirical research on dynamic oligopoly, see Maskin and Tirole (1988a, 1988b), McGuire and Pakes (1994), Ericson and Pakes (1995, 1998), and Fershtman and Pakes (2000).

Lave (1962), in work that started as a B.A. thesis at Reed College, conducted an experiment with undergraduates who participated in a repeated two-person, two-strategy, multiperiod prisoners' dilemma game. The players were placed so they could not see each other, making explicit communications impossible. Nonetheless, the vast majority of players were apparently able to communicate indirectly and thereby achieve the cartel solution. In various versions of the experiment, 75 to 100 percent of the outcomes were the cartel solution. As predicted theoretically, in the last period, when players knew the experiment was going to end, many (though not all) deviated from the cartel outcome because there could be no retaliation at that point.

Fouraker and Siegel (1963) conducted duopoly and triopoly (three-firm) experiments (compare Holt 1985). Each subject was given a payoff table showing that profits depended on a player's output choice and the output of the rival(s).

Each of 16 pairs of undergraduates played the game 25 times. Fouraker and Siegel used the players' decisions in the 21st period to evaluate the equilibrium. The duopoly outputs were distributed fairly uniformly over the range from slightly less than the collusive (cartel) level to the competitive (Bertrand) level. Five were closest to competition, 7 were closest to Cournot, 1 was between Cournot and cartel, and 3 were closest to the cartel equilibrium. The median output was the Cournot output.

In the triopoly game, however, the median output was only slightly below the competitive output. In 5 cases, the industry output was closest to Cournot, and in 6, it was closest to competition.

Fouraker and Siegel also conducted experiments in which subjects chose prices, as in the Bertrand game, instead of output levels. According to the payoff table, a player who chose a price above a rival's made no sales and suffered a small loss of profits.

When players had incomplete information (they knew whether their price was higher or lower than a rival's, but did not know the rival's profits), the price converged to the competitive equilibrium (or just above it) within 14 periods in 17 out of 18 cases. When duopoly players had full information (each knew all past prices and all players' profits), the results were more varied. In 6 cases, the market was at the competitive equilibrium by the 14th period, and in 3 more, the price was just above it. In 4 cases, the price was exactly midway between the competitive and the cartel price; in the remaining 4, it was at or adjacent to the cartel price.

In the triopoly case, whether players had incomplete or complete information, the market converged to the competitive level virtually every time. Thus, with full information, competitive behavior seems likely in three-person price games, but not in two-person ones. With incomplete information, competitive equilibrium is also more likely with three-person games.

One possible reason that competitive behavior was not observed in full-information duopoly games is that profits were near zero at the competitive level, so players had little to lose by choosing other strategies. Holt (1985) conducted a similar experiment of repeated duopoly games, in which the profits at the competitive or Bertrand equilibrium were positive.[23] He found that the outcomes were between cartel and Cournot, and closest to the Cournot outcome.

[23] This experiment differs from the Fouraker and Siegel experiment in a number of ways, most importantly in that it was constructed so that no possible strategy ensured a profit that always exceeded the positive competitive payoff. The end of the game was determined by the throw of a die, so as to avoid endgame effects.

Realizing that the repetition of the game favors cartel behavior, Holt tried a single-period experiment, which he felt would favor the Cournot equilibrium or possibly even more competitive behavior. Twelve experienced subjects engaged in a series of single-period games (with no guarantee as to which players were paired in a given game). In early games, the output choices were quite diverse, but eventually, virtually all the players chose the Cournot equilibrium output.

Holt concluded that in full-information duopoly games, whether or not there are multiperiod markets, the Cournot equilibrium is more likely than the competitive, or Bertrand, equilibrium. The only effect of experience and repeated games seems to be to raise prices.

Where explicit signaling is permitted, we would expect higher prices to be more likely. In a series of experiments, Friedman (1967) allowed players to transmit two written messages before privately making a price decision. Cartel outcomes were attained over 75 percent of the time. Further, 75 percent of these cartel agreements maximized each player's profits (with no side payments allowed). As should be expected, once the players succeeded in achieving the cartel solution, the probability of another cartel solution was 96 percent.

These experimental results have generally withstood the test of time and are still widely cited. There is now a large body of research on experimental methods and industrial organization. Based on his survey of the experimental literature, Holt (1995) drew the following conclusions. The one-period Cournot model often emerges as a good predictor of outcomes in one-period games. In multiperiod games with three or more sellers, the outcomes are often more competitive than a static Cournot model would predict. The price approaches the competitive level as the number of sellers increases if buyers and sellers propose prices simultaneously (a *double auction*) or if individual negotiations between each buyer and seller occur. Both these trading mechanisms result in lower prices than occur if sellers announce or post the prices at which they are willing to trade. Cooperation between players often increases as the number of times the game is repeated rises, but there has been no direct evidence that trigger-price strategies will result in cooperative outcomes without communications between players.

SUMMARY

Although most economists agree about the basic characteristics of oligopolistic markets, they do not agree about the best way to model these markets. Oligopoly models make very different assumptions about how firms behave; as a result, they make very different predictions about the nature of the equilibrium. Several conclusions can be drawn, however.

First, cartel outcomes are more likely in markets that last a long or uncertain period of time than in those that exist for only a short, known period of time. Experimental evidence supports the conclusion that cartel pricing is most likely to occur in repeated games. Explicit contact between firms increases the probability of achieving a monopoly price.

Second, most models (except the single-period, homogeneous-good Bertrand model) predict that the more firms in the industry, the more competitive is the equilibrium. The Bertrand model in which firms have constant marginal costs predicts the competitive equilibrium, regardless of the number of firms.

Third, experimental evidence indicates that the Cournot equilibrium is often (but not always) observed, especially in duopoly games. This evidence and the relative ease of using the model explain its continuing popularity.

Fourth, the reemergence of game theory has led to a better understanding of when strategies are credible to other firms. Research is ongoing to restrict the number of possible equilibria that can occur in multiperiod games with and without uncertainty.

All models in this chapter assume that the number of firms is fixed, firms produce homogeneous products, and firms maximize their profits by setting their expected marginal revenues equal to their marginal costs. The models differ only in the way in which firms calculate their expected marginal revenues. The next chapter extends these models to include product differentiation and the entry of new firms.

PROBLEMS

1. How does the Cournot equilibrium change if each firm faces a fixed cost of F as well as a constant marginal cost per unit?
2. Show the payoff matrix and explain the reasoning in the prisoners' dilemma example where Jeff and Dennis, possible criminals, will get one year in prison if neither talks. If one talks, one goes free and the other gets five years; and if both talk, both get two years. (*Note:* The payoffs are negative because they represent years in jail.)
3. What are the best strategies for Players 1 and 2 if each chooses between setting a low price or a high price and the payoffs are 5 if both firms charge the high price and zero for all other combinations of strategies? [*Hint:* Write down the 2 × 2 normal-form representation of this game and look for dominant strategies.]
4. What happens to price and output in the Cournot, Bertrand, and Stackelberg models if marginal costs increase by 10 percent?
5. For $n = 2$, 5, 10, 50, and 1,000, add columns to Table 6.2 for
   a. Market elasticity, ϵ, which equals $(\mathrm{d}Q/\mathrm{d}p)(p/Q)$.
   b. Lerner's measure of market power, $(p - MC)/p$.
   c. Consumer surplus.
   d. Social welfare = consumer surplus + industry profits.
   e. Deadweight loss (the amount by which social welfare is less than the optimum).

   Confirm that Lerner's measure of market power, $(p - MC)/p$, equals $1/(n\epsilon)$.
6. What is the relationship between the Stackelberg model and the dominant-firm-competitive-fringe model (Chapter 4)?
7. Using the data in Example 6.6, calculate the market demand elasticity for automobiles in the mid-1950s. For large changes in price and quantity, an *arc elasticity* is used. One common method of calculating an arc elasticity is to use the midway point between the two price-quantity pairs: (p, q) and (p^*, q^*). Thus, the formula for an arc elasticity is

$$\left(\frac{q - q^*}{q + q^*}\right) \Big/ \left(\frac{p - p^*}{p + p^*}\right).$$

   Is that number consistent with the theory that there was a profit-maximizing cartel in 1954? Why or why not?

Answers to the odd-numbered problems are given at the back of the book.

SUGGESTED READINGS

For a clear presentation of traditional oligopoly models and an introduction to game theory that is only slightly more technical than this textbook, see Shubik with Levitan (1980), Friedman (1983), and Ulph (1987). Williams (1966) has a good, relatively nontechnical discussion of simple games. Dixit and Nalebuff (1991) contains wonderful, nontechnical applications of game theory to a variety of economic problems. Binmore (1992) and Gibbons (1992) are relatively accessible game theory texts. Fudenberg and Tirole (1989) and Shapiro (1989) provide excellent surveys of oligopoly models. See Gaudet and Salant (199l) for a discussion of the effect of firms in related product markets.

Two short surveys of dynamic game issues are Kreps and Spence (1984) and Fudenberg and Tirole (1986b). Some of the relatively technical, recent textbooks on game theory are Shubik (1982, 1984), Friedman (1977, 1986), Mas-Collel, Whinston, and Green (1995), Tirole (1988), Fudenberg and Tirole (1991), and Myerson (1991).

APPENDIX 6A

A Mathematical Derivation of Cournot and Stackelberg Equilibria

This appendix uses calculus to derive the Cournot and Stackelberg equilibria prices and quantities for a general functional form and a linear example. Assume that there are n firms, where n is exogenously determined. The output of the ith firm is q_i and the total output, Q, is the sum of the (homogeneous) output of each firm: $Q = q_1 + \ldots + q_n$. The demand and cost functions are

	General Functional Form	Linear Example
Market demand	$p(Q)$	$p = a - bQ$
Firm's cost	$C(q_i)$	$C(q_i) = mq_i$

where a, b, and m are constants. In the example, demand is linear and marginal cost is constant. The competitive and monopoly solutions are

	General Functional Form	Linear Example
Competition	$MC \equiv C'(q_i) = p(Q)$	$m = a - bQ = p$ $Q = \dfrac{a - m}{b}$
Monopoly	$MC = C'(Q) = p'(Q)Q + p(Q) = MR$	$m = a - 2bQ = MR$ $Q = \dfrac{a - m}{2b}$ $p = \dfrac{a + m}{2}$

where a, b, and m are constants.

To analyze a Cournot industry, one starts by examining the behavior of a representative firm. Firm 1 tries to maximize its profits through its choice of q_1:

$$\max_{q_1} \pi_1(q_1, q_2, \ldots, q_n) = q_1 p(q_1 + \cdots + q_n) - C(q_1). \tag{6A.1}$$

The first-order condition is $MR = MC$, or

$$p(q_1 + \cdots + q_n) + q_1 p'(q_1 + \cdots + q_n) \times \left(1 + \frac{\partial q_2}{\partial q_1} + \cdots + \frac{\partial q_n}{\partial q_1}\right) = C'(q_1). \tag{6A.2}$$

If the firms play Nash-in-quantities (Cournot), these partial derivatives, $\partial q_i/\partial q_1$, are zero. Thus, the first-order condition may be rewritten as

$$p(q_1 + \cdots + q_n) + q_1 p'(q_1 + \cdots + q_n) = C'(q_1). \qquad (6A.3)$$

Rearranging terms in Equation 6A.3, multiplying and dividing the right-hand side by n, and noting that $p' = \mathrm{d}p/\mathrm{d}Q$ and $Q = nq_1$ (given that all firms are identical), one obtains the Lerner Index,

$$\frac{p - C'}{p} = -\frac{1}{n}\frac{\mathrm{d}p}{\mathrm{d}Q}\frac{Q}{p} = -\frac{1}{n\epsilon}, \qquad (6A.3')$$

where the second equality holds because the elasticity of market demand, ϵ, is $(\mathrm{d}Q/\mathrm{d}p)(p/Q)$. The left-hand side of Equation 6A.3′ is Lerner's measure of market power: the ratio of the price markup over marginal cost to the price. If the market is competitive, then $p = C'$, and Lerner's measure is zero. The larger the measure, the greater the market power. With symmetric firms, the elasticity facing any one firm is $n\epsilon$. Notice that, holding the market elasticity constant, as the number of firms increases, Lerner's measure falls. As n approaches ∞, the elasticity facing any one firm approaches $-\infty$, so Lerner's measure approaches $1/\infty$ or 0, and the market is competitive (see, however, Ruffin 1971).

Equation 6A.3 shows how the profit-maximizing q_1 depends upon $q_2, \ldots, q_n$. One can rearrange this expression, solving for q_1, to derive Firm 1's best-response function:

$$q_1 = R_1(q_2, \ldots, q_n). \qquad (6A.3'')$$

With our linear example, the first-order condition for profit maximization 6A.2 is

$$MR = a - b(2q_1 + q_2 + \cdots + q_n) = m = MC. \qquad (6A.4)$$

In equilibrium, because all firms have the same cost function, $q_2 = q_3 = \cdots = q_n \equiv q$. Solving Equation 6A.4 for q_1, the best-response function 6A.3″ for the first firm is

$$q_1 = R_1(q_2, \ldots, q_n) = \frac{a - m}{2b} - \frac{n - 1}{2}q. \qquad (6A.5)$$

The intersection of the best-response functions determines the Cournot equilibrium. In the example, that occurs where $q_1 = q_i = q(i = 2, \ldots, n)$. Setting $q_1 = q$ in Equation 6A.5 and solving for q gives

$$q = \frac{a - m}{(n + 1)b}. \qquad (6A.6)$$

Total output, nq, equals $n(a - m)/[(n + 1)b]$. The corresponding price is obtained by substituting $Q = nq$ into the demand function:

$$p = \frac{a + nm}{n + 1}. \tag{6A.7}$$

Setting $n = 1$ in the last two equations yields the monopoly quantity and price. As n becomes large, the quantity and price approach the competitive levels. That is, using Equations 6A.6 and 6A.7, as n grows large, total output, nq, approaches $(a - m)/b$ and price approaches m.

Similarly, using Equation 6A.7, Lerner's measure of market power, $(p - C')/p$, equals $(a - m)/(a + nm)$. As n grows large, the denominator goes to ∞, so Lerner's measure goes to 0, and there is no market power.

A Stackelberg leader (say, Firm 1) takes the Cournot best-response functions of the follower firms as constraints. That is, its objective is

$$\begin{aligned} &\max_{q_1} \pi_1(q_1, q_2, \ldots, q_n) = q_1 p(q_1 + \cdots + q_n) - C(q_1) \\ &\text{s. t. } q_i = R_i(q_1, Q_i) \qquad i = 2, \ldots, n, \end{aligned} \tag{6A.8}$$

where Q_i is the sum of the output of all the firms except Firm 1 and Firm i. Substituting the best-response functions into the profit expression for each q_i and differentiating with respect to q_1, we obtain the first-order condition for a profit-maximum.

For example, with a duopoly, the first-order condition for the Stackelberg leader is

$$p(q_1 + R_2(q_1)) + q_1 p'(q_1 + R_2(q_1))[1 + R'_2(q_1)] = C'(q_1), \tag{6A.9}$$

where R'_2 is the partial derivative of the best-response function of Firm 2 with respect to q_1. The follower's output is determined by setting the q_1 determined by Equation 6A.9 into the follower's best-response function.

In the linear example, each follower's best-response function is of the form of Equation 6A.5:

$$q_i = \frac{a - m}{2b} - \frac{(n - 2)q}{2} - \frac{q_1}{2}, \qquad i = 2, \ldots, n, \tag{6A.10}$$

where the output of the firms other than Firm i and Firm 1 is $(n - 2)q$. Because all the follower firms produce the same amount of output, q, the best-response function of the followers, Equation 6A.10, can be written as

$$q = \frac{a - m}{nb} - \frac{q_1}{n}. \tag{6A.10$'$}$$

The leader maximizes its profits, taking the best-response functions of the followers as given. The leader's first-order condition for profit maximization is given in Equation 6A.9. By differentiating Equation 6A.10′ with respect to q_1, one finds the slope of the followers' best-response functions, $\mathrm{d}R_i/\mathrm{d}q_1 = -1/n$, which we substitute into Equation 6A.9. That is, for every unit the leader firm's output rises, each follower firm's output falls by $1/n$, so the followers' collective output falls by $(n - 1)/n$. Thus, substituting the linear demand curve expression for p into Equation 6A.9 and solving for q_1, one obtains the output of the leader:

$$q_1 = \frac{a - m}{2b}. \qquad \text{(6A.11)}$$

In this linear model, q_1 is independent of the number of follower firms and equals the monopoly output. The output for the $n - 1$ (≥ 1) follower firms is

$$q = \frac{a - m}{2bn}. \qquad \text{(6A.12)}$$

Thus, $q_1 > q$ for any number of firms, n (≥ 2).

Total industry output is

$$Q = \frac{a - m}{2b}\left(\frac{2n - 1}{n}\right), \qquad \text{(6A.13)}$$

which exceeds the Cournot market output of $[(a - m)/\mathrm{b}][n/(n + 1)]$, using the Cournot q from Equation 6A.6. The market price is

$$p = \frac{a + m(2n - 1)}{2n}. \qquad \text{(6A.14)}$$

Thus, as the number of firms, n, grows large, price and total quantity approach the competitive levels: $p \rightarrow m$ and $Q \rightarrow (a - m)/b$.

APPENDIX 6B

Mixed Strategies

In the text, only *pure* strategies, where a player picks a single price or quantity, were considered. Players, however, may use *mixed* strategies where they choose their actions randomly. For example, suppose Figure 6.9 were changed so that the payoff were:

		Firm 1	
		Low price	High price
Firm 2	Low price	0 2	2 1
	High price	7 0	6 6

Given these payoffs, Firm 2 wants to match Firm 1's price, but Firm 1 does not want to match Firm 2's price. Both of them setting a low price is not a Nash equilibrium because if Firm 2 believes Firm 1 will set a low price, it wants to set a high price. Nor is Firm 2 setting a low price and Firm 1 setting a high price a Nash equilibrium, because Firm 2 would want to change its behavior. Similarly, the other possible pairs of strategies are not Nash equilibria. The only possible Nash equilibrium is for the players to choose their actions randomly.

Let α be the probability that Firm 1 sets a low price and β be the probability that Firm 2 sets a low price. If the firms choose their prices independently, then $\alpha\beta$ is the probability that both set a low price, $(1 - \alpha)(1 - \beta)$ is the probability that both set a high price, $\alpha(1 - \beta)$ is the probability that Firm 1 prices low and Firm 2 prices high, and $(1 - \alpha)\beta$ is the probability Firm 1 prices high and Firm 2 prices low.

Firm 2's expected payoff, $E(\pi_2)$ is

$$\begin{aligned} E(\pi_2) &= 2\alpha\beta + (0)\alpha(1 - \beta) + (1 - \alpha)\beta + 6(1 - \alpha)(1 - \beta) \\ &= (6 - 6\alpha) - (5 - 7\alpha)\beta. \end{aligned}$$

Similarly, Firm 1's expected payoff is

$$\begin{aligned} E(\pi_1) &= (0)\alpha\beta + 7\,\alpha(1 - \beta) + 2(1 - \alpha)\beta + 6(1 - \alpha)(1 - \beta) \\ &= (6 - 4\beta) + (1 - 3\beta)\alpha. \end{aligned}$$

Each firm must form a belief about its rival's behavior. For example, suppose Firm 1 believes that Firm 2 will choose a low price with a probability β^e. If β^e is less than 1/3 (that is, Firm 2 is relatively unlikely to choose a low price), it pays for Firm 1 to choose the low price because the second term in $E(\pi_1)$, $(1 - 3\beta)\alpha$, is positive, so that as α increases, $E(\pi_1)$ increases. Because the highest possible α is 1, Firm 1 chooses the low price with certainty. Similarly, if Firm 1 believes β^e is greater than 1/3, it sets a high price with certainty ($\alpha = 0$).

If Firm 2 believes that Firm 1 thinks β^e is slightly below 1/3, then Firm 2 believes Firm 1 will choose a low price with certainty, and hence Firm 2 will also choose a low price. That outcome, $\beta = 1$, however, is not consistent with what Firm 1 expects (β^e is a fraction). Indeed, it is only rational for Firm 2 to believe Firm 1 believes Firm 2 will use a mixed strategy if Firm 1's belief about Firm 2 makes Firm 1 unpredictable. That is, Firm 1 uses a mixed strategy only if it is *indifferent* between setting a high or a low price. It is only indifferent if it believes β^e is exactly 1/3. By similar reasoning, Firm 2 will use a mixed strategy only if its belief is that Firm 1 chooses a low price with probability $\alpha^e = 5/7$. Thus, the only possible Nash equilibrium is $\alpha = 5/7$ and $\beta = 1/3$.

It can be shown that every game with a finite number of players, each of which has a finite number of pure strategies, has at least one Nash equilibrium, possibly in mixed strategies. Proofs are provided in the game theory texts cited in the recommended readings.

Many game theorists, however, do not like the concept of mixed strategies in static games. They believe that few people actually randomize over their pure strategies. It is hard to imagine a manager of a firm rolling dice to decide what price to charge tomorrow. One (weak) response to this objection is that the firms only have to *appear* to be unpredictable to each other.

A few mixed strategy models have been estimated. See, for example, Golan, Karp, and Perloff (2000).

CHAPTER 7

Product Differentiation and Monopolistic Competition

Good taste is better than bad taste, but bad taste is better than no taste.

—Arnold Bennett

In many markets firms engage in *monopolistic competition:* Firms have *market power,* the ability to raise price profitably above marginal cost, yet they make zero economic profits. Such a market structure combines attributes of monopoly (market power) and competition (zero economic profits). An industry has monopolistic competition if there is *free entry* and each firm faces a *downward-sloping demand curve.* If firms enter the industry whenever positive profits are available, each firm makes zero economic profits in the long run, as in a competitive industry. If a firm faces a downward-sloping demand curve, it has market power.

An important reason why a firm faces a downward-sloping demand curve is that consumers view its product as different from those of other firms in the industry. In previous chapters, we concentrated on industries with **homogeneous** or **undifferentiated** goods: Products are viewed as identical by consumers. That is, consumers view the products as *perfect substitutes* for each other. In many industries, however, products are typically **heterogeneous** or **differentiated**: Consumers consider products or brands of various firms to be imperfect substitutes. If consumers view brands in an industry as imperfect substitutes, a firm may raise its price above that of its rivals without losing all its customers.

The models analyzed in this chapter differ in two ways from the oligopoly models in Chapter 6. First, entry is impossible in an oligopolistic market (by definition), but firms can freely enter and exit an industry characterized by monopolistic competition. In this chapter, the number of firms is determined within the model by entry behavior rather than arbitrarily determined outside the model as in the oligopoly chapter. Second, in Chapter 6, we assumed that oligopolistic firms produce identical products, whereas in this chapter products differ across firms.[1]

In the models of Chapter 6, in which firms produced homogeneous goods, an increase in the number of oligopolists benefited consumers because the additional competition led to lower prices. If firms produce differentiated products, the entry of a new firm helps consumers for two reasons: It lowers prices and increases the variety of products from which to choose.

Both these effects are illustrated in models of monopolistic competition. There are two major types of monopolistic competition models with free entry and differentiated products. In one, the *representative consumer model,* all firms compete equally for all consumers who typically buy from each firm. This model might be used to study the restaurant market, in which firms produce differentiated products (such as different ethnic cuisines), but all compete for the same customers.

In the other, the *spatial* or *location* model, each consumer prefers products that have certain characteristics or are sold by firms located near him or her and is willing to pay a premium for these preferred products. Moreover, the consumer may not care greatly about the price of some other goods in the market. For example, a consumer whose favorite cereal is Kellogg's corn flakes is more sensitive to the relative price of Post's corn flakes than to the relative price of Nabisco's sugar-coated shredded wheat. The other brand of corn flakes is a much better substitute than other types of cereal.

These models differ in the type of demand each firm faces. In the representative consumer model, a firm's demand varies continuously with the prices of all firms. A small change in any one firm's price causes a relatively small change in the demand facing a firm. In the location model, as the cereal example suggests, the demand for one brand may be either independent of some other brand's price because they are not close substitutes, or highly dependent on another brand's price because they are close substitutes. Moreover, a firm may, at some very low price, gain a large number of extra consumers as it captures all the customers of another firm that produces a very similar product.

Either model can be used to study the welfare of consumers and firms by comparing the monopolistic competition equilibrium to the social optimum in terms of price and variety. This chapter asks whether there are too many or too few brands in the monopolistic competition equilibrium. The answer to this question depends on how much more consumers are willing to pay for greater variety (it is expensive to produce many types of products). Which would you prefer: a choice of three different-flavored soft drinks at 50¢ per drink or only one flavor at 25¢? The answers to such questions determine the optimal variety-price combination.

The first section of this chapter explains why product differentiation affects the demand curve facing a firm. Then the two most widely used models of monopolistic

[1] Differentiated products are sold in most oligopolistic markets. This heterogeneity was ignored in the previous chapter for simplicity. The analysis of differentiation in this chapter may be applied to those oligopoly models.

competition are discussed. Representative consumer models of monopolistic competition with both homogeneous and heterogeneous products are examined. The discussion shows how the homogeneous product, Cournot-Nash oligopoly model presented in the previous chapter changes when free entry is allowed and then describes how the equilibrium price in this model compares to the social optimum. Next, the model is modified to allow for product differentiation, and price and variety in the monopolistic competition equilibrium are compared to the socially optimal combination.

The discussion then turns to a location model. Product differentiation is inherent in location models, so no homogeneous product model is presented. Again, the welfare implications are examined. Finally, hybrid models that have elements of both models are used to explain why these two types of models have different properties.

The key questions in this chapter are

1. Why does product differentiation increase firms' market power?
2. What number of firms maximizes welfare if all brands are perfect substitutes (homogeneous)?
3. What number of firms maximizes welfare if consumers view brands as imperfect substitutes for each other?
4. What number of firms maximizes welfare if consumers only value some of the brands in the market?

Differentiated Products

The study of an industry of differentiated products is based on two key concepts. First, products are differentiated because consumers *think* they differ. That is, even though aspirin brands may be chemically identical, if consumers believe that the products differ and shop accordingly, then the products are effectively differentiated. For example, "commodity" beans (pintos, Great Northerns, and so forth) from Golden Grain/Mission sell for about 69¢ per pound at your local grocery, whereas those from Melissa's, which are in a pretty package, go for about $4.59 per pound.[2]

Similarly, many consumers strongly prefer Coke to Pepsi or vice versa, yet they have trouble differentiating them by taste. When regular cola drinkers were given samples of Coca Cola Classic, Pepsi, Diet Coke, and Diet Pepsi, only 37 percent could correctly identify the brand they said they preferred. Only 26 percent of diet cola consumers could identify their brand.[3]

[2] M. A. Mariner, "Consumers Are Willing to Pay a Lot for a Pretty Package," *San Francisco Chronicle,* April 30, 1997: Food 3.

[3] Consumer Union, "The Cola Wars," *Consumer Reports* 56, August 1991:518–25. In a similar study at Williams College ("Diet Cola Advertising Gets Put to the Test." *San Francisco Chronicle,* January 31, 1990:C1), only an "insignificant" number of subjects could consistently tell the difference between Diet Coke and Diet Pepsi. More than one-third of those professing a preference for one brand chose the other in the test.

Conversely, if consumers view chemically or physically different products as identical, then for economic purposes they are homogeneous: "The consumer is always right." See **www.aw-bc.com/carlton_perloff** "Spurious Product Differentiation: A Drug on the Market."

Second, the pricing of one brand exerts a greater constraint on another brand's pricing when the two brands are close substitutes than when they are not. For example, few would dispute that Pepsi Cola and Coca Cola are close substitutes. Indeed, Canada Dry Ginger Ale may also compete with Coke and Pepsi, because they are all soft drinks with sugar. But are soft drinks without sugar close substitutes? What about noncarbonated drinks like milk and water?[4] See Example 7.1.

An example of a market with homogeneous products is wheat: Consumers do not care which farm produced a particular bushel of wheat. It is harder to think of industries with a small number of firms whose products consumers view as perfectly identical, but there are a number of industries whose products consumers may view as nearly identical. Delivery services in a city may be viewed as quite similar. Different brands of beach balls may also strike most consumers as very close substitutes. An industry has relatively homogeneous products if consumers do not care which brand they buy.

There are two approaches to analyzing differentiation. In the standard consumer theory of basic microeconomics books, consumers have preferences regarding commodities: They choose between ice cream and cake or between brands of ice cream and cake. In an alternative formulation, consumers have preferences regarding the attributes, or characteristics, of commodities. For example, some consumers love chocolate, a characteristic of some ice creams and cakes. These consumers prefer either chocolate ice cream or chocolate cake to vanilla ice cream or white cake.

The Effect of Differentiation on a Firm's Demand Curve

> *A cynic is a man who knows the price of everything, and the value of nothing.*
>
> *—Oscar Wilde*

In industries with undifferentiated products, the demand facing a particular firm depends only on the total supply of its rivals, whereas in an industry with differentiated products, the demand facing a firm depends on the supply of each of its competitors separately. For industries with either differentiated or undifferentiated goods, we can write the inverse demand curve facing Firm i as:

$$p_{\mathrm{i}} = D(q_1, \ldots, q_{\mathrm{n}}). \tag{7.1}$$

[4] The definition of a market is often a crucial issue in antitrust and merger cases (see Chapter 19). Often, expert witnesses in these cases contend that if products are "close substitutes," they are part of the same *market.* Throughout this book, unless otherwise noted, the term *market* is used loosely, without reference to legal definitions. This chapter assumes each firm's product is in the market being discussed, in the sense that at least some consumers view it as a substitute for at least some other products in the market. That is, it assumes the products are "adequately close" substitutes without defining what "adequately close" means.

EXAMPLE 7.1 *All Water Is Not the Same*

Until recently, few people thought of water as a product that could be differentiated. But with clever marketing, firms have convinced consumers that water is indeed a differentiated product.

In 2000, 5.0 billion gallons of bottled water were consumed (only about one-third of the consumption of carbonated beverages) and sold for $6 billion. By 2002, sales exceeded $7.7 billion. The many brands of bottled water appeal to different segments of the market and sell for very different prices:

Brand	Price per Quart (bottle size)	Source	2002 U.S. Wholesale Sales
Aquafina	$0.88 (1.5 liter)	Purified tap water	$838.0 million
Dasani	$1.58 (20 oz)	Purified tap water	$765.0 million
Poland Spring	$0.92 (24 oz)	Spring in Maine	$621.5 million
Deer Park	$1.32 (24 oz)	Springs in Florida, Maryland, & Pennsylvania	$311.1 million
Crystal Geyser	$0.77 (six pack of 1 liter bottles)	Springs in California & Tennessee	$270.0 million
Evian	$1.46 (1 liter)	Spring in the French Alps	$191.1 million

As the table shows, the best-selling brand, Aquafina, is moderately priced, while the second-place top brand, Dasani, costs nearly twice as much. Coca-Cola, Pepsi, Nestlé, and the other companies that sell bottled water each typically offer an array of different-price brands. Recently, water bottlers have added flavors in order to differentiate their products further. These flavored waters have become increasingly important, and their sales rose tenfold between 1999 and 2002.

Sources: J. Jordan and S. He, "Size Counts: The Economic Value of Bottled Water," *Choices*, September 22, 2002; International Bottled Water Association, "Marketing Statistics Gallonage by Segment," **www.bottledwater.org/public/gallon_byseg.htm**; "U.S. Soft Drink Sales Slow in 2002," Beverage Marketing Corporation of New York news release (July 24, 2003); Phil Lempert, "Navigating the Sea of Bottled Water," *Today Show*, June 17, 2003; prices are from peapod.com for Washington, DC, on August 27, 2003; source information comes from brand Web sites and Betsy McKay and Robert Frank, "Coke, Danone to Announce Venture—Pact to Market, Distribute Bottled Spring Water Could Challenge PepsiCo, Nestlé," *Wall Street Journal*, June 17, 2002:B5; sales data come from "Bottled Water Moves Up in the Rankings, Says Beverage Marketing Corporation," Beverage Marketing Corporation of New York news release (May 19, 2003).

That is, the price, p_i, that Firm i may charge depends on the quantity of its brand sold and the quantities of all other $n - 1$ brands. Where products are differentiated, this expression cannot be simplified. One can also write the demand curve facing Firm i as a function of the prices of each rival product, $q_i = \tilde{D}(p_1, p_2, \ldots, p_n)$.

If consumers view all products as identical, or perfect substitutes, however, the demand curve may be written more simply. Consumers are unwilling to pay more for one firm's product than another's. Thus, all firms must charge the same price, p, if all are to sell their products. With undifferentiated products, only total market output, $Q = q_1 + q_2 + \cdots + q_n$, matters in determining the price, p.[5] In this case, the inverse demand equation may be written as

$$p_i = p = D(q_1 + q_2 + \cdots + q_n) = D(Q). \tag{7.2}$$

As an example, suppose there are two firms in an industry. If the two products are viewed by consumers as identical, the price each firm may charge ($p = p_1 = p_2$) might be written as

$$p = a - bQ = a - b(q_1 + q_2) = a - bq_1 - bq_2, \tag{7.3}$$

where a and b are positive constants. That is, an increase in either firm's output reduces the market price—and hence the price for each firm—by an equal amount.

In contrast, if consumers view the products as imperfect substitutes, Firm 1's demand curve may be

$$p_1 = a - b_1q_1 - b_2q_2, \tag{7.4}$$

where $a > 0$ and $|b_1| > |b_2|$. That is, an increase in Firm 1's output has a greater effect on its price than an increase in Firm 2's output. Indeed, the more a firm succeeds in differentiating its product, the more insulated its demand is from the actions of other firms. For example, a change in the quantity sold or the price of Ripple or Thunder Bird, which are inexpensive wines in screw-top bottles, may have negligible effects on the price or demand for expensive wines.

Oligopolies or monopolistic competition markets may have differentiated goods, but, in a perfectly competitive market, products are not differentiated. If a firm's product is differentiated, it faces a downward-sloping demand function, which is inconsistent with a competitive firm's price-taking behavior.

Preferences for Characteristics of Products

In Lancaster's (1966, 1971, 1979) and Becker's (1965) consumer theories, consumers have preferences over the characteristics of commodities. Each commodity is a bundle of characteristics. For example, candy bars and ice cream vary in sweetness, temperature, texture, and so forth. Rather than comparing the products as such, consumers choose on the basis of the more fundamental characteristics.

[5] If the brands of two firms are perfect substitutes, a consumer's indifference curve for the goods is a straight line with a slope of −1. That is, a consumer is indifferent between having 20 units of Brand 1 and 0 units of Brand 2, or 10 units of each, or 0 units of Brand 1 and 20 units of Brand 2. The consumer's utility depends only on the sum of the output of the two brands.

To illustrate how products can be compared by examining their characteristics, suppose the only important characteristic of a soft drink is how sweet it is. Soft drinks are located in "sweetness" space:

Not Sweet ←——→ Sweet

In this space, Schwepps Club Soda is located to the left of Classic Coke, which is to the left of Pepsi. That is, the sweeter products are, the further to the right they are located. Soft drinks, then, can be said to be located in a characteristic space: There is an axis showing the amount of each characteristic or attribute (here there is only one attribute, sweetness), and each brand can be located in this space according to its characteristics.

Of course, a product may have many characteristics: Cereal brands may differ by sweetness and "mouth feel." If those are the only important characteristics, then cereal brands can be located in a characteristic space that has sweetness on one axis and mouth feel (from soggy to crunchy) on the other.

The representative consumer model may use either the product or characteristic approach; location models inherently use a characteristic approach. We examine both models in turn.

The Representative Consumer Model

I alone am here the representative of the people. —*Napoleon Bonaparte*

The first monopolistic competition model was developed by Chamberlin (1933). In this representative consumer model, the typical consumer views all brands as equally good substitutes for each other; hence, brands are treated symmetrically. This representative consumer model can be used to examine industries with either differentiated or undifferentiated products. We start by examining undifferentiated product markets and then extend the analysis to markets in which products are heterogeneous. The analysis shows that whether or not products are differentiated, the equilibrium prices and number (variety) of brands in a monopolistic competition equilibrium are not generally socially optimal.

A Representative Consumer Model with Undifferentiated Products

In the simplest version of the representative consumer model, the various brands are homogeneous: All brands have the same characteristics. This model differs from the oligopoly models of the previous chapter only in the way the number of firms in the industry is determined. Both the oligopoly models and the monopolistic competition model determine the output of each firm. In both models, profit-maximizing behavior determines the output of each firm. That is, each firm chooses its output so that its

TABLE 7.1 **Comparison of Oligopoly and Monopolistic Competition Models**

Model	Profit Maximization by Individual Firms	Number of Firms (n) Determined by Entry
Noncooperative oligopoly	marginal revenue = marginal cost	No entry: number of firms is fixed at n
Monopolistic competition	marginal revenue = marginal cost	Free entry: firms enter until profit = 0, so n is endogenously determined

marginal revenue corresponding to its residual demand curve, MR_r, equals its marginal cost, MC.

Entry is treated differently in the two models. In the oligopoly models, the number of firms is arbitrarily determined outside the model: The existing firms, the government, or some other force prevents new entry. In Chamberlin's model, firms freely enter the industry as long as it is profitable for them to do so. This *entry condition* determines the number of firms in the industry within the model. The two conditions that determine the oligopolistic and monopolistic competition equilibria, profit maximization and entry, are shown in Table 7.1.[6]

The monopolistic competition model requires that firms face downward-sloping demand curves. Although product differentiation leads to such demand curves, high fixed costs can have the same result by limiting the number of firms that enter the industry, as the following example shows.

A Cournot Example. To illustrate how the monopolistic competition model with homogeneous goods differs from an oligopoly model, the Cournot-Nash model of a noncooperative oligopoly is modified to allow entry; otherwise the same assumptions are made as in the oligopoly example of Chapter 6:

- *Cournot equilibrium:* In equilibrium, no firm wants to change its output level, and each firm expects its rivals to produce at their actual level of output.
- *Homogeneity:* Output is homogeneous.
- *Demand:* The quantity that the market demands, Q, is a function of the market price, p:

$$Q = 1{,}000 - 1{,}000p. \tag{7.5}$$

- *Costs:* Each firm has a cost function of

$$C(q) = 0.28q + F, \tag{7.6}$$

[6] The shape of the marginal revenue curve referred to in Table 7.1 depends on the game (Bertrand or Cournot) being played.

where q is the firm's output and F is its fixed cost. As in Chapter 6, marginal cost is constant at 28¢.

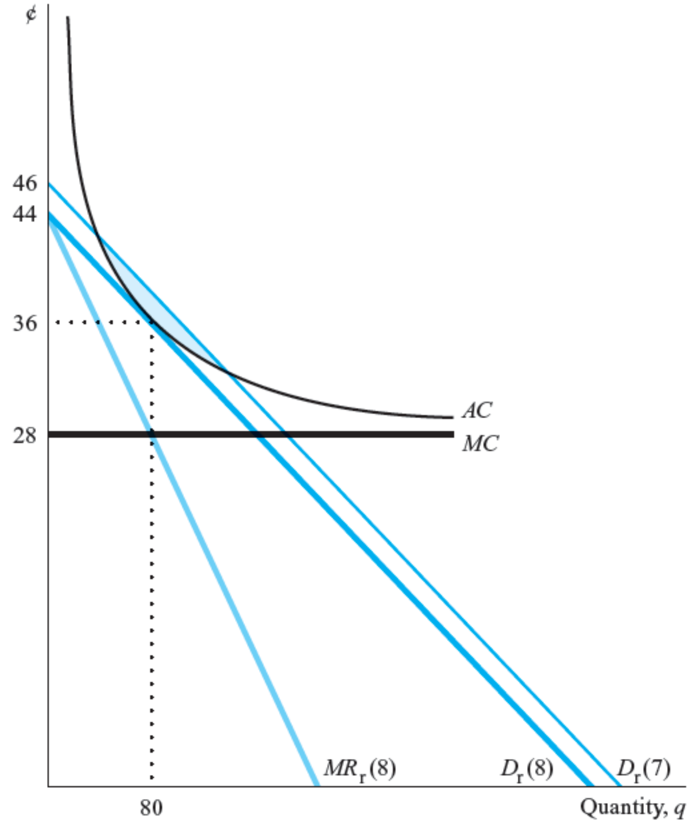
The assumption of the oligopoly model in the previous chapter of a fixed number of firms is replaced by the entry condition: Firms enter the market when profits are positive and exit when profits are negative.

The marginal cost, *MC*, and average cost, *AC*, curves are shown in Figure 7.1. Marginal cost is a horizontal line at 28¢. The average cost may be calculated by dividing total cost from Equation 7.6, $C(q)$, by output. That is,

$$AC = \frac{C(q)}{q} = 0.28 + \frac{F}{q}.$$

Thus, average cost is the sum of average variable costs ($\$0.28 = [\$0.28\ q]/q$) and average fixed cost (F/q). As output grows, fixed costs are spread over more and more

FIGURE 7.1 Monopolistically Competitive Equilibrium



units, so average fixed costs fall, and the average cost consists primarily of average variable costs. As a result, AC is well above MC at low output levels and approaches MC (which is the same as average variable costs) as q gets large, as shown in Figure 7.1.

The entry condition says that firms enter the industry as long as profits are positive. Thus, firms enter the industry until economic profits are driven to zero:[7]

$$\pi = pq - C(q) = 0. \tag{7.7}$$

Thus, in long-run equilibrium, each firm makes zero profit overall; hence, it makes zero profit per unit, and each firm's average cost equals its price, $AC = p$.[8]

To determine the equilibrium number of firms, we use a two-step procedure. We first determine the Cournot equilibrium output for each possible number of firms (see Chapter 6). Second, we determine the number of firms by examining these equilibria and picking the one in which firms make zero profits.

To illustrate how the number of firms is determined in a monopolistic competition industry, suppose that each firm has a fixed cost of \$6.40. Thus, a firm enters this industry if profits are positive or, equivalently, if price is greater than average cost, $AC = 0.28 + 6.40/q$.

Table 7.2 shows market price, firm output, and profit for various numbers of firms. If there are initially five firms in the industry, each produces 120 units of output, and the market price is 40¢. Each firm makes a profit of \$8.00 [$= (p - AC)q = (\$0.40 - \$0.3333)120$].

If another firm enters, profit per firm falls to \$4.18. Because profits are still positive, more firms enter. Entry continues until eight firms are in the industry, and each one exactly breaks even. Because no firm is losing money, none has an incentive to leave the industry. No additional firm has an incentive to enter. As Table 7.2 shows, if a ninth firm enters, each firm loses \$1.22, so there is an incentive for firms to exit the industry. Thus, in this industry, the equilibrium number of firms is eight.

Graphic Analysis. This equilibrium can be determined graphically. Figure 7.1 shows the residual demand curve, $D_r(8)$, that each of the eight Cournot firms believes it faces, and the corresponding marginal revenue curve, $MR_r(8)$. The firm maximizes its profits by producing $q = 80$ units of output so that its $MR_r = MC$, as shown. It sells its output at the market price of 36¢. The firm's average cost curve is tangent to the demand curve ($p = 36¢ = AC$) where $q = 80$. As a result, the firm makes zero profit.

[7] The following discussion assumes that the profits of the last entrant are exactly zero. That condition does not always hold if there must be a whole number of firms. If there cannot be a fractional number of firms, profits may be positive in equilibrium, but if one more firm entered, all would make losses. Seade (1980) shows that the basic results discussed here hold even when one assumes that there must be a whole number of firms.

[8] We can write profit as $\pi = pq - C(q) = (p - C(q)/q)q = (p - AC)q$. Thus, if profit is zero, $\pi = 0$, then (dividing through by q) average profits must equal zero, $p - AC = 0$; hence, price equals average cost, $p = AC$. In our particular example, average profits are zero if $p = AC = 0.28 + F/q$.

TABLE 7.2 Cournot Monopolistic Competition Example with Different Fixed Costs (*F*)

			$F = \$6.40$		$F = \$1.60$	$F = \$0.00$
Number of Firms	Price (¢)	Firm Output	Average Costs (¢)	Firm Profit ($)	Firm Profit ($)	Firm Profit ($)
1	64	360	29.8	123.20	128.00	129.60
2	52	240	30.7	51.20	56.00	57.60
3	46	180	31.6	26.00	30.80	32.40
4	42.4	144	32.4	14.34	19.14	20.74
5	40	120	33.3	8.00	12.80	14.40
6	38.3	102.9	34.2	4.18	8.98	10.58
7	37	90	35.1	1.70	6.50	8.10
8	36	80	36.0	0.00	4.80	6.40
9	35.2	72	36.9	−1.22	3.58	5.18
10	34.5	65.5	37.8		2.68	4.28
11	34	60	38.7		2.00	3.60
12	33.5	55.4	39.6		1.47	3.07
13	33.1	51.4	40.4		1.04	2.64
14	32.8	48	41.3		0.70	2.30
15	32.5	45	42.2		0.42	2.03
16	32.2	42.4	43.1		0.19	1.79
17	32	40	44.0		0.00	1.60
18	31.8	37.9	44.9		−0.16	1.44
20	31.4	34.3	46.7			1.18
100	28.7	7.1	118			0.05
500	28.1	1.4	473			0.002
1,000	28.1	0.7	918			0.001
∞	28	~0	~∞			0.00

Note: The negative profits shown in the table represent the profits that would occur if the number of firms indicated produced at their profit-maximizing (loss-minimizing) levels, given that exit was impossible and fixed costs were sunk. If costless (no sunk costs) exit is possible, these firms shut down to avoid making losses.

Figure 7.1 shows that if only seven firms are in the industry, it pays for a firm to enter. The demand facing one of the seven Cournot firms, $D_r(7)$, cuts the average cost curve, so that there is a shaded region where average costs are lower than the price on the residual demand curve. A firm that operates at a point within this region makes a positive profit because its price is above its average cost. As Table 7.2 shows, each of the seven firms maximizes its profit at 90 units of output, so the market price is 37, which is greater than each firm's $AC = 35.1$¢.

Lower Fixed Costs. How does this monopolistic competition equilibrium change if each firm incurs lower fixed costs? If fixed costs are $1.60, the new equilibrium has 17 firms (compared to eight when fixed costs are $6.40), as Table 7.2 shows.

Thus, the lower the fixed costs, the higher the equilibrium number of firms in a monopolistic competition industry. The reason for the increase in the equilibrium number of firms is that the lower the fixed costs, the higher the profits for any given number of firms in an industry. Additional firms must enter the industry to drive profits to zero.

How do we know that each firm's profit is higher (holding the number of other firms constant), the lower is its fixed cost? The reason is that a reduction in a firm's fixed cost does not affect its total revenues but does lower its total costs. Although fixed costs affect a firm's decision about whether to produce at all, they do not influence output levels if the firm actually produces. Each firm sets its output where $MR_r = MC$, and neither MR_r nor MC are affected by a change in the firm's fixed cost. A producing firm sells the same output regardless of the level of fixed costs, so its total revenues and total variable costs are not affected by a change in fixed costs. Total costs equal variable costs plus fixed costs, so holding variable costs constant and lowering fixed costs causes total costs to fall. Because total revenues remain constant as total costs fall, profits rise.

Graphically, with lower fixed costs, the average cost curve lies strictly below the one in Figure 7.1. For the new average cost curve to be tangent to a firm's demand curve, the demand must be lower as well. The only way to get a lower demand curve is to have more firms in the industry. It follows from this reasoning that if fixed costs fall to zero, the number of firms becomes unlimited, and this Cournot monopolistic competition industry becomes perfectly competitive, as the last column of Table 7.2 shows.

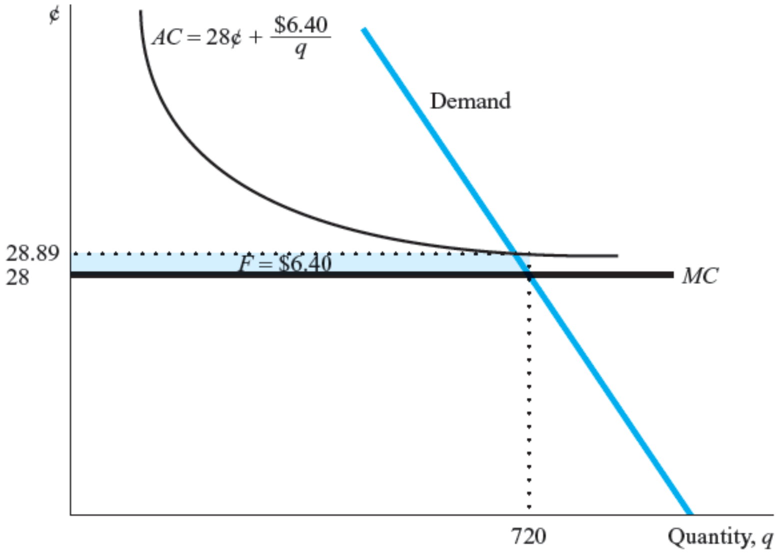
To summarize: High fixed costs cause price to be above marginal cost. Where there are no fixed costs, enough firms enter the industry to drive price to marginal cost: the competitive solution. See Example 7.2 on the effects of entry on price.

Welfare with Undifferentiated Products. How does this equilibrium compare to the social optimum in which welfare is maximized? Two welfare or efficiency problems arise with this monopolistic competition equilibrium. First, because price is above marginal cost, the industry produces too little total output: An extra unit of this product is worth more to consumers than it costs firms to produce it. Second, the number of firms is excessive when marginal costs are nonincreasing (constant or falling with quantity). Each additional firm must pay a fixed cost, F, so fixed costs to society are excessive.

EXAMPLE 7.2 *Entry Lowers Prices*

In the first year after United Airlines entered the short-hop market along the West Coast of the United States, prices plunged as much as 70 percent. The lowest fare United, Delta, and US Air were offering between San Francisco and Los Angeles in September 1994, before entry, was $133. A year later, after United had entered the market, the lowest rate was $39. Similarly, rates from San Jose to Seattle went from $79 on Alaska Airlines and $59 on Reno Air in 1994 to $49 on all lines in September 1995 after United and Southwest entered this market. Entry was very likely responsible for these price declines, as the average fares across all U.S. pairs rose during this period.

FIGURE 7.2 First Best



In the preceding example, each firm's cost function is $C(q) = mq + F$, where m is a firm's constant marginal cost. Here, society's optimal solution is to subsidize one firm to produce all the output and to require that price be set equal to marginal cost. The best possible solution (ignoring costs of administration) is referred to as the **first-best optimum** (see Appendix 7A).

Figure 7.2 illustrates the first-best solution. It shows a single firm's marginal and average cost curves and the market demand curve, based on the preceding example with fixed costs of \$6.40. In the proposed first-best equilibrium, the firm is regulated so that it sets its price equal to marginal cost, $m = 28¢$, and consumers purchase $q^* = 720$ units of output. The socially optimal output is 80 units (12.5 percent) more than the monopolistic competition output of 640.

At that price, the firm loses money because the price is less than the average cost ($p = \$0.28 < m + F/q^* = \$0.28 + \$6.40/720 = \$0.2889$). Thus, the government must subsidize the firm if it is to stay in business.[9] The shaded area in Figure 7.2 represents the subsidized loss $= F = \$6.40 = \$0.0089 \times 720 = (F/q^*)q^*$. The firm sells its product at its marginal cost or average variable cost, so it covers its out-of-pocket production expenses, but it does not recover its fixed costs.

[9] The government could raise the necessary revenues by taxing away the consumer surplus by charging consumers a fixed fee (to consume at all) and a price equal to the marginal cost for each unit consumed.

The consumer surplus at the social optimum is \$259.20.[10] If we define welfare as the sum of consumer surplus plus revenues minus costs, welfare at the social optimum is \$252.80. In contrast, in the monopolistic competition equilibrium, consumer surplus and welfare are \$204.80. Thus, welfare at the social optimum is 23.4 percent higher than in the monopolistic competition equilibrium.

If there is only one firm that can set its price however it likes, it will act like a monopoly, setting its p_m = 64¢ and selling q_m = 360 units. Its price is above its average cost (29.78¢), so it makes positive profits of \$123.20. Here, consumer surplus is \$64.80 and welfare is \$188. Thus, welfare in the monopolistic competition equilibrium is 8.9 percent higher and welfare at the social optimum is 34.5 percent higher than in the monopoly equilibrium.

When a single firm has a downward-sloping average cost curve, it is called a *natural monopoly* (Chapter 4) because one firm could fulfill all consumers' demands more cheaply than could two or more firms. Each firm could produce at the same marginal cost, but entry by an additional firm requires an additional expenditure of fixed costs, *F.* Thus, in the monopolistic competition equilibrium, not only is price above marginal cost, but if there are eight firms, too much has been spent in fixed costs. That is, one firm could produce the total monopolistic competition output for \$44.80 ($= 7F$) less than eight firms could because of the savings on fixed costs. In this example, unnecessary fixed costs represent 20 percent of total industry costs.

Even if firms have U-shaped average cost (*AC*) curves, there are too many firms in a homogeneous Cournot equilibrium. With U-shaped curves, the equilibrium occurs where each firm's residual demand curve is tangent to its *AC* curve (profits are zero). Because the residual demand curve is downward sloping, this tangency occurs in the downward-sloping (increasing returns to scale) section of the *AC* curve. Thus, the firms operate at a smaller output than the output that minimizes their *AC.* That is, monopolistically competitive firms have "excess capacity." There are too many small firms producing the output compared to the social optimum: The same output could be more efficiently produced with fewer firms.

Typically, the government cannot regulate an industry so as to achieve a first-best solution and maximize society's welfare. For example, it may be politically infeasible to subsidize a monopoly such as a local electric company. In some industries, the government may be able to control the number of firms, but it may not be able to force them to produce more than the profit-maximizing quantity if it is unwilling to subsidize them. Many cities control the number of taxicabs, for example.[11] By choosing the optimal number of firms, the government can achieve the **second-best optimum**: the best possible outcome subject to a constraint that violates one of the conditions for a

[10] Consumer surplus equals the triangle under the demand curve above 28¢. If demand is $p = a - bq$, then consumer surplus at quantity q is $1/2[a - p(q)]q = 1/2[a - (a - bq)]q = 1/2bq^2$. In our example, consumer surplus is $0.0005q^2$.

[11] Many economists argue that the number of taxicabs is restricted to drive up the profits of those lucky enough to be allowed to operate (see the evidence in Chapter 20). That is, rather than trying to maximize social welfare, the government is trying to enrich existing cab companies.

first-best outcome. That is, welfare is raised to the highest level possible given that the government does not subsidize firms.

The government faces a trade-off. If it allows more firms to enter, it can drive the market price down, yet additional firms increase total expenditures on fixed costs. It can be shown (Appendix 7A) that, under some plausible conditions, there are too many firms in the monopolistic competition equilibrium. That is, welfare could be increased by restricting the number of firms.

By restricting entry, the government obtains the second-best optimum. Although welfare is not as high as in the first-best optimum, it is higher than in the unrestricted, monopolistic competition equilibrium. Table 7.3 shows the sum of consumer surplus and industry profits from Table 7.2, where $F = \$6.40$. The monopolistic competition equilibrium number of firms is eight, but the sum of consumer surplus and profits is maximized at three firms. By lowering the number of firms from eight to three, society reduces its expenditures on fixed cost (by $5F = \$32$) at the expense of a higher output price (46¢ instead of 36¢).

A Representative Consumer Model with Differentiated Products

The essence of the monopolistic competition model just discussed remains unchanged if all firms produce differentiated (heterogeneous) products. Profit maximization is still determined by $MR_r = MC$, and entry still occurs only so long as profits are positive. The only modification to the model of the previous section caused by product differentiation is that a firm's demand curve (and hence its MR_r curve) depends on the individual quantities produced by each of its competitors rather than on just the total quantity.

Adding product differentiation complicates the model. Each firm's demand curve may differ from another's so that it may not be sufficient to study a representative firm. It is possible, however, that although products are differentiated, the general form of the demand curves facing each firm is identical.

For example, all the firms in the industry could have demand curves of the form of Equation 7.4 where, due to product differentiation, a firm's price is more sensitive to changes in the quantity of its own product than to those of its competitors:

TABLE 7.3 Second-Best Optimum

Number of Firms	Price (¢)	Firm Output	Industry Profits ($)	Consumer Surplus ($)	Welfare ($)
1	64	360	123.20	64.8	188.00
2	52	240	102.40	115.20	217.60
3	46	180	78.00	145.80	223.80
4	42.4	144	57.34	165.89	223.25
5	40	120	40.00	180.00	220.00
6	38.3	103.9	25.08	190.34	215.42
7	37	90	11.90	198.45	210.35
8	36	80	0.00	204.80	204.80

Note: Parameters are the same as in Table 7.2, with fixed costs of $6.40.

$$p_i = a - b_1 q_i - b_2 \sum_{j \neq i} q_j, \tag{7.8}$$

where $\sum_{j \neq i} q_j$ means the sum of the output of all firms except Firm i.

The representative firm model with homogeneous products can be modified to handle this demand curve, and many of the qualitative results are the same as in the homogeneous model. For example, as each firm's fixed cost falls, the number of firms in the industry increases, and price may fall.

The primary impact of differentiation is that each firm faces a more steeply downward-sloping demand curve than it does otherwise, because other products are less close substitutes. This greater slope gives the firm more market power—the power to raise price profitably above marginal cost. See Example 7.3 on entry and product differentiation in the jeans market.

Welfare with Differentiated Products. The optimal welfare solution changes when products are differentiated.[12] In general, a monopolistic competition equilibrium with differentiated products has two problems: Neither the price nor the *variety* (number of brands) is optimal. As before, price is above marginal cost. However, there may be either too little or too much variety where products are differentiated.[13]

Two factors determine the variety in a monopolistic competition equilibrium. One of them leads to too few brands, but the other may lead to too many brands. The first factor is that highly desirable products may not be produced even though price is greater than firms' variable costs if fixed costs are so great that firms lose money. That is, consumer surplus would rise if more products were produced, but the high fixed costs keep the number of brands below the optimal level.

The second factor—the effect on other firms—is an offsetting force. When a firm introduces a new brand, it ignores the effect of its increased competition on the profits of other firms. When its product is a *substitute* for other brands, as Coke is for Pepsi, part of its profits come from these other brands. Because firms ignore these effects on other firms, they have a tendency to produce too many products at too low prices.[14] Because the two factors work in opposite directions, there may be too many or too few brands compared to the social optimum.

[12] Probably the first, and certainly among the best studies of welfare with differentiated products are Spence (1976) and Dixit and Stiglitz (1977). These models have been criticized by Pettingill (1979) and Koenker and Perry (1981), respectively. This section and the corresponding Appendix 7A are based, in part, on these articles and on unpublished lecture notes of Steven C. Salop, whom we thank.

[13] There is an analogous literature on the optimal amount of variety chosen by a single firm (Katz and Shapiro 1985; Farrell and Saloner 1985, 1986). A firm has to trade off the gains from standardization (such as economies of scale and compatibility with different manufacturers' products) against the benefits from variety.

[14] However, if products were *complements* such as bread and butter, there would be a tendency to have too few brands, with some prices too high because firms fail to account for the positive effect of their low prices and brands on the demand for other complementary products. Henceforth we assume brands are substitutes.

EXAMPLE 7.3

The Jeans Market

A lot of money is spent on jeans—and more every day. In 1996, sales of jeans in the United States grew 8 percent to $10.6 billion.

As the size of this market has ballooned over the last several decades, many firms have entered what used to be largely Levi's market. According to a survey of teenagers in the fall of 1996, 56% bought Levi's; 29%, Lee; 27%, Arizona; 21%, Guess; 19%, Gap; 18%, Calvin Klein; 16%, Bugle Boy; 15%, Wrangler; 13%, Union Bay; and 9% each, Chic and Tommy Hilfiger. Encouraged by the Gap's success and J. C. Penney's Arizona jeans, other large retailers started aggressively pushing their own private label jeans, such as Sears' Canyon River Blues.

In addition to entry by many new firms, greater product differentiation is occurring. Calvin Klein, Ralph Lauren, Donna Karan, and Tommy Hilfiger are spending substantial sums on promoting their designer jeans. Small startup firms, many based in Los Angeles, such as JNCO and Menace, sell offbeat cuts to younger consumers. These designer jeans sell for a premium over plain blue jeans.

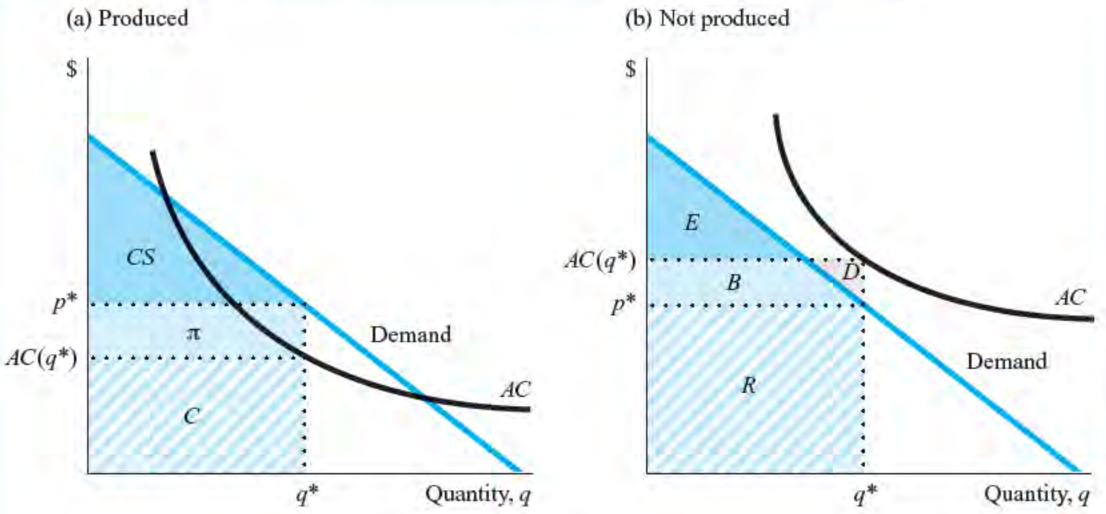
Entry of new firms and product differentiation are important forces in this market. Increased competition has hurt the giant, Levi Strauss, which laid off about a thousand workers in 1997 after its share of the market fell 5 percentage points from the previous year, and suffered a drop in U.S. sales from $5.1 billion in 1999 to $4.1 billion in 2002. The other two largest jeans producers, Lee and Guess, also are shifting resources from manufacturing jeans. To combat their new rivals, the big three all engaged in new product differentiation and other actions to boost sales and profits.

Source: Jennifer Steinhauer, "Squeezing into the Jeans Market," *New York Times*, March 14, 1997:C1, C15; Alexandra Jardine, "As Levi's Celebrates Its 150th Birthday," *Marketing*, September 4, 2003.

If goods are homogeneous, as in the previous example, there are definitely too many firms because there is no benefit to having more than one firm that is regulated to set price equal to marginal cost, assuming such regulation is possible. However, variety is desirable with differentiated products. Thus, regulating the markets so that there is only one firm charging marginal cost is unlikely to be optimal. The following section considers this analysis in more detail, first illustrating that fixed costs tend to result in underproduction of certain types of goods, and then discussing how the optimal number of brands is determined.

Fixed Costs Lead to Too Little Variety. When firms operate in the increasing–returns-to-scale section of their average cost curves, they tend to produce too few products, all else the same. If a firm's marginal cost does not rise rapidly, and it has

FIGURE 7.3 When Does a Market Produce a Product?


(a) Produced

(b) Not produced

$CS + \pi + C$	Social Benefit ($CS + R$)	$E + B + R$
$CS + \pi$	Welfare ($CS + R - C$)	$E - D$
CS	Consumer Surplus	$B + E$
$\pi + C$	Revenue	R
C	Cost	$R + B + D$
π	Profit	$-(B + D)$

large fixed costs, it operates in the downward-sloping, or increasing-returns, section of its average cost curve. Figure 7.3 illustrates why some desirable products are produced and others are not when the average cost curve is strictly falling.

In both diagrams in Figure 7.3, society is better off if the products are produced: Social benefit exceeds the social costs. In Figure 7.3a, the average cost crosses the demand curve, so it is profitable to produce. The firm's profit, π, is positive at quantity q^* because the average cost per unit is less than average revenue or price, p^*. Social benefit (the sum of consumer surplus, *CS*, and revenue, $\pi + C$) minus social (and private) cost (C) equals welfare ($CS + \pi$), which is positive.

In Figure 7.3b, the average cost curve is everywhere above the demand curve. Thus, total costs exceed total revenues at all output levels, so the product is not produced. It is, however, socially desirable to produce this product. Social benefit (consumer surplus, $E + B$, plus revenues, R) minus cost ($R + B + D$) equals welfare ($E - D$), which is positive because area *E* is greater than area *D*.

The reason the product is not produced, even though it is socially desirable to do so, is that the firm does not obtain the entire social benefit even though it pays the entire social cost. That is, the firm ignores consumer surplus when it makes its decision whether or not to produce. It would suffer a loss (negative profit, $B + D$) if it produced. Most customers would enjoy consumer surplus (the amount by which the product is worth more than p^*) if it were sold; whereas, the firm's price, p^*, is the value the marginal consumer (the one who has no consumer surplus) places on the good.[15] Thus, the example in Figure 7.3b shows that firms may not find it profitable to produce all goods that are socially desirable.

The product that is most likely to be produced is one for which the demand curve is a right angle: Consumers have an inelastic demand up to a cutoff price, p^*, at which their demand becomes perfectly elastic. With such a demand curve, there is no difference between total revenue and total social benefit, because there is no consumer surplus at price p^*. The firm's decision to produce or not is identical to society's criterion of total social benefit. Thus, all else the same, the smaller the ratio of consumer surplus to total revenues, the more likely is a firm to produce a socially desirable good.[16]

The crucial point is that this distortion—the underproduction of certain products—is due to the presence of fixed costs and the firm's inability to capture consumer surplus. For example, if there are no fixed costs and constant marginal costs, then average cost equals marginal cost. With constant marginal costs and no fixed costs, if it is socially optimal for a product to be produced, it pays for firms to produce it.

Optimal Diversity. The optimal equilibrium reflects the trade-off between product *variety,* the number of brands, and the *quantity* of each brand produced, which is determined by the price. For simplicity, assume that the number of brands, n, fully reflects the value of variety: The more firms or brands, the better off are consumers, all else the same. If all goods are produced with the same cost function and face the same demand curve, then the number of units of output, q, is the same for each brand in equilibrium. The essential facts about the equilibrium can be summarized by the number of brands, n, and the output per brand, q.

To illustrate the trade-off between variety and quantity, suppose the economy has 100 units of input, each unit of output can be produced at a constant MC of 1, and the fixed cost is 5. Table 7.4 shows some possible combinations of number of brands

[15] If a firm could perfectly price discriminate (Chapter 10), it could capture the entire consumer surplus. That is, it would charge each consumer the maximum that consumer would pay for the product, so that there would be no consumer surplus. Because its revenues would be larger than costs, the firm would find it profitable to produce. See also Romano (1991).

[16] For constant elasticity (ϵ) demand curves, $q = p^{-\epsilon}$, where $\epsilon > 1$, the higher the elasticity, the smaller the ratio of consumer surplus to revenues. Revenues are $R \equiv pq = p^{1-\epsilon}$, and consumer surplus is

$$CS = \int_p^\infty s^{-\epsilon} ds = \frac{p^{1-\epsilon}}{\epsilon - 1}.$$

Thus, the ratio of consumer surplus to revenues, $CS/R = 1/(\epsilon - 1)$, is decreasing in ϵ.

TABLE 7.4 Variety and Quantity

Number of Brands, n	Quantity of Each, q
1	95
2	45
3	28.33
4	20
5	15
6	11.67
7	9.29
8	7.5
9	6.11
10	5

and quantity (n, q). The **production possibility frontier** (*PPF*) is the feasible combinations of number of brands and quantity per brand that can be produced with society's total inputs (Figure 7.4 and Table 7.4).[17]

Society's preferences concerning the choice between quantity and variety are summarized by the indifference curves shown in Figure 7.4. Point $O = (q^*, n^*)$, the tangency between the *PPF* and an indifference curve, represents society's optimal choice. At any point on any indifference curve that lies below the indifference curve through point *O*, society is worse off. Points on indifference curves that lie above point *O* are above the *PPF* and hence cannot be produced. The point *B* on the *PPF* represents a possible monopolistic competition equilibrium. At that point, the industry is producing too few products, but more output per product than at the optimum. At point *A* on the *PPF*, the industry is producing more brands than at the optimum, but less output per brand.

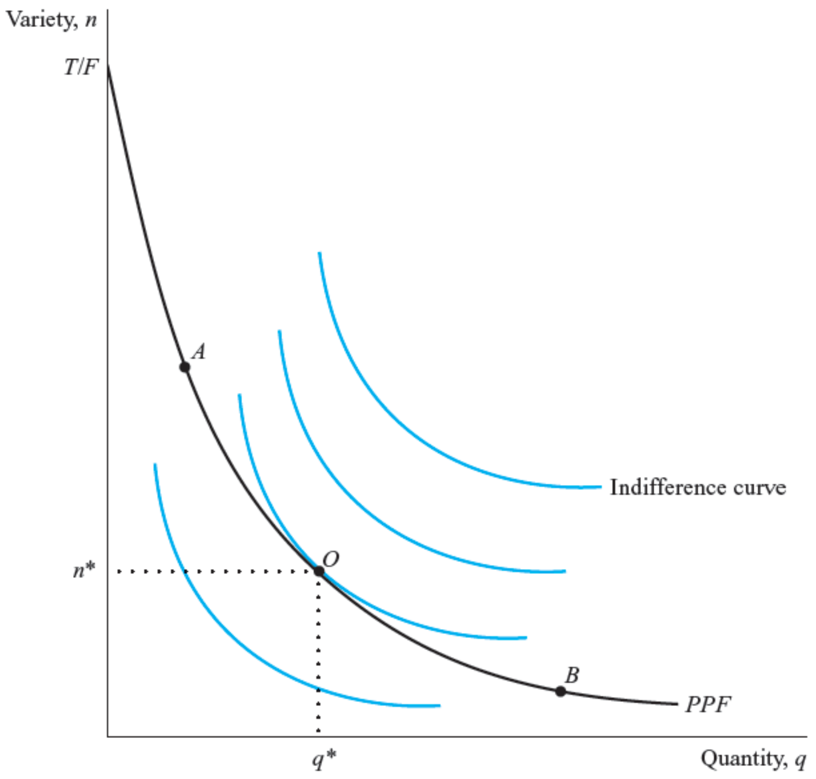
Whether the monopolistic competition equilibrium is at a point like *A*, *B*, or *O* depends on the preference of the representative consumer and the production function. Appendix 7B discusses the factors that determine the relative position of the monopolistic competition equilibrium. In general, any of these outcomes is possible.

Conclusions About Representative Consumer Models

In the Chamberlinian representative consumer monopolistic competition equilibrium, price is too high and the number of firms is nonoptimal. With undifferentiated products, there are almost certainly too many firms. With differentiated products, there may be too many or too few firms.

[17] As Appendix 7B shows, the *PPF* in Figure 7.4, which equates total cost to total resources, is $(F + mq)n = (5 + q)n = 100$, where $F = 5$ is the fixed cost and $m = 1$ is the constant marginal cost. Equivalently, the *PPF* is $n = 100/(5 + q)$.

FIGURE 7.4 Optimal (*O*) and Monopolistically Competitive (*A* and *B*) Equilibria



Typical representative consumer models assume that all products are equally good substitutes for each other. To apply such a model to the ice cream market, for example, one must believe that Breyer's ice cream competes equally with Häagen-Dazs and with Baskin-Robbins. That is, one must not believe that Breyer's is a closer substitute to Baskin-Robbins than to Häagen-Dazs. This extremely strong assumption makes the model relatively easy to use, but unrealistic in some markets.

Location Models

Brands compete more vigorously with brands that are close substitutes than with those that consumers view as less close substitutes. Consumers view certain brands as closer substitutes than others. For example, certain brands have a particular common characteristic that other brands lack: Some cereals are sugar coated and others are not. That is, each brand is "located" at a particular point in product characteristic space. As another example, products sold at nearby stores are close substitutes. That is, each firm is located at a particular address or point in geographic space.

Location *(spatial)* models are monopolistic competition models in which consumers view each firm's product as having a particular location in geographic or product (characteristic) space. The closer two products are to each other in geographic or characteristic space, the better substitutes they are. In these models, consumers also have locations in geographic or product space. It costs consumers more to shop at stores farther from home, or, alternatively, they receive less pleasure from products whose characteristics deviate from their ideal. Because firms or products only compete directly with others near them, each has some market power. The market power stems from the preference of consumers to make a purchase at the nearest firm or to purchase their preferred product.

The following discussion first examines the original location model and then uses a newer location model to analyze the impact of increased competition on the market equilibrium. Finally, the welfare implications of this equilibrium are analyzed.

Hotelling's Location Model

Hotelling (1929) developed a model to explain the location and pricing behavior of firms.[18] Although he concentrated on geographic space, his model can be used to study monopolistic competition by viewing products as being located in product or characteristic space. In Hotelling's location (spatial) model, products differ in only one dimension, such as the location of the stores that sell them. However, Lancaster (1966, 1971, 1979) and others have shown that this model can be extended to examine products that differ in more dimensions.

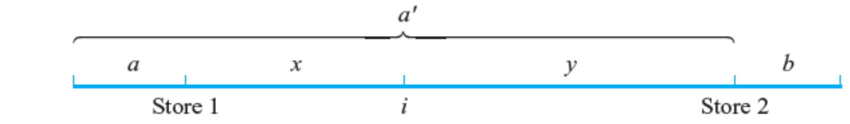
Consider a long, narrow city with only one street, Main Street, that is a fixed length. Consumers are uniformly distributed along this street, so that in any block there are an equal number. All consumers are identical except for location, and each consumer buys 1 quart of milk in each time period.

Two stores sell identical bottles of milk in this town. Store 1 is located a miles from one end of town (the left end in Figure 7.5), and Store 2 is located b miles from the other (right) end of town. Consumers have no preference for either store except that consumers prefer to purchase from the nearest store because each consumer faces a transportation cost of c per mile. That is, each consumer buys from the least expensive store, taking transportation costs into account. Consider Consumer i who lives at the location shown in Figure 7.5. She lives x miles from Store 1 and y miles from Store 2. Because x is less than y (see Figure 7.5), she goes to Store 1 to minimize her transportation costs. Only someone who lives exactly halfway between the two stores is indifferent as to which store to patronize.

Suppose that the government sets the price of milk. How should Store 1 choose its location to maximize its profits if Store 2 is already located b miles from the right end

[18] See also, Eaton (1976), D'Aspremont, Gabszewicz, and Thisse (1979), Novshek (1980), and Friedman (1983). For analytic simplicity, in the literature most representative consumer models assume firms play Cournot and most location models assume they play Bertrand, but either oligopoly concept can be used in either model.

FIGURE 7.5 Hotelling's Town



of the city and cannot change its location? Because consumers only care about how far they must travel, Store 1 wants to be the nearest store for the greatest possible number of consumers. Store 1 maximizes its profits by locating just to the left of Store 2, a' miles from the left end of the city. There, it gets all the customers to its left, which is the majority.

If Firm 2 could costlessly relocate after Firm 1 locates, however, it would move slightly to the left of Firm 1's new location. This process would be repeated until both firms were in the middle of the town, with each firm having half the customers. You may have noticed the propensity of firms to locate near each other in a variety of markets. For example, several gas stations often locate on the corners of a busy intersection.

Thus, if price is given, the location of two firms can be determined. This equilibrium is Nash in location strategies (see Chapter 6). That is, when firms are set at their equilibrium locations, no firm wants to change its location. Similarly, by fixing location and letting the firms vary prices, a Nash equilibrium in prices can be determined (similar to the Bertrand equilibrium discussed in the previous chapter).

Hotelling's model illustrates an important point: The properties of the Bertrand equilibrium discussed in the previous chapter hold only when two firms sell perfectly homogeneous products. In the Bertrand model with homogeneous products, if one firm undercuts the other, the high-priced firm loses all its customers. The same thing happens in Hotelling's town if both firms are permanently located in the center of town.

However, suppose that the two stores are permanently located some distance apart at a and b in Figure 7.5. If Store 1 charges less than Store 2, Store 2 still gets a number of customers. The reason is that Store 2 is much closer for several customers than Store 1, and some shoppers will pay more for the convenience.

Thus, Hotelling's model illustrates that the Bertrand equilibrium price equals the marginal cost only if the products are homogeneous (located at the same place in product or geographical space). In a more general model of differentiated products, firms with Bertrand expectations may charge different prices and all prices are above marginal cost.[19] In short, differentiation gives firms market power.

Unfortunately, it can be shown that when firms can costlessly change their prices *and* their locations (for example, reformulate their product), there is a *nonexistence of*

[19] Mergers of a subset of the firms in the industry have no effect in a Bertrand model with homogeneous goods, but are profitable for the merging firms in a Bertrand model with heterogeneous goods (Deneckere and Davidson 1985). Compare this result to those in Example 6.3.

equilibrium (D'Aspremont, Gabszewicz, and Thisse 1979).[20] This result is analogous to the Edgeworth example in the previous chapter, in which the two firms continuously change their behavior, never settling down to a single price (and location). The existence of an equilibrium, however, can be shown in modified versions of this model. One modification allows for nonlinear transportation costs. Another approach is studied next.

Salop's Circle Model

A circle is the longest distance to the same point. —*Tom Stoppard*

A number of models modify Hotelling's basic model so that an equilibrium exists. One of the most interesting and best known of these is Salop's (1979a) circle model, which introduces two major changes in Hotelling's model.

First, in this model, firms are located around a circle instead of along a line. The reason for this change is that a circle has no end-points. That is, a circle is roughly equivalent to an infinitely long line in that neither has end-points. It can be shown that a major cause of the nonexistence of equilibrium in Hotelling's model is the presence of end-points.

Second, Salop's model takes explicit account of a second, or outside, good. For example, the differentiated product might be brands (flavors) of ice cream (the products located around the circle), and the outside good might be chocolate cake, which is an undifferentiated product competitively supplied by another industry.

How Consumers Choose a Product. Assume that customers are uniformly located around the circle that is of unit circumference. For simplicity, each customer buys exactly one scoop of ice cream. A customer's location, t^*, represents that customer's most preferred type of ice cream. For example, suppose one location on the circle is chocolate ice cream, another vanilla, and a point between chocolate and vanilla is chocolate-chip ice cream. Each flavor of ice cream is a possible brand and is described by its location on the circle.

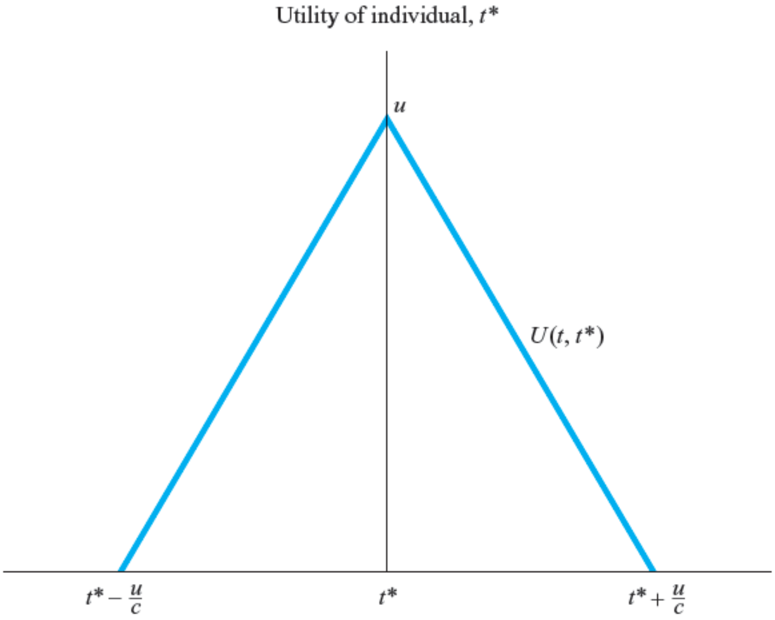
The pleasure (utility) a consumer gets from eating a scoop of a brand of ice cream located at t is

$$U(t, t^*) = u - c|t - t^*|, \tag{7.9}$$

where u is the utility from the consumer's favorite flavor of ice cream (the flavor located at the same point, t^*, along the circle as the consumer); $|t - t^*|$ (the absolute value of the difference between t and t^*) is the distance brand t is from the customer's favorite flavor t^*; and c is the rate at which a deviation from the optimal brand lowers the consumer's pleasure.

[20] A randomized (mixed strategy) equilibrium exists, where each firm chooses its action probabilistically.

FIGURE 7.6 Consumer's Utility Function



The consumer's utility function is shown in Figure 7.6, where a segment of the circle has been straightened out into a line. The figure shows that at $t = t^* + u/c$ and at $t = t^* - u/c$ the consumer has a utility of zero. The figure shows that the pleasure a consumer receives from a brand located either to the left or to the right of the optimal brand is lower than from the optimal brand.

Each consumer attempts to maximize consumer surplus, which is the difference between the consumer's pleasure from eating a brand located at t and the price: $U(t, t^*) - p$. In other words, if your favorite flavor of ice cream is chocolate, but chocolate chip ice cream costs half as much, you might buy the chocolate chip because the loss in taste or utility is less than the gain from buying the cheaper product. Thus, you purchase the *best buy:* the product with the greatest surplus—the best combination of price and quality.

Instead of buying one of the brands of ice cream, however, the consumer may decide to buy the outside good, chocolate cake, if it is a *better buy* in the sense that it gives more pleasure for a given amount of money. Suppose the surplus from the cake (pleasure from eating it less the price) is $\underline{u}$. The consumer only buys a scoop of the best-buy brand, i, of ice cream if its surplus is at least equal to $\underline{u}$:

$$\max_i \,[U(t_i, t^*) - p_i] \geq \underline{u}, \tag{7.10}$$

where the expression on the left side of the equation is the surplus from the best-buy brand of ice cream (maximize the surplus through choice of brand i), and the right side is the surplus from cake. That is, the consumer should only buy ice cream if the surplus from the best-buy brand of ice cream is at least as great as the surplus from cake.

If a consumer's ideal ice cream is produced (located at t^*) and sold at p^*, the greatest surplus the consumer can get is $u - p^*$. The consumer is only willing to buy that brand if its surplus is equal to or greater than that from cake: $u - p^* \geq \underline{u}$, or, rearranging terms, $u - \underline{u} \geq p^*$. As a result, the consumer has a *reservation price*, $v = u - \underline{u}$, which is the highest price that the consumer is willing to pay for that brand of ice cream.

Alternatively stated, a consumer buys a scoop of ice cream only if the *net surplus* from the best-buy brand, the surplus from the best-buy brand minus the surplus from cake, is positive:

$$\max_{i}[v - c\,|t_i - t^*| - p_i] \geq 0. \tag{7.11}$$

Equation 7.11 is obtained by subtracting u from both sides of Equation 7.10, substituting for $U(t, t^*)$ from Equation 7.9, and using $v = u - \underline{u}$.

Firms' Behavior. The symmetric equilibrium in this model depends on where firms are located and how they set price.[21]

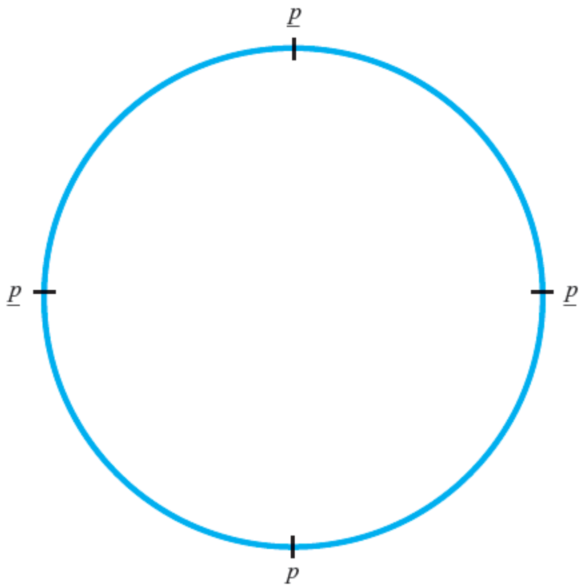
All else the same, each firm wants to locate as far from its nearest competitors as possible. The further away other stores are from your store, the greater the market power you have with respect to the customers located near your store. As a result of trying to locate as far apart as possible, the stores locate equidistant from each other. If there are n ice cream brands located at equal distances around the circle, the distance between two brands is $1/n$ (because the circle is of unit circumference).

Salop starts his analysis by assuming that the stores are already located equidistant from each other and then asks what price each store charges. Suppose a typical brand (the one at the bottom of the circle) charges price p, and its two nearest competitors charge $\underline{p}$, as Figure 7.7 shows. How should the producer of the typical brand set price? The answer depends on how many brands there are. We first consider the case in which there are relatively few firms, and then consider a market with many more firms.

Monopoly Region. If there are relatively few brands, they do not compete with each other for the same consumers. Each brand is a local monopoly and sells to all consumers living close enough so that their net surplus is positive. That is, each monopoly sells only to consumers who receive more surplus from that brand than they get from cake.

[21] Economides (1986, 1989) examines existence of the full subgame-perfect equilibrium in the Hotelling, Salop, and two-dimensional space of characteristics models. His intuition is that the equilibrium price does not converge to marginal cost as the locations of two competing firms become nearly identical in the Hotelling model, so there is a strong tendency to undercut a rival's price when the locations are very close (see D'Aspremont, Gabszewicz, and Thisse 1979). If utility is quadratic in distance (or for a two-dimensional space even with linear utility), however, prices do converge to marginal cost as locations become nearly identical, which eliminates the undercutting and nonexistence problem.

FIGURE 7.7 Circular Market



Consider a consumer located a distance $x = |t - t^*|$ from the brand at t with price p. The consumer is willing to buy that brand only if the consumer's net surplus is nonnegative: $v - cx - p \geq 0$ (using the expression for surplus in Equation 7.11). Thus, by rearranging this expression, the maximum distance, x_m, a consumer can be located from that brand and still buy it is

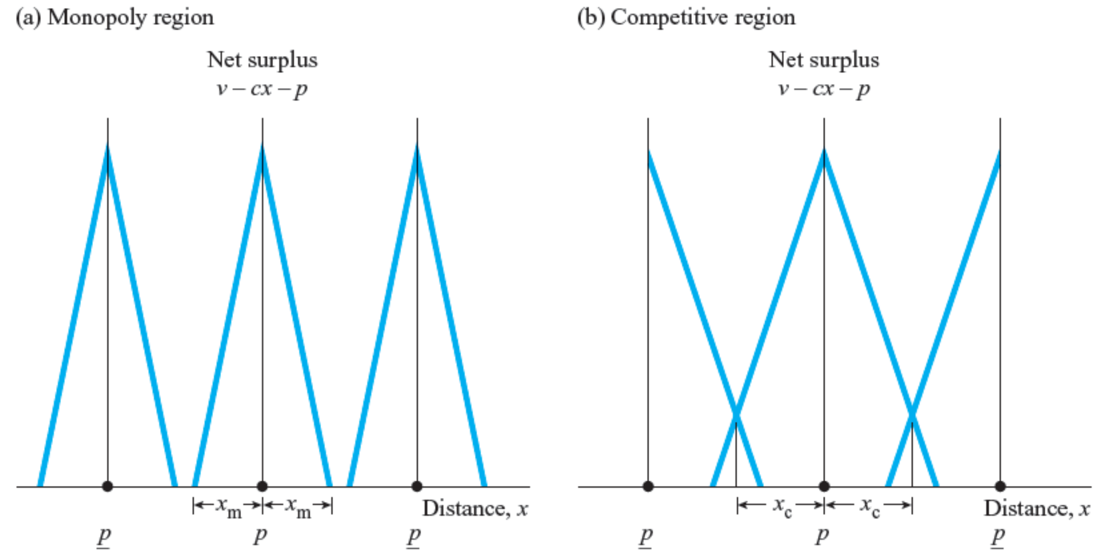
$$x_m = \frac{v - p}{c}. \tag{7.12}$$

This distance, x_m, is determined graphically in Figure 7.8a. The vertical axis in the figure is the net surplus from that brand and the horizontal axis is the distance, x, a consumer is from the most preferred brand (labeled with the price, p, which is assumed to be slightly more than $\underline{p}$) . The greater the distance, x, a brand is from the consumer's most preferred product, the lower the consumer's net surplus. When the brand is x_m distance from the consumer's most preferred location, the consumer's net surplus from that brand equals zero (where the net surplus line hits the x-axis) so that the consumer is indifferent between buying and not buying.

The brand captures all the consumers who are no further than x_m distance on each side of its location, or all the consumers in a $2x_m$ segment of the circle. If there are L consumers located uniformly around the circle, the monopoly demand facing this brand, q_m, is $2x_mL$, or, substituting for x_m from Equation 7.12:

$$q_m = \frac{2L}{c}(v - p). \tag{7.13}$$

FIGURE 7.8 Two-Market Structure



The monopoly quantity demanded of the firm, as shown in Equation 7.13, falls by $-2L/c$ as its price rises by $1. If the firm sets its price equal to the reservation price, v, of the customer who most prefers this product, its sales fall to zero.

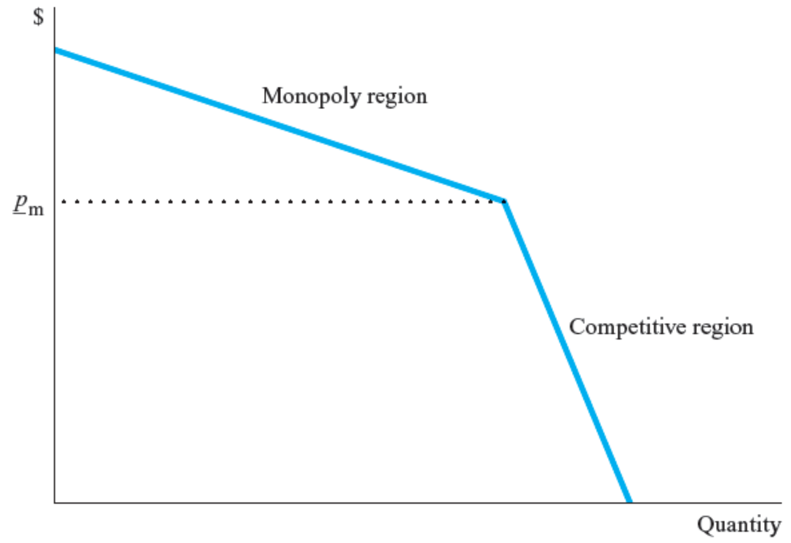
Competitive Region. If there are more firms, so that they are located closer together and compete for the same consumers, then each firm must take into account the price its rivals charge in setting its own price as in the homogeneous-good Bertrand model of Chapter 6. When firms compete with each other, a firm does not capture all the customers who prefer its ice cream to cake: It loses some to its two nearest rivals. Those customers located in the potential market of each of two brands buy from the one offering the highest net surplus.

Both of the typical brand's closest competitors are $1/n$ distance away and charge $\underline{p}$. How much does this brand sell if it sets its price at p? It captures all the consumers within a distance x_c, where x_c is the distance such that consumers get the same utility from this brand as from that of one of its closest rivals:

$$v - cx_c - p = v - c\left(\frac{1}{n} - x_c\right) - \underline{p}. \tag{7.14}$$

The left side of Equation 7.14 is the net utility from this brand, and the right side is the net utility from the other brand (because a consumer who is a distance x_c from this brand is $1/n - x_c$ distance from the rival brand). Figure 7.8b shows how the limit of the competitive region, x_c, is determined by the point where a consumer is just indifferent between the two goods—where Equation 7.14 holds with equality. Where the net surplus lines from two rival brands intersect, a consumer is indifferent between buying either brand.

FIGURE 7.9 Demand in Salop's Circle Model



Solving Equation 7.14 for x_c, and noting that the quantity demanded of a competitive firm is $q_c = 2x_cL$, the competitive demand equation is

$$q_c = \frac{L}{c}\left(\frac{c}{n} + \underline{p} - p\right). \tag{7.15}$$

Thus, the competitive quantity demanded falls by $-L/c$ as p rises by \$1 (holding $\underline{p}$ constant). That is, the slope of the competitive demand curve is only half as steep as that of the monopolistic demand curve.

Types of Equilibria in the Circle Model. At high prices, the demand regions of the firms do not overlap. Each firm is a local monopolist. As the price falls, so that more consumers are interested in ice cream, the regions overlap, and competition between the firms begins. The monopoly and competitive demand regions are shown in Figure 7.9. At prices above $\underline{p}_m$ the demand region is monopolistic: A brand's customers do not consider buying any other brand. Below $\underline{p}_m$ the brand competes with its nearest neighbors.[22]

[22] As Equation 7.14 shows, at prices below $\underline{p} - c/n$, all of the customers between a given firm and its neighbor prefer buying from the low-priced firm, even if they do not like that brand as much. Consider the consumer located at the same point as the neighbor firm: a distance of $1/n$ from the low-priced firm. That consumer loses c/n utility by consuming the low-priced brand rather than the preferred brand. But the price saving is greater than that loss. This type of behavior is extremely aggressive (Salop calls it *supercompetitive*), because the low-priced firm captures all the new neighbor brand's contested customers. For simplicity, Figure 7.9 does not show this region.

Salop shows (in an argument analogous to the one used for the representative consumer models) that where firms have constant marginal and fixed costs, there exists a symmetric Nash equilibrium in which no firm wants to alter its price and no additional firms want to enter. That is, all firms charge the same price in equilibrium and are located $1/n$ distance from each other. Suppose that free entry is allowed and that firms can costlessly relocate so that they are equidistant from each other. Then, in a monopolistic competition equilibrium, the entry of one more firm causes all firms' profits to be negative. Example 7.4 discusses what happens when firms cannot costlessly relocate as new entry occurs.

EXAMPLE 7.4 *A Serial Problem*

In 1972, the U.S. Federal Trade Commission (FTC) charged the four largest U.S. manufacturers of ready-to-eat breakfast cereal (RTE cereal) with several antitrust violations, including conspiring through brand proliferation and differentiating similar products to prevent entry into the industry. Although the FTC failed to win its case, this argument is theoretically interesting.

Richard Schmalensee and F. M. Scherer used localized competition models to explain the FTC's argument. In such models, consumers choose cereals based on their characteristics, such as sweetness and "mouth feel." Each brand is located in a characteristic space. A given brand must compete for customers with other nearby brands located in that part of product space. If the company owning that brand can surround it with other similar brands of its own, then its brands compete with each other. For example, Kellogg's Corn Flakes and Special K may be very close substitutes.

If a firm creates enough of these surrounding, or *defensive*, brands, there may not be enough customers left for any other firm to profitably establish a brand in that area of product space, according to this brand-proliferation theory. Similarly, several firms could conspire to establish a large number of brands (more brands than are profit maximizing in the short run) in a given area of product space to prevent entry by new firms.

Whether the firms were conspiring or not, the top six firms had 95 percent of the sales of cereal. Moreover, between 1950 and 1972, the six leading producers introduced over 80 brands into distribution beyond test marketing.

In the early 1970s, however, "health" cereals started selling well. Because existing firms had not located in this area of product space, new firms (which included such giants as Colgate, International Multifoods, Pet, and Pillsbury) were able to enter. By mid-1974, these "natural" cereals had 10 percent of the market. But apparently, the previous positioning of these new firms did not prevent the established firms from entering this section of product space. As a result of the entry of the established firms and the decline in demand for that segment of the market from its 1974 height, all but one of the new entrants (Pet) was driven from that area of product space by late 1977.

Sources: Schmalensee (1978b), Scherer (1979), and (for a different view) Williamson (n.d.).

Changes in Costs and Welfare in the Circle Model. Salop shows that, as in the representative consumer models, in the competitive region, as fixed costs rise, there are fewer firms or brands, so price rises and equilibrium variety falls. In that region, as the constant marginal cost rises, price rises by an equal amount (all increases in costs are shifted to consumers), but equilibrium variety remains unchanged.

At the kink in the demand curve at $\underline{p}_m$ in Figure 7.9, however, an increase in either fixed or marginal cost reduces the number of firms (variety) but, perversely, lowers price (the kink shifts down and to the right in Figure 7.9). Thus, if the economy is at such a point, a tax that raises firms' costs lowers prices and decreases variety. Salop shows, however, that welfare rises, even if the proceeds of the tax are ignored.

Indeed, welfare in this circle market can be studied in the same manner as in the representative consumer models. Salop shows that the first-best optimal variety is less than the variety in either the monopolistic or competitive equilibria. With fewer brands, the savings in fixed costs exceed the losses due to higher prices. Thus, in the circle model's monopolistic competition equilibrium, there are unambiguously too many brands; whereas in the differentiated products, representative consumer model equilibrium, there can be too many or too few brands.

Salop shows that the second-best optimum, given the government's only regulatory policy is to control entry, is either the market equilibrium or complete monopoly. That is, the optimal entry policy is either free entry or entry restricted so that each brand has a complete monopoly market.

Hybrid Models

We have drawn the distinction between representative consumer models and location models. Although the vast majority of all monopolistic competition models fall cleanly into one or the other of these two categories, increasing use is made of hybrid models that combine some of the properties of each model.[23]

One of these hybrid models, Deneckere and Rothschild (1986), includes the circle model and a version of the representative consumer model as special cases of Bertrand equilibrium. Using their hybrid model, Deneckere and Rothschild show that prices are lower in a representative consumer model than in the circle model because there is more competition in a representative consumer model. They also show that adding another brand benefits relatively few consumers in the circle model, whereas consumers benefit substantially from the introduction of extra brands in the representative consumer model. It is for these reasons that there are too many brands in the equilibrium of the circle model, but there may be too many or too few brands in the equilibrium of the representative consumer model.

[23] Hybrid models based on the work of Anderson and de Palma (1992a, 1992b), Besanko, Perry, and Spady (1990), Deneckere and Rothschild (1986), Perloff and Salop (1985), and Sattinger (1984) are discussed in **www.aw-bc.com/carlton_perloff** "Hybrid."

Estimation of Differentiated Goods Models

The properties of the differentiated goods models we have described depend critically on the pattern of substitutability among the products. In recent years, advances in statistical techniques and more powerful computers have made possible the simultaneous estimation of demand functions for many brands within a single market.

Nonetheless, it remains difficult to simultaneously estimate the demand curves facing all the differentiated products in a market because of the large number of parameters that must be estimated. For example, there are at least 174 firms selling 230 brands and 613 different products of canned juices in U.S. grocery stores. If we had to estimate 613 simultaneous equations of the quantity demanded for each good conditional on 613 prices with one coefficient for each price in each equation, we would have to estimate at least 375,769 parameters. Even if we use utility theory to restrict these parameters, we would still have to estimate many thousands of parameters.

To reduce the number of parameters that must be estimated, researchers use restricted demand systems in which they impose relationships among the various demand curves, some of which stem from theory and others that are arbitrary, such as the functional form. Analysts use many different functional forms when estimating demand equations, such as logit, nested logit, and the almost ideal demand system, but they all have the purpose of imposing constraints on the pattern of substitution among the differentiated products in order to limit the number of parameters that have to be estimated.[24] Further, researchers frequently estimate demand curves for only the major products within a market.

Two approaches are used frequently. The traditional approach is to estimate a system of demand curves—one for each product—but to impose enough restrictions so that the parameters in the demand curves can be estimated (for example, Hausman and Leonard, 1997). An alternative approach is to use a logit or the more general random parameter logit (or probit) model, in which the econometrician tries to explain the fraction of sales for each product conditional on its characteristics (such as, flavor and size of container) as well as relative prices.[25] Researchers must be careful in making assumptions about the structure of the demand curves that they do not implicitly assume their conclusions by the manner in which they restrict the substitution patterns.

After estimating the demand system, researchers frequently want to make "what if" prediction about how changes in a market (such as a merger) will affect prices. To do so, they need a model of oligopoly behavior. Researchers usually lack information about marginal costs. However, by making strong assumptions, they can estimate marginal costs.

[24] See, for example, Baker and Bresnahan (1988), Trajtenberg (1989), Hausman, Leonard, and Zona (1994), Bresnahan, Stern, and Trajtenberg (1997), Hausman and Leonard (1997), Hendel (1999), and Peters (2003).
[25] See Berry (1994), Berry, Levinsohn, and Pakes (1995), Nevo (2000, 2001), and Petrin (2002).

EXAMPLE 7.5 Combining Beers

Consumers can choose from among many different brands and types of beer, the latter including premium, light (low calorie), and imported. Postulating a demand curve for each brand and type of beer, an analyst would have to estimate a very large number of parameters. Instead, Hausman, Leonard, and Zona (1994) postulate that there is a three-stage decision process to beer consumption. First consumers decide how much beer to consume in the aggregate. Then they decide what share of each type to consume. Finally, they pick the share of each brand for each type of beer. By specifying demand in this way, the analyst is reducing the number of demand parameters at the cost of imposing restrictions on demand substitution patterns.

The analyst then uses the estimated demand elasticities from this demand system to determine the markup of price over marginal cost based on the assumption that the firms engage in a Bertrand game and have constant marginal costs. A logical question to ask is what will happen to prices if two of the firms merge. To answer this question, the analyst rewrites the profit-maximizing equations describing the equilibrium with fewer firms. Previously, the two firms set prices independently. Now the new merged firm takes into account that it controls the pricing of products previously set by the other firm. Using this technique, Hausman et al. calculate that a merger of Coors and Labatts, two brewers of premium beers, would raise the Coors price by 4.4 percent and the Labatt's price by 3.3 percent. However, they also conclude that, if the merger results in a 5 percent gain in efficiency so that marginal costs fall by 5 percent, the merger will lead to lower, not higher, prices.

Unfortunately, usually they do not know marginal cost, so they use the assumptions that the firms maximize profit, engage in a Bertrand game, and have a constant marginal cost, m, to estimate the marginal cost. For example, the Lerner Index markup of price over marginal cost for a profit maximizing monopoly is $(p - m)/p = -1/\epsilon$ (Equation 4.3). This equation shows that the price markup depends only on the elasticity of demand, ϵ. Moreover, given that we can observe p and have estimated the demand equation and have an estimate of ϵ, we can use this equation to estimate m. This profit-maximizing condition generalizes for any profit-maximizing firm, even in markets with many differentiated products, given that one knows the type of game, such as Bertrand, that the firms play and all own and cross-price elasticities. Even in these generalized expressions, the price markup depends only on the own and cross-price elasticities of demand of all the products, so that again one can infer the constant marginal costs (see, for example, Hausman et al. 1994). Using this approach, researchers predict the price effects of a merger (see Example 7.5) and the value of new products (see Example 7.6).[26]

[26] In a comparison of how the various methods worked in predicting the effect of airline mergers, Peters (2003) found that estimation of the demand system combined with assumed oligopoly behavior did not perform significantly better than a "reduced form" estimation that directly relates prices to concentration (which Peters treated as an endogenous variable).

EXAMPLE 7.6 *Value of Minivans*

In 1984, Chrysler introduced the first minivan, the Dodge Caravan. The minivan was a successful innovation because it handled like a passenger car even though it was substantially larger than one. General Motors and Ford soon introduced their own minivans. Minivans' success came partially at the expense of station wagons, whose sales fell by over 60 percent during the next seven years.

How much was the introduction of the minivan worth to consumers? To answer this question, Petrin (2002) estimated the demand curve for minivans so as to calculate the increased consumer surplus that consumers receive. But even consumers who do not purchase the minivan can benefit if the minivan simulates competition and lowers the price of the car they do purchase.

Petrin estimated a random coefficient discrete choice demand system, which was pioneered by Berry et al. (1995). Here, the demand estimation allows for differences in tastes across individuals. In this approach, the researcher assumes that each individual chooses that product that yields him or her the highest utility, and that firms maximize profits in competition with others in Bertrand competition. Even though Petrin lacked information on purchases by individuals, he had data on the average income of a buyer of each type of car, which he used to better estimate his model.

Petrin calculated the amount of money consumers would have had to receive in the absence of minivans in order to attain the same utility as when minivans did not exist. He estimated that the average benefit per consumer over a four-year period was $1,247 with over 40 percent coming as benefits to non-minivan purchasers from increased competition. Petrin calculated, if one ignored taste heterogeneity, that the benefit would have been overestimated to equal $13,652. His best estimates of the benefits to consumers of minivans totaled $2.8 billion over a four-year period. However, his estimates ignore the negative externality—more deaths—imposed by minivans and sport utility vehicles on drivers of normal size cars and pedestrians (White 2002).

SUMMARY

This chapter examines product differentiation and monopolistic competition. Product differentiation creates at least some market power for a firm. The greater the perceived difference between two firms' products, the more each firm can charge.

If free entry is allowed, firms enter markets until profits are driven to zero. A monopolistic competition equilibrium is one in which firms face downward-sloping demand curves and earn zero profits.

There are two basic types of monopolistic competition models. In Chamberlin's representative consumer model, a typical consumer views all products as equally good substitutes for each other. Price is above marginal cost, and there may be too much or too little variety. Entry (due, for example, to a reduction in fixed costs) tends to reduce the prices of all firms.

Hotelling's location (spatial) model postulates that consumers' preferences and brands are located in product or geographic space. Consumers prefer brands near them. As a result, firms have some market power. The pricing behavior of other firms has little effect if the consumers who buy from a given firm do not like the products of those firms. In the localized competition circle model, price is above marginal cost, and there is unambiguously too much variety. New entry does not lower the price a given consumer pays unless a firm enters near the firm that the consumer patronizes because consumers are uninterested in brands that are very dissimilar to the ones they like best.

PROBLEMS

1. Explain how, in a monopolistically competitive industry, high fixed costs can result in too little variety.
2. In the Salop circle model, if all consumers get more pleasure from ice cream (*u* increases), how does the equilibrium change?
3. Explain and illustrate the following claim: "In our example, a monopolistic competition industry with homogeneous products cannot be more than one firm away from the output sold at price equals marginal cost."
4. In Hotelling's town, if all firms are required to charge the same fixed price, describe the equilibrium location of three firms. Explain your answer. Now describe the equilibrium for four firms.
5. What is the effect of a cost-saving technological change on a monopolistic competition industry in which the cost curves facing each firm are $C(q) = mq + F$, where m is the constant marginal cost, and F is the fixed cost? *Hint:* A cost-saving technological change may be modeled as reducing *m,* reducing *F,* or reducing both.
6. Show graphically that if a firm's $MC = AC =$ a constant, it will produce a product if it is socially desirable for that product to be produced.

Answers to odd-numbered problems are given at the back of the book.

SUGGESTED READINGS

Friedman (1983) has a good survey and discussion of most of the models in this chapter. In the 1930s, there was a lively (and relatively nontechnical) debate between Chamberlin (1933) and Robinson (1934) and Kaldor (1935) concerning the necessary conditions for a firm to possess market power (the power to set price above marginal cost). For an excellent survey of the technical literature on product differentiation, see Eaton and Lipsey (1989).

APPENDIX 7A

Welfare in a Monopolistic Competition Model with Homogeneous Products

> *Why, a four-year-old child could understand this report. Run out and find me a four-year-old child. I can't make head or tail of it.* —Groucho Marx

Two problems arise in a monopolistic competition equilibrium with homogeneous goods:[1]

1. Because price is greater than marginal cost, the industry produces too little output.
2. If marginal cost is constant, the industry bears excess fixed costs.

First-Best Optimum

> *All is for the best in the best of possible worlds.* —Voltaire

Given constant marginal cost, the first-best optimum requires a single firm that charges a price equal to marginal cost, $p = m$, and a subsidy of the firm's losses. We illustrate this result using a simple, general equilibrium model. In this model, there is no important distinction between the partial and general equilibrium because the general equilibrium's income effect is the same as in the partial equilibrium.

The representative consumer's utility function is

$$U(Q, y) = u(Q) + y, \tag{7A.1}$$

where Q is the output of the monopolistic competition industry and y represents all other goods. Let y be produced at constant cost, and, by normalizing, let this constant cost equal 1 so that the competitive price is also 1.

The consumer maximizes his or her utility subject to the budget constraint

$$I = pQ + y, \tag{7A.2}$$

where I is the consumer's income and p is the price of a unit of Q. From Equation 7A.2, $y = I - pQ$. Substituting that expression for y into the consumer's utility function (Equation 7A.1), the consumer's utility maximization problem is

[1] Appendixes 7A and 7B draw heavily on Steven C. Salop's unpublished lecture notes, Dixit and Stiglitz (1997), and Spence (1976).

$$\max_{Q} u(Q) + I - pQ. \tag{7A.3}$$

The first-order condition for utility maximization is

$$u'(Q) = p. \tag{7A.4}$$

That is, the consumer picks Q so that marginal utility equals the marginal cost of Q, which is p. As a result, the consumer's demand function may be written as $p = p(Q) = u'(Q)$. Because marginal utility is positive, $p > 0$. The second-order condition, $u'' < 0$, implies there is diminishing marginal utility so that the demand curve is downward sloping: $p' < 0$.

If there are n identical firms in the Q-industry, each produces an equal amount of output, $q = Q/n$. The economy's resource constraint is

$$T = (nF + mQ) + y, \tag{7A.5}$$

where T is the total resources of the economy (maximum production), F is the fixed cost each Q-firm must sink to be in business, and $nF + mQ$ is the total cost of producing Q units of output. For example, if T is the total hours of labor available and y is leisure, then the total time spent producing output plus leisure equals T.

Society's problem is to maximize Equation 7A.1 subject to Equation 7A.5 through its choice of Q, y, and n. By substituting for y in Equation 7A.1 using Equation 7A.5, we may write this problem as

$$\begin{aligned} \max_{Q,\, n}\ & u(Q) + T - nF - mQ \\ \text{s. t.}\quad & n \geq 1, \\ & Q > 0 \\ & (p - m)\frac{Q}{n} - F \geq 0, \end{aligned} \tag{7A.6}$$

where the last condition is that each firm makes nonnegative profits so that they do not shut down. Equation 7A.6 says society should maximize the objective function, utility, $u(Q) + T - nF - mQ$, by choosing Q and n appropriately, subject to the restrictions that there is at least one firm ($n \geq 1$) and that some positive amount of Q is produced ($Q > 0$).

The Lagrangian may be written as

$$\mathcal{L} = u(Q) + T - nF - mQ - \lambda(n - 1) - \mu Q, \tag{7A.7}$$

with Lagrangian multipliers λ and μ. If any of the constraints are nonbinding (hold as a strict inequality), then the associated Lagrangian multiplier is zero.

The Kuhn-Tucker first-order conditions with respect to n and Q imply that[2]

$$n = 1, \tag{7A.8}$$

because $\mathscr{L}_{\text{n}} = 0$ implies that $-F = \lambda$, and

$$u'(Q) = m, \tag{7A.9}$$

because $Q > 0$. Thus, as shown in Figure 7.2, the first-best optimum requires the following:

- One firm produces all the output: $n = 1$, from Equation 7A.8.
- Because a positive amount of the monopolistic competition good is produced ($Q > 0$), price equals marginal cost, $u' = p = m$, from Equations 7A.9 and 7A.4.
- The single firm's losses are subsidized to prevent it from shutting down. The subsidy is necessary because the losses $= -F$.

This solution is that of a regulated natural monopoly (see Chapter 20). There are economies of scale everywhere; that is, the firm always operates in the downward-sloping portion of its average cost curve. For this solution to be optimal, funds for the subsidy must be raised in a nondistorting manner. Given a representative consumer, one efficient method of raising funds is a lump-sum tax.

Second-Best Optimum

Now assume that the government cannot achieve the first-best optimum because its actions are constrained:

- The government can control *only* the number of firms, n.
- The government cannot force the firms to produce more than the profit-maximizing quantity; that is, it may not subsidize firms.

In the second-best optimum, assuming firms play Cournot, each firm chooses its (positive) output level such that its marginal revenue equals its marginal cost:[3]

$$\frac{Q}{n} p'(Q) + p(Q) = m, \tag{7A.10}$$

where Q/n is the output of a single firm.

[2] The Kuhn-Tucker conditions are that $\mathscr{L}_{\text{n}} \leq 0$, and if strictly less than zero, $n = 1$ (which occurs here); $\mathscr{L}_{\text{Q}} \leq 0$, and if strictly less than zero, $Q = 0$ (not true here); $\mathscr{L}_{\lambda} \geq 0$, and if strictly greater than zero, $\lambda = 0$; and $\mathscr{L}_{\mu} \geq 0$, and if strictly greater than zero, $\mu = 0$.

[3] A single firm's revenue is $p(q^* + \underline{Q})q^*$, where q^* is its output and $\underline{Q}$ is the output of the $n - 1$ other firms. Differentiating revenue with respect to q^* using the Cournot assumption, and noting that in equilibrium $q^* = \underline{Q}/(n - 1) = Q/n$, we obtain Equation 7A.10.

Firms enter the industry if the marginal firm earns a nonnegative profit. That is, price is at least as great as average cost:

$$p(Q) \geq m + \frac{F}{Q/n}. \tag{7A.11}$$

Equations 7A.10 and 7A.11 determine Q and n.

To find out how much total output changes as the number of firms increases, we can totally differentiate Equation 7A.10:

$$\frac{\mathrm{d}Q}{\mathrm{d}n} = \frac{Q}{n}\left[\frac{p'}{(n+1)p' + Qp''}\right]. \tag{7A.12}$$

The denominator of this expression is negative by the second-order condition, so an extra firm increases industry output: $\mathrm{d}Q/\mathrm{d}n > 0$. A sufficient condition for the second-order condition to hold is $p'' \leq 0$. Because $p''(Q) = u'''(Q)$, p'' can have any sign in general;[4] however, for specificity, we assume $p''(Q) \leq 0$ in what follows.

As a result, increasing output through the market mechanism requires additional firms, and hence additional fixed costs. That is, output may be written as a function of the number of firms, $Q(n)$, where $Q'(n) > 0$. There is a trade-off between the total cost of production (more firms) and lower price (more output).

Society's problem is

$$\max_{n} u(Q(n)) + T - nF - mQ(n), \tag{7A.13}$$

subject to Equations 7A.10 and 7A.11. This problem differs from the problem in Equation 7A.6 in that society is maximizing with respect to n and not with respect to Q and n. Thus, this second-best optimization is constrained in the sense that society can only control Q indirectly through its choice of n.

Ignoring the constraint Equation 7A.11 for the moment, the first-order condition for welfare maximization is

$$(p - m)Q'(n) = F, \tag{7A.14}$$

where $u'(Q)$ is replaced with p using Equation 7A.4. This condition states that the difference between price and marginal cost times the change in output as n increases by one firm $[Q(n+1) - Q(n) \approx Q'(n)]$ equals fixed cost. The left side is the gain from more output from an extra firm, and the right side is the (fixed) cost from one more firm.

[4] See Seade (1980) for a discussion of how stability conditions rule out certain possibilities.

Equation 7A.14 is the appropriate optimality condition if the constraint in Equation 7A.11 is not binding. That is, industry profits are nonnegative for each of the *n* firms, where the government specifies *n*. We can show that profits are positive if the constraint does not bind. First, rewrite Equation 7A.14 as

$$p = m + \frac{F}{Q'(n)}. \tag{7A.14$'$}$$

For profits to be positive, we need $p = m + F/Q'(n) > m + F/[Q/n] = AC$, or

$$Q'(n) < \frac{Q(n)}{n} \quad \text{or} \quad \frac{nQ'(n)}{Q(n)} < 1. \tag{7A.15}$$

That is, the elasticity of *total output with respect to entry* is less than 1.

From Equation 7A.12,

$$\frac{nQ'(n)}{Q(n)} = \frac{p'}{(n+1)p' + Qp''}, \tag{7A.16}$$

so, because $p'' \le 0$,

$$\frac{nQ'(n)}{Q(n)} \le \frac{p'}{(n+1)p'} = \frac{1}{n+1}, \tag{7A.17}$$

or $nQ'(n)/Q(n) < 1$, as required. That is, if $p'' \le 0$, the constraint that $p \ge AC$ (Equation 7A.11) is not binding. The free-entry equilibrium has too many firms (in the sense that the zero-profits constraint is not binding). Thus, society would benefit if it could costlessly restrict the number of firms to the optimal number.

APPENDIX 7B

Welfare in a Monopolistic Competition Model with Differentiated Products

The object of government is the welfare of the people.
—Theodore Roosevelt

We start by considering an economy with only a monopolistic competition industry. We then extend the model to include an outside good.

An Economy with a Single Monopolistic Competition Market

For simplicity, suppose that the degree of product variety is fully reflected by the number of different brands, n. If all firms have identical cost functions, in a symmetric equilibrium, each produces the same amount of output, q.

Each firm's cost function is $C = F + mq$, where C is total costs, F is fixed costs, and mq is the variable costs associated with output level q. Thus, both variable and marginal costs equal m.

Society's production possibility frontier (PPF) is the set of points (q, n) that can be produced with society's total resources, T:

$$(F + mq)n = T, \tag{7B.1}$$

where the left side of Equation 7B.1 is the total cost of n firms producing q units of output each. Equivalently, the PPF is $n = T/(F + mq)$, which is plotted in Figure 7.4. Totally differentiating this equation, we find that the slope of the PPF, $\mathrm{d}n/\mathrm{d}q = -mT/(F + mq)^2$, is negative, as shown in Figure 7.4. Using the numerical example in Table 7.4, $\mathrm{d}n/\mathrm{d}q = -100/(5 + q)^2$. Further, as q rises, the slope becomes less negative, $\mathrm{d}^2n/\mathrm{d}q^2 = 2m^2T/(F + mq)^3$, so that the PPF is concave, as shown in the figure.

Consumers have preferences regarding quantity, q, and variety, n. That is, they are willing to trade off some output of each brand for more brands. For example, the utility function over all *potential* brands, $i = 1, 2, \ldots, \infty$ is

$$U(q_1, q_2, \ldots, q_m, \ldots) = W\left(\sum_{i=1}^{\infty} u_i(q_i)\right). \tag{7B.2}$$

In the symmetric case with n firms in the industry, $u_i(q) \equiv u(q)$, for all i; $q_i = q$, for $i = 1, 2, \ldots, n$; and $q_i = 0$, for $i > n$, so we can rewrite Equation 7B.2 as follows:

$$U(q, q, \ldots, q, 0, \ldots, 0) = W(nu(q)). \tag{7B.3}$$

A consumer's indifference curve corresponding to utility level $\underline{w}$ is

$$W(nu(q)) = \underline{w}. \tag{7B.4}$$

The optimum output-variety combination, O = (q^*, n^*), is determined by the tangency of an indifference curve with the *PPF,* as shown in Figure 7.4. Points on lower indifference curves are less desirable and points on higher indifference curves are unobtainable because they lie above the *PPF.*

Points *A* and *B* on the figure represent possible market equilibrium. That is, one does not know if the monopolistic competition equilibrium lies to the left or the right of the optimum.

A Simple General Equilibrium Model

What is algebra exactly; is it those three-cornered things? —*J. M. Barrie*

To compare the market equilibrium to the optimum, an explicit general equilibrium model should be used. Figure 7.4 only considers the trade-off between output and variety of a monopolistic competition industry. If there is another good, *y,* one needs to consider the trade-off between the two industries. Again, we assume that the outside good, *y,* is produced at a constant cost equal to 1 and that its competitive price is 1.

The Optimum. If the utility function is additively separable in *y,* society's maximization problem is

$$\max_{q_i,\, y} W\left(\sum_{i=1}^{n} u(q_i)\right) + y \tag{7B.5}$$

subject to

$$y = T - \sum_{i=1}^{n} (mq_i + F).$$

If all firms are identical in the sense that they have the same cost function, $q_i = q$, society's problem may be rewritten as maximizing surplus:

$$\max_{q,\, n} W(nu(q)) + T - n(mq + F). \tag{7B.6}$$

There are two first-order conditions for a maximum. The first condition is obtained by differentiating Equation 7B.6 with respect to n, setting the derivative equal to zero, and rewriting it as:

$$W'u(q) = mq + F. \tag{7B.7}$$

This condition says that brands should be added until the marginal gain in welfare from an extra brand, W'u, equals the opportunity cost of the outside good ($mq + F$ is the cost of one more firm in terms of forgone consumption of the outside good).

The other first-order condition is obtained by differentiating Equation 7B.6 with respect to q, setting the derivative equal to zero, dividing through by n and rewriting it as

$$W'u'(q) = m. \tag{7B.8}$$

Equation 7B.8 says that each brand's output, q, should be increased until the marginal gain in utility from an extra unit of output, $W'u'(q)$, equals the marginal cost, m, of an additional unit of output. Using the same type of reasoning as in Appendix 7A, $p = W'u'(q)$. Thus, Equation 7B.8 says that price should equal marginal cost: $p = m$.

Equations 7B.7 and 7B.8 determine the optimal output per brand and number of brands (q^*, n^*). Dividing Equation 7B.7 by Equation 7B.8 and multiplying by $1/q$, we obtain

$$\frac{\dfrac{u(q)}{q}}{u'(q)} = \frac{\dfrac{mq + F}{q}}{m} = \frac{AC}{MC}. \tag{7B.9}$$

That is, at the optimum, the ratio of the average to the marginal utility equals the ratio of the average to the marginal costs.

If utility is concave ($u' > 0$, $u'' < 0$), average utility always exceeds marginal utility. As a result, Equation 7B.9 implies that average cost is greater than marginal cost at the optimum. That is, the optimum lies on the downward-sloping portion of the average cost curve. This condition is automatically met for the specific cost function we chose. It can be shown that this result holds even when average cost curves are U-shaped. Thus, firms should not produce at minimum average cost, as in a competitive industry. The optimum has more variety than would be the case if firms produced at full capacity (the bottom of a U-shaped average cost curve).

The Equilibrium. The equations that describe the Cournot monopolistic competition equilibrium are different from those that describe the optimum, Equations 7B.7 and 7B.8. We now derive the corresponding equations for the equilibrium.

The profits of a representative firm are

$$\pi = qW'u'(q) - mq - F, \tag{7B.10}$$

because $W'u'(q) = p$. Ignoring the integer problem, firms enter until profits are zero ($\pi = 0$) or revenue equals cost:

$$qW'u'(q) = mq + F. \tag{7B.11}$$

This equation differs from the corresponding condition for an optimum, Equation 7B.7, by having $qW'u'(q)$ instead of $W'u(q)$ on the left side.

By differentiating Equation 7B.10 with respect to *q,* we find that the Cournot firm maximizes profits where marginal revenue equals marginal cost, *m:*

$$W''(u'(q))^2nq + W'(qu''(q) + u'(q)) = m. \tag{7B.12}$$

The left side of this equation is different from Equation 7B.8, the condition for an optimum. Thus, because the conditions for an optimum (Equations 7B.7 and 7B.8) differ from those for the equilibrium (Equations 7B.11 and 7B.12), the optimum differs from the equilibrium.

These two sets of conditions are identical only if $W(\cdot)$ and $u(\cdot)$ are linear and $u' = u/q$. That case is uninteresting, however, because each brand is a perfect substitute for every other brand. Demands are therefore perfectly elastic, and no market equilibrium even exists. That is, prices are driven to marginal cost and profits are negative (due to fixed costs), as shown in Appendix 7A.

In general, it can be shown that the equilibrium may lie on either the left or the right of the optimum (in Figure 7.4). To determine the exact relationship, more structure on the utility function is required. A number of articles (Spence 1976; Dixit and Stiglitz 1977; and Koenker and Perry 1981) have worked out the relationship for particular utility functions similar to the one used here. These articles also show that price regulation with a zero-profit constraint leads to the market equilibrium.

CHAPTER 8

Industry Structure and Performance

Merely corroborative detail, intended to give artistic verisimilitude to an otherwise bald and unconvincing narrative. —*W. S. Gilbert*

Theories on competitive and noncompetitive markets hold that the less competition a firm faces, the greater its *market power*: the ability to set price profitably above marginal cost. Thus, market power (and hence price and profits) should be higher in industries with substantial entry barriers that reduce actual and potential competition. Economists conduct empirical investigations to test two of the implications of these theories:

1. How much market power do particular firms (industries) exercise?
2. What are the major factors that determine market power?

For many decades, economists have conducted *structure-conduct-performance* (SCP) studies that concentrate on the second question, which concerns the relationship between market performance and market structure. Market *performance* is the success of a market in producing benefits for consumers (for example, a market is performing well if prices are near the marginal cost of production). Market *structure* consists of those factors that determine the competitiveness of a market. Market structure affects market performance through the *conduct* or behavior of firms. Traditionally, SCP researchers presume that market power or performance can be measured relatively easily, and concentrate on the relationship between performance and structure.

In contrast, many economists now believe that readily available statistics often do *not* accurately reflect either market performance or structure. They rely on new data and techniques to better measure the degree of market power, and its relationship to market performance.

This chapter starts with a summary of the theories on the major market structures based on Chapters 3 through 7. Then, it turns to SCP research and discusses the traditional SCP studies' measures of performance and analyses of the relationship between performance and structure. The main findings are that many industries appear to depart considerably from perfect competition, yet the degree of this departure apparently is not strongly related to industry concentration (the share of sales made by the largest firms in the industry), which presumably reflects the structure of the industry. Finally, the chapter examines modern studies of market power.

Theories of Price Markups and Profits

The relationship between price, p, and marginal cost, MC, and the existence and persistence of economic profits depend on the market structure (Table 8.1). In a competitive industry composed of identical firms with free entry, price equals short-run marginal cost; short-run profits, π_{SR}, are either positive or negative; and long-run profits, π_{LR}, are zero, where capital is charged at its rental price based on the competitive return (or normal return) that capital earns in a competitive industry. Even if firms are price takers (competitive), each firm's profit equals zero in the long run only if each firm has equal access to the same technology and inputs. If some firms have lower costs than others, their profits will not be eroded completely by entry. Free entry guarantees only that the profit of the least profitable firm to enter (the marginal firm) equals zero in the long run.

In monopoly or oligopoly, price exceeds marginal cost, profit in the short run is either positive or negative, and long-run profit is either zero or positive. In monopolistic competition, price is above marginal cost and entry drives long-run profit to zero.

Based on the relationships summarized in Table 8.1, two important conclusions can be drawn. First, testing whether long-run profits are positive is a test of free entry, not of (perfect) competition. Free entry guarantees that long-run profits equal zero, but

TABLE 8.1 Predictions Based on Market Structure

	$p - MC$	π_{SR}	π_{LR}
Competition	0	+ or −	0
Monopolistic competition	+	+ or −	0
Monopoly	+	+ or −	+ or 0
Oligopoly	+	+ or −	+ or 0

p = price, MC = marginal cost (short run), π_{SR} = short-run profits, and π_{LR} = long-run profits.

not that price equals marginal cost: Firms in a monopolistically competitive industry may earn zero profit even though price is above marginal cost. To determine whether price exceeds marginal cost, one must examine price data, not profit data. Second, short-run profits reveal very little about the degree of competition in an industry because, in all market structures, short-run profits can be either positive or negative.

Although Table 8.1 shows only four market structures, many more structures are possible. Moreover, for any given market structure, industries can differ substantially. For example, an oligopoly with four firms may set prices differently than one with only two firms. Generally, one would expect price-cost margins and profits to vary with the number of rivals and the size of barriers to entry. It is this generalization that provides the foundation for the SCP approach.

Structure-Conduct-Performance

Edward S. Mason (1939, 1949) and his colleagues at Harvard introduced the structure-conduct-performance (SCP) approach, which revolutionized the study of industrial organization by introducing the use of inferences from microeconomic analysis. In the SCP paradigm, an industry's *performance*—its success in producing benefits for consumers—depends on the *conduct* or behavior of sellers and buyers, which depends on the structure of the market. The *structure* in turn depends on basic conditions such as technology and the demand for a product.

Because the nature of these connections is usually not explained in detail, many economists criticize the SCP approach for being descriptive rather than analytic. George J. Stigler (1968) and others argued that economists, rather than employ the SCP approach, should use price-theory models based on explicit, maximizing behavior by firms and governments. Others suggested replacing the SCP paradigm with analyses that emphasize game theory (von Neumann and Morgenstern 1944). We discuss modern approaches later in this chapter.

Most of the earliest SCP works were case studies of an individual industry (for example, Wallace 1937). The first empirical applications of the SCP theory were by Mason's colleagues and students, such as Joe S. Bain (1951, 1956). In contrast to the case studies, these studies made comparisons across industries.

A typical SCP study has two main stages. First, one obtains a measure of performance (through direct measurement rather than estimation) and several measures of industry structure. Second, the econometrician uses cross-industry observations to regress the performance measure on various measures of structure so as to explain the difference in market performance across industries. We first discuss the measurement of performance and structure variables and then examine the evidence relating performance to structure.

Measures of Market Performance

Measures of market performance try to provide an answer to our first key question as to whether market power is exercised in an industry. Two different measures that directly or indirectly reflect the profits or the relationship of price to costs are commonly used to gauge how close an industry's performance is to the competitive benchmark:

- The *rate of return,* which is based on profits earned per dollar of investment.
- The *price-cost margin,* which should be based on the difference between price and marginal cost, although, in practice, researchers often use some form of average cost in place of marginal cost.

A third measure, *Tobin's q,* is less commonly used. Tobin's *q* is the ratio of the market value of a firm to its value based on the replacement cost of its assets. (See **www.aw-bc.com/carlton_perloff** "Tobin's *q*" for more details.)

Rates of Return

A **rate of return** is a measure of how much is earned per dollar of investment. This section explains the relationship between economic profits and rates of return. The correct calculation of rates of return can be difficult, and sometimes compromises must be made that bias the final results. We discuss several different rate-of-return measures.

The Relationship Between Rates of Return and Economic Profit. The theories summarized in Table 8.1 make predictions about profit, and a rate of return is a measure of profit. The predictions in Table 8.1 refer to *economic* profit, which is revenue minus opportunity cost, not *accounting* profit (which is measured by accountants, using standard accounting principles). To test the predictions of Table 8.1, an economist's first step should be to adjust accounting profit to reflect economic profit before calculating the rate of return.

There are several important distinctions between economic and accounting profits. The main distinction concerns long-lived capital assets, like plant and equipment. Economic profit equals revenue minus labor, material, and an appropriate measure of capital cost. Measuring revenue, labor cost, and material cost is generally easy. The problem is measuring annual **capital cost**, which equals annual rental fees if all the capital assets were rented. The total rental fees equal the rental rate per unit times the number of units of capital. That is, the appropriate cost measure of capital is a *flow* (the price of renting capital per time period) and not a *stock* (the cost of capital, such as a machine, which lasts for many periods). If well-developed rental markets exist—for example, for used equipment—it is easy to calculate the relevant rental rate on capital and economic profits. When rental rates are not readily available, the economist must implicitly calculate a rental rate before calculating economic profit.

In the calculation of the implicit rental rate of capital used to determine long-run economic profits, capital assets should be valued at **replacement cost**, which is the long-run cost of buying a comparable-quality asset. If capital is valued at its replacement cost, then a low rate of return is a signal that no new capital should enter the industry. It does not mean that the firm should shut down or that it made an error in its past investment decisions. For example, a firm that bought machinery when it was cheap could earn a low rate of return based on the replacement cost and still have enjoyed a huge profit on its initial purchase. A high rate of return is a signal that new capital should enter the industry.

Researchers often divide economic profits by the value of the capital of the firm to obtain an earned rate of return on capital, which is a measure of profitability that controls for differences in capital across firms. There is a close relationship between economic profits, the earned rate of return on capital, and rental rates on capital. To develop this relationship requires an understanding of what a rental rate on capital really is: A rental rate must provide an owner of capital with a particular rate of return *after* depreciation has been deducted on the equipment.

Depreciation is the decline in economic value that results during the period the capital is used.[1] For example, if you rent your house for $1,000 a year, and the wear and tear on the house is $300 a year, then the depreciation is $300, and your net annual rental after accounting for the depreciation is $700. If the house is worth $10,000 initially, then your rate of return is 7 percent, and the depreciation is 3 percent. What matters to the investor is the return after depreciation has been deducted. For that reason, a rental rate (per dollar of capital) can be expressed as an earned rate of return, r, plus a rate of depreciation, δ.

Your profit is

$$\pi = R - \text{labor cost} - \text{material cost} - \text{capital cost},$$

where R is revenue and capital costs are the rental rate of capital times the value of capital. The value of capital is $p_{\text{k}}K$, where p_{k} is the price of capital and K is the quantity of capital. If the rental rate is $(r + \delta)$, then profit is

$$\pi = R - \text{labor cost} - \text{material cost} - (r + \delta)p_{\text{k}}K. \tag{8.1}$$

The earned rate of return is that r such that economic profit is zero. Setting π equal to 0 and solving for r in Equation 8.1 yields

$$r = \frac{R - \text{labor cost} - \text{material cost} - \delta p_{\text{k}}K}{p_{\text{k}}K}. \tag{8.1$'$}$$

Thus, the earned rate of return is net income divided by the value of assets, where *net income* is revenue minus labor cost minus material cost minus depreciation.[2]

The Relationship Between Rates of Return and Price. By how much would price or revenues have to fall in a highly profitable industry in order for that industry to earn a normal rate of return? To see how excess rates of return translate into price overcharges, suppose that a firm earns a rate of return r^* that is 5 percentage points higher than normal: $r^* = r + .05$. That is, the firm's invested capital earns excess rev-

[1] An accountant's definition of depreciation may be based on a formula involving historical cost and age. This measure of depreciation is likely to differ from an economist's measure of depreciation, which is based on opportunity cost.

[2] Another rate-of-return measure is the internal rate of return, which is that interest rate such that the discounted present value of cash flows equals zero. The value of the internal rate of return is that it concisely summarizes the return earned by a project lasting several years. When profitability is changing over time, it may be misleading due to its aggregate nature. Because an internal rate of return depends primarily on the observed cash flows each year (except for the initial and terminal value of the firm), it frees the economist from having to calculate the value of capital each year.

enues of 5 percent times the value of its capital above what it would earn if it were in a competitive industry. If the firm's revenue is R^*, then its rate of return is

$$r^* = [R^* - \text{labor cost} - \text{material cost} - \delta p_k K]/p_k K = \mathrm{r} + .05.$$

Let R be the revenue that would yield a normal rate of return, r. The amount by which revenue must decline to yield the normal return is $R - R^*$, all else constant. Using Equation 8.1′ for r and the expression for r^*, we know that $r - r^* = -.05 = (R - R^*)/(p_k K)$. Multiplying both sides by $p_k K$, we find that $R - R^* = -.05 p_k K$. Thus, to get the normal rate of return, revenue would have to fall by 5 percent of the value of capital.

In many manufacturing industries, the ratio of the value of capital to the value of revenue is roughly 1. In such industries, revenues must fall by 5 percent in order for the industry to earn a normal return. Alternatively, all else constant, price needs to fall by 5 percent. Therefore, if a firm is earning a real rate of return 5 percent higher than the normal rate of return (which was roughly between 5 percent and 10 percent over the period 1948–1976), the competitive price is roughly 95 percent (= 1 − .05) of its current value. That is, industries that earn a rate of return 1.5 times higher than the return earned by competitive industries (say, 15 percent instead of 10 percent) have prices that are only 5 percent above those that generate a normal return. This price overcharge is the same as would occur if a monopoly faced an elasticity of −21. In other words, even large differences in rates of return on capital between concentrated and unconcentrated industries do not necessarily imply that prices in concentrated industries are much above the competitive level. In industries with a low ratio of capital to revenue, even large excess returns can translate into tiny price overcharges.

Pitfalls in Calculating Rates of Return. There are eight major problems in calculating rates of return correctly (see Fisher and McGowan 1983). First, capital is usually not valued appropriately because accounting definitions are used instead of the economic definitions. An economist measures the annual capital cost flow as the annual rental fee if all the capital assets were rented.[3] In contrast, the accounting value of capital, or *book value,* is based on the historical cost of the capital combined with accounting assumptions about depreciation. Capital should be valued at replacement cost (the long-run cost of replacing existing assets with comparable assets) to determine whether the rate of return is above the competitive level (in which case the firm or industry should expand) or below the competitive level (in which case the firm or industry should contract).[4] Because historical cost is often very different from the actual replacement cost of the capital, using the book value of capital rather than the economic value can severely bias the measurement of rate of return.

[3] See **www.aw-bc.com/carlton_perloff**, Chapter 2, "Turning an Asset Price into a Rental Rate."

[4] In all but dying industries, the current value of capital depends on replacement cost. In dying industries, the value of capital is permanently less than replacement cost. The low value of capital is a signal that the industry should not invest in new facilities. In an expanding industry, the current value of capital can exceed replacement cost. The high value of capital is a signal that the industry should invest in new capital. The speed (and cost) of adjustment determines how long the current value can differ from replacement cost.

Second, depreciation is usually not measured properly. Accountants use several fixed formulas to measure the depreciation of an asset. One common formula, called *straight-line depreciation,* assumes that the asset's value declines in equal annual amounts over some fixed period (the *useful life* of the asset). For example, a machine that costs $1,000 and is assigned a useful life of 10 years would incur $100 of depreciation annually for its first 10 years of life. If it lasts more than 10 years, it incurs no additional depreciation. The fixed formula's predictions of the amount of depreciation may be unrelated to the asset's decline in economic value, which is the measure of its economic depreciation. As a result, the estimate of the rate of return may be biased. (See **www.aw-bc.com/carlton_perloff** "Accounting Bias in the Rate of Return.")

Third, valuing problems arise for advertising and research and development (R&D) for the same reason as for capital: All have lasting impacts on either a firm's demand or its costs. The money a firm spends on advertising this year may generate benefits next year, just as a plant built this year provides a benefit next year. If consumers forget about an advertisement's message slowly over time, the advertisement's effect on demand may last for several years. If a firm *expensed* (initially deducted its entire cost of) annual advertising expenditures and then made no deductions in subsequent years, its earned rate of return would be misleadingly low in the initial year and too high in later years. A better approach is to calculate the advertising cost based on the interest rate and the annual decline in the economic value of the advertising. Unfortunately, it is difficult to determine the correct rates of depreciation for advertising expenses.

Similar problems arise with R&D expenditures. Research and development can have a long-lasting impact. In addition, because R&D is risky, we need to be careful in interpreting rates of return. For example, suppose that a firm's research to discover new products is successful one time in ten. If the firm's expected profit is zero, then the profit on the successful product must be high enough to offset the losses on the nine failures. It is misleading to conclude that there are excessively high profits based on an examination of the profit of the one successful product.

Fourth, proper adjustment must be made for inflation. The earned rate of return can be calculated as either a *real* rate of return (a rate of return adjusted to eliminate the effects of inflation) or as a *nominal* rate (which includes the effects of inflation). One should be careful to compare rates that are either all real or all nominal.

If one is using a real rate, income in the numerator of the rate of return should not include the price appreciation on assets from inflation—it should only include the gain in the value of assets beyond that due to general price inflation. For example, if capital is initially worth $100, annual income (before depreciation) is $20, and the annual depreciation rate is 10 percent (so depreciation is $10), then the earned rate of return is 10 percent [(20 − 10)/100]. If inflation was 20 percent during the year, the value of the capital at the end of the year equals $90 ($100 − 10 percent depreciation) times 1.2 (to adjust for inflation), or $108. The firm has incurred a "gain" of $18 on its capital, but it is illusory; it does not represent an increase in purchasing power because all prices have risen as a result of the inflation.

Fifth, monopoly profits may be inappropriately included in the calculated rate of return. This problem stems from using book value in the calculation, because book value sometimes includes *capitalized* (the present value of future) monopoly profit.

Suppose that the monopoly earns excess annual economic profits of $100 above the competitive rate of return and the annual interest rate is 10 percent. The owner of the monopoly sells the firm (and its future stream of monopoly profits) for $1,000 more than the replacement cost of its assets. The owner willingly sells the firm because that extra $1,000 will earn $100, or 10 percent, a year in a bank. The new owner makes only a competitive rate of return because the monopoly profit per year is exactly offset by the forgone interest payments from the extra $1,000. The extra $1,000 paid for the monopoly is the capitalized value of the monopoly profit, *not* the replacement cost to society of replacing the monopoly's capital. Thus, if the reported value of capital inappropriately includes capitalized monopoly profit, the calculated rate of return is misleadingly low if one wants to determine whether an industry is restricting output and is thereby earning an above-normal rate of return.

Sixth, the before-tax rates of return may be calculated instead of the appropriate after-tax rates of return. Corporations pay taxes to the government, and only what is left is of interest to individual investors. That is, after-tax rates of return govern entry and exit decisions. Competition among investors causes after-tax rates of return to be equated on different assets. If assets are taxed at different rates, the before-tax rate of return could vary widely even if all markets are competitive. For that reason, we should use after-tax rates of return and after-tax measures of profit, especially when comparisons are made across industries that are subject to different tax rates.

Seventh, rates of return may not be properly adjusted for risk. To determine whether a firm is earning an excess rate of return, the proper comparison is between the rate of return actually earned and the competitive **risk-adjusted rate of return**, which is the rate of return earned by competitive firms engaged in projects with the same level of risk as that of the firm under analysis. Investors dislike risk and must be compensated for bearing it: The greater the risk, the higher the expected rate of return.[5]

Eighth, some rates of return do not take debt into account properly. Researchers often use the rate of return to the stockholders as a measure of the firm's profitability. If a firm issues debt in addition to equity, both debtholders and equity holders (stockholders) have claims on the firm's income (Chapter 2). Because the assets of the firm are paid for by both debtholders and stockholders, the rate of return on the firm's assets equals a weighted average of the rate of return to the debtholders and the stockholders. The rate of return to debtholders is typically lower than the rate of return to stockholders, because debt is less risky than stock and debtholders get paid before stockholders when a firm is in financial distress. The return to stockholders increases with debt because the

[5] One commonly used approach to adjusting for risk is based on the Capital Asset Pricing Model. According to this model, the expected return on an asset equals the rate of return on risk-free investments (U.S. government Treasury bills are an example of a relatively risk-free investment) plus a number (called *beta*, the Greek letter β) times the difference between the market return (for example, return on the portfolio of all stocks) and the risk-free rate (Brealey and Myers 2003, Ch. 8). Beta reflects how closely the returns on one asset move with the returns on all other assets (the general economy). Risks that are related to movements in the general economy must generate higher returns than the riskless rate of return in order to attract investors.

income received by stockholders in a *highly leveraged* firm (one with a high ratio of debt to equity) is risky, so stockholders in such firms demand high rates of return.[6]

Therefore, it is improper to compare the rates of return to stockholders in two firms in order to measure differences in the degree of competition if the two firms have very different ratios of debt to equity. The debt/equity ratio has nothing to do with whether the firm is earning excess rates of return on its *assets*. Differences among firms in their rates of return to stockholders could reflect differences in competition facing firms or differences in their debt/equity ratios. Even though the rate of return calculated by dividing net income by assets differs from the rate of return from dividing income to stockholders by the value of stockholders' equity, they tend to be highly correlated (Liebowitz 1982b).

Comparing Rates of Return. To judge a rate of return, one must compare it to alternative rates of return. For example, if a firm has 100 units of capital each worth \$10, revenues of \$110, combined labor and material costs of \$10, and capital depreciation of 2 percent per year, then its earned rate of return is 8 percent per year: (110 − 10 − 20)/1,000. If investments in competitive industries yield a 5 percent rate of return, the firm is earning an *excess* rate of return.

There is an equivalent way to reach the same conclusion. If the rental rates on capital were based on the competitive rate of return of 5 percent, then the rental rate would equal 7 percent (5 percent plus depreciation of 2 percent). Calculating economic profit as revenue minus labor cost, material cost, and capital (rental) cost yields a *positive* economic profit of \$30 (110 − 10 − [.07 × 1,000] = 30). Thus, earning *positive* economic profit and earning *excess* rates of return (above the competitive or normal level) are equivalent ways of expressing the same idea. Excess economic profit exists if the earned rate of return exceeds the competitive rate.

Fraumeni and Jorgenson (1980) calculated the after-tax economic rate of return for a large sample of American industries over the period 1948–1976. In their calculations, they were careful to avoid many of the pitfalls described above. They found that over this period, the median manufacturing industry earned a nominal (unadjusted for inflation) rate of return of approximately 11 percent (Table 8.2). Over this same pe-

[6] Suppose that a firm initially has no debt and finances an investment with \$1,000 raised through sale of stock. Next year, the investment returns the \$1,000 plus either \$80 or \$200 with equal probability, so that stockholders' rate of return is either 8 percent or 20 percent, for an average return of 14 percent. Suppose, instead, that the firm raises the \$1,000 for the investment by issuing debt of \$500 that pays 10 percent interest and selling stock worth \$500. Debtholders must receive payment of interest before stockholders receive any income. Therefore, whether the firm earns \$80 or \$200, debtholders receive \$500 plus \$50 of interest. Stockholders receive \$500 plus either \$30 or \$150, so that the total amount paid to both debtholders and stockholders is \$1,000 plus either \$80 or \$200. Stockholders therefore earn either 6 percent (= 30/500) or 30 percent (= 150/500), for an average return of 18 percent, while debtholders earn 10 percent. Stockholders now earn a higher average rate of return and face a wider range of outcomes, even though the income potential of the firm is unchanged.

TABLE 8.2 **Average Annual Returns, 1948–1976**

Industry	Nominal Rate of Return	Own Rate of Return*	Nominal Rate of Return on Stockholders' Equity
Agriculture	.07	.04	
Crude petroleum	.12	.08	
Food	.10	.07	.10
Tobacco	.14	.11	.13
Textiles	.09	.06	.08
Chemicals	.13	.10	.14
Motor vehicles	.29	.25	.15
All manufacturing - median industry	.11	.08	.11
Railroads	.07	.03	
Telephone and telegraph	.15	.11	
Retail trade	.10	.07	

*The own rate of return subtracts from income the effects of increases in the price of capital for each industry. If the price of capital changes only with inflation, the own rate of return is a real (inflation-adjusted) rate of return.

Sources: Fraumeni and Jorgenson (1980); Federal Trade Commission, *Quarterly Financial Reports*, 1948–1976.

riod, the average rate paid on three-month U.S. government Treasury bills was roughly 3.6 percent, so the rate of return in manufacturing significantly exceeded the rate of return on Treasury bills, possibly to compensate for the increased risk.

Studies that calculate rates of return often differ in their methodologies and, because of data constraints, are commonly forced to calculate something other than economic rates of return. Nevertheless, they can still be valuable in investigating whether the rate of return in one industry is higher than that in another, as long as the biases in the calculated rates of return are similar across different industries. It is dangerous, however, to compare the absolute levels of rates of return from one study with the absolute levels of rates of return from another study if the studies follow different methodologies for calculating rates of return.

To illustrate the differences that can arise when different concepts are used to calculate rates of return, the last column of Table 8.2 presents the returns on the book value of stockholders' equity (the difference in the book values of assets and liabilities) published by the Federal Trade Commission (FTC).These rates of return are calculated as after-tax corporate income (which deducts interest payments on debt) divided by the stockholders' equity in the company. Table 8.2 shows that different methodologies can lead to different rates of return. For example, the nominal rate of return in motor vehicles is about 29 percent according to Fraumeni and Jorgenson (1980), but is 15 percent according to the FTC. Nonetheless, the relative rates of return between industries follow the same pattern using both methodologies. For example, tobacco earns a higher rate of return than textiles according to Fraumeni and Jorgenson as well as the FTC.

Price-Cost Margins

To avoid the problems associated with calculating rates of return, many economists use a different measure of performance, the Lerner Index or *price-cost margin,* $(p - MC)/p$, which is the difference between price, p, and marginal cost, MC, as a fraction of the price. The predictions in the first column of Table 8.1 about the relationship of price to marginal cost are stated in terms of the price-cost margin. Because the correlation between accounting rates of return and the price-cost margin can be relatively low (Liebowitz 1982b), it makes a difference which of these two performance measures is used.

The price-cost margin (Chapter 4) for a profit-maximizing firm equals the negative of the reciprocal of the elasticity of demand, ϵ, facing the firm:

$$\frac{p - MC}{p} = -\frac{1}{\epsilon}. \tag{8.2}$$

A competitive firm sets $p = MC$ because its residual demand price elasticity is negative infinity (it faces a horizontal demand curve).

Unfortunately, because a marginal cost measure is rarely available, many researchers use the price-average variable cost margin instead of the appropriate price-marginal cost margin.[7] Their approximation to the price-average variable cost margin is typically calculated as sales (revenues) minus payroll minus material cost divided by sales. That is, they tend to ignore capital, research and development, and advertising costs.[8]

This approach may lead to serious biases. Suppose that marginal cost is

$$MC = v + (r + \delta)\frac{p_k K}{Q}, \tag{8.3}$$

where r is the competitive rate of return, δ is the depreciation rate, and the cost of the labor and materials needed to produce 1 unit of output, Q, is v. Equation 8.3 describes a technology that requires K/Q units of capital (at a cost of p_k per unit of capital) to produce 1 unit of output. Using v in place of marginal cost can lead to serious bias, however, as can be seen by substituting MC from Equation 8.3 into Equation 8.2 to obtain

$$\frac{p - v}{p} = -\frac{1}{\epsilon} + (r + \delta)\frac{p_k K}{pQ}. \tag{8.4}$$

Thus, $(p - v)/p$ differs from the correct measure $(p - MC)/p = -1/\epsilon$ by the last term in Equation 8.4, $(r + \delta)p_k K/(pQ)$, which is the rental value of capital divided by the value of output.

[7] A few studies (Keeler 1983, Friedlaender and Spady 1980) estimate marginal cost based on cost functions.

[8] See Fisher (1987) for a critique of the typical price-cost margin. An even more serious error that is sometimes made is to use average total cost.

Measures of Market Structure

To examine how performance varies with structure, we also need measures of market structure. A variety of measures are used, all of which are thought to have some relation to the degree of competitiveness in an industry. We now describe some of the common measures of market structure.

Industry Concentration. In most SCP studies, industry concentration is the structural variable that is emphasized. Industry concentration is typically measured as a function of the market shares of some or all of the firms in a market.

By far, the most common variable used to measure the market structure of an industry is the four-firm concentration ratio, C4, which is the share of industry sales accounted for by the four largest firms. It is, of course, arbitrary to focus attention on the top four firms in defining concentration ratios. Other concentration measures are used as well. For example, the U.S. government also has published eight-firm concentration ratios, C8.

Alternatively, one could use a *function* of all the individual firms' market shares to measure concentration. The most commonly used function is the **Herfindahl-Hirschman Index**, HHI, which equals the sum of the squared market shares of each firm in the industry. For example, if an industry has three firms with market shares of 50, 30, and 20 percent, the HHI equals 3,800 (= 2,500 + 900 + 400). More attention has been paid to the HHI since the early 1980s, when the Department of Justice and Federal Trade Commission started using it to evaluate mergers. The government publishes HHI statistics by industry.

Typically, empirical studies produce similar results for both the HHI and a four-firm concentration index. It has been shown theoretically (Appendix 8A) that the HHI is the appropriate index of concentration to explain prices if firms behave according to the Cournot model.

Rather than aggregating information about the relative sizes of firms into a single measure, one could examine the effects of the market shares of the first, second, third, fourth, and smaller firms on industry performance. For example, one could determine whether increases in the market share of the second firm raise prices by as much as increases in the share of the leading firm. Using this approach, Kwoka (1979) showed that markets with three (relatively equal-size) firms are much more competitive than those with only two firms.

Table 8.3 shows three concentration measures—C4, C8, and HHI—for several manufacturing industries. Aside from concentration in individual industries, one can examine concentration in manufacturing in general. The 1997 Census of Manufactures reports concentration ratios for 470 manufacturing industries. In 1997, the concentration ratio of the four largest firms was below 40 percent in more than half of the industries, between 41 and 70 percent in about one-third of the industries, and over 70 percent in about one-tenth of the industries, based on value of shipments.

There are now more industries with low four-firm concentration ratios and fewer with high four-firm concentration ratios than in 1935. In 1935, about 47 percent of industries had a four-firm concentration ratio below 40 percent, and about 16 percent of industries had ratios above 70 percent. Since World War II, however, the distribution of concentration ratios in manufacturing has not changed much. Comparisons

TABLE 8.3 1997 Concentration Ratios in Selected Manufacturing Industries

Product Grouping	C4	C8	HHI*
Meat products	35	48	393
Breakfast cereal	83	94	2,446
Distilleries	60	77	1,076
Cigarettes	99	NR	NR
Men's and boy's suits and coats	42	56	846
Sawmills	15	20	87
Folding paperboard boxes	25	38	246
Book printing	32	45	364
Petroleum refining	29	49	422
Tires and inner tubes	68	86	1,518
Blast iron and steel mills	33	53	445
Household refrigerators and freezers	82	97	2,025
Motor vehicles and car bodies	87	94	NR
Computers	40	68	658

*Herfindahl-Hirschman Index for the 50 largest companies. *NR* indicates that the index is not reported.

Source: Census of Manufactures: Concentration Ratios in Manufacturing (2001, Table 2).

based on value of shipments, and not on the number of industries, produce similar conclusions.

Table 8.4 shows that there has not been a trend toward increasing aggregate concentration in the manufacturing sector based on *value added* (revenue minus the cost of fuel, power, and raw materials) accounted for by the largest firms. The table shows that aggregate domestic concentration has increased since 1947, but remained relatively constant between 1967 and 1992 and fell slightly in 1997. Moreover, these domestic concentration statistics overstate concentration because they ignore imports, which have grown in importance.

Most of what we know about concentration ratios concerns manufacturing industries, which comprised only about 14 percent of the GDP in 2001.[9] What about concentration in the other sectors of the economy? Unfortunately, data on concentration ratios are not readily available for most individual industries outside of manufacturing. It is generally believed that ease of entry keeps most of agriculture, services, retailing and wholesale trade, and parts of manufacturing and finance, real estate, and insurance relatively unconcentrated.

Unfortunately, concentration measures have two serious problems. First, many factors influence seller concentration measures. For example, profitability may affect the degree of concentration in an industry by affecting entry. One of the key questions

[9] Table B-12, Economic Report of the President, 2003.

TABLE 8.4 **Percent Aggregate Concentration in the Manufacturing Sector (measured by value added)**

Top Firms	1947	1954	1963	1967	1972	1977	1982	1987	1992	1997
50 largest	17	23	23	25	25	24	24	25	24	21
100 largest	23	30	30	33	33	33	33	33	32	29
200 largest	30	37	38	41	42	43	43	43	42	38

Sources: 1982, 1987, 1992, and 1997, *Census of Manufactures: Concentration Ratios in Manufacturing,* Table 1.

posed in the introduction concerns whether a less competitive market structure "causes" higher profits. A test of this hypothesis is meaningful only if structure affects profits, but not vice versa. That is, this theory should be tested using *exogenous* measures of structure, where exogenous means that the structure is determined before profitability and that profitability does not affect structure.[10]

Most commonly used measures of market structure are not exogenous. They depend on the profitability of the industry. For example, suppose that we use the number of firms as a measure of the structure of an industry, arguing that industries with more firms are more competitive. However, entry occurs in extraordinarily profitable industries if there are no barriers to such entry. Although, in the short run, an inherently competitive industry may have a small number of firms, in the long run, many additional firms enter if profits are high.

An exogenous barrier to entry is a better measure of structure than the number of firms. For example, if a government historically prevented entry in a few industries, those industries with the barrier should have higher profits, but the higher profits do not induce additional entry.

Most SCP studies have ignored the problem with obtaining exogenous measures of market structure. In particular, the commonly used concentration measures, such as C4, are definitely *not* exogenous measures of market structure.

The second serious problem is that many concentration measures are biased because of improper market definitions. The relevant *economic market* for a product includes all products that significantly constrain the price of that product (see Chapter 19). In order for industry concentration to be a meaningful predictor of performance, the industry must comprise a relevant economic market. Otherwise, concentration in an industry has no implication for pricing.

For example, the concentration ratio for an industry whose products compete closely with those of another industry may understate the amount of competition. If plastic bottles compete with glass bottles, the concentration ratio in the glass-bottle industry may reveal very little about market power in that industry. The relevant concentration measure should include firms in both industries. Similarly, firms classified in

[10] If measures of structure are determined by profitability, the measures are said to be *endogenously* determined. Failure to use exogenous measures of structure leads to what statisticians call the "simultaneous equations estimation problem."

one industry that can modify their equipment and easily produce products in another industry are potential suppliers that influence current pricing, but are not reflected in the relevant four-firm concentration ratio.[11]

Unfortunately, concentration ratios are published by the government for specific industries and products, and the definitions used do not necessarily coincide with relevant economic markets. Concentration measures are often based on aggregate national statistics. If the geographic extent of the market is local because transport costs are very high, national concentration statistics may misleadingly indicate that markets are less concentrated than is true. Some researchers use distance shipped to identify markets in which the use of national data is misleading: If the distance shipped is short, the concentration in the local market may be much different from the national market concentration.

Similarly, concentration measures are often biased because they ignore imports and exports. For example, the 1997 four-firm concentration ratio for U.S. automobiles was 80 percent. This figure indicates a very concentrated industry; however, it ignores the imports of British, Japanese, and German cars, which were over 23 percent of total 1997 sales in the United States. The use of improper concentration measures, of course, may bias the estimates of the relationship between performance and concentration.

Just as seller concentration can lead to higher prices, buyer concentration can lead to lower prices. When buyers are large and powerful, their concentration can offset the power of sellers. For that reason, several researchers include buyer concentration as a market structure variable explaining industry performance. The same type of market definition problems can affect this measure. However, this measure is more likely to be exogenous than is seller concentration.

Barriers to Entry. Probably the most important structural factor determining industry performance is the ability of firms to enter the industry (Chapter 4). In industries with significant long-run entry barriers, prices can remain elevated above competitive levels.

Commonly used proxies for entry barriers include minimum efficient firm size, advertising intensity, and capital intensity, as well as subjective estimates of the difficulty of entering specific industries. Chapter 3 makes a distinction between a long-run barrier to entry and the speed with which entry can occur. Most empirical studies do not distinguish these two concepts, and so any measure they use for entry barrier typically reflects both concepts.

Fraumeni and Jorgenson (1980) show that differences in rates of return across industries persist for many years. If there are no long-run barriers to entry or exit, rates of return across industries should converge. Their results indicate that there are long-run barriers, or that the rate of entry and exit is very slow so that convergence in rates of return is slow across industries, or that there are persistent differences across industries in the levels of risk that are reflected in rates of return.

[11] If the producers of some Product B could profitably switch production to Product A (Product B is a *supply substitute* for Product A), then the producers of Product B should also be considered in the market for Product A.

Again, many of the proxies to barriers to entry, such as advertising intensity, are not exogenous. Others, such as subjective measures, have substantial measurement bias.

Unionization. If an industry is highly unionized, the union may be able to capture the industry profits by extracting them through higher wages. Moreover, the higher wages would drive prices up. Therefore, unionization may raise prices to final consumers even though profits of the firms in the industry are not excessive. It is also possible that unions could raise wages and prices and also raise profits to the industry. By making it costly to expand the labor force, unions can prevent industry competition from expanding output and driving profits down. Unionization may not be exogenous if unions are more likely to organize profitable industries.

The Relationship of Structure to Performance

There are hundreds, if not thousands, of studies that attempt to relate market structure to each of the three major measures of market performance. This section first discusses the key empirical findings for each of the performance measures based on U.S. data.[12] Then, SCP studies based on data from other countries and on data for individual industries are examined. Finally, the section summarizes the major critiques of the results and their interpretation.

Rates of Return and Industry Structure. Joe Bain deserves credit for pioneering work that led to the voluminous literature on the relationship between rates of return and industry structure. Bain (1951) investigated 42 industries and separated them into two groups: those with an eight-firm concentration ratio in excess of 70 percent and those with an eight-firm concentration ratio below 70 percent. The rate of return (calculated roughly as income divided by the book value of stockholders' equity) for the more concentrated industries was 11.8 percent compared to 7.5 percent for less concentrated industries.

Bain (1956) classified industries by his subjective estimate as to the extent of barriers to entry. His hypothesis was that profits should be higher in industries with high concentration and high barriers to entry. The evidence that Bain presented is consistent with his hypothesis.

Brozen (1971) criticized Bain's findings for two reasons. First, as Bain recognized, the industries that Bain studied could be in disequilibrium. Brozen showed that the industries Bain identified as highly profitable suffered a subsequent decline in their profits, while the industries of lower profitability enjoyed a subsequent increase in profits. In fact, for the 42 industries of Bain's initial 1951 study, the profit difference of 4.3 percent that he found between the highly concentrated and less concentrated groups diminished to only 1.1 percent by the mid-1950s (Brozen 1971). Second, Brozen pointed out that Bain's use, in some of his work, of the profit rates of the leading firms, rather than the profit rate of the industry, could have skewed his results.

[12] See also **www.aw-bc.com/carlton_perloff** "Tobin's *q*" for a discussion of research using Tobin's *q*.

TABLE 8.5 Average Profit Rates (selected industries)

Eight-Firm Concentration Ratio over 70 Percent		Eight-Firm Concentration Ratio below 70 Percent	
Industry	Profit Rate (%)	Industry	Profit Rate (%)
Auto	15.5	Shoes	9.6
Cigarettes	11.6	Beer	10.9
Ethical drugs	17.9	Bituminous coal	8.8
Liquor	9.0	Canned fruits and vegetables	7.7
Steel	9.0	Average for all industries studied	9.0
Average for all industries studied	13.3		

Source: Mann (1966, 299).

Using 1950–60 data, Mann (1966) reproduced many of Bain's original findings (Table 8.5). Using the same 70 percent concentration ratio criterion as Bain used to divide his sample into two groups, Mann found that the rate of return for the more highly concentrated group was 13.3 percent compared to 9.0 percent for the less concentrated group.

Mann also investigated the relationship between profit and his own subjective estimates of barriers to entry. He found that industries with "very high" barriers to entry enjoy higher profits than those with "substantial" barriers, which in turn earn higher profits than those with "moderate to low" barriers. He confirmed Bain's predictions and earlier findings that concentrated industries with very high barriers to entry have higher average profit rates than concentrated industries that do not have very high barriers to entry.

There have been many econometric estimates of the relation between rates of return, concentration, and a variety of other variables, such as those measuring barriers to entry (see the surveys by Weiss 1974 and Schmalensee 1989). Econometric studies attempt to measure the effects of several variables on rates of return. Such an estimated relationship is called a *regression.* Regression studies provide not only an estimate of the effect of one variable on another but also a statistical measure of whether the estimated effect could be different from zero.

Based on his survey of many of these studies, Weiss (1974) concluded that there was a significant relationship between profit, concentration, and barriers to entry. Studies based on more recent data tend to find only a weak relationship or no relationship between the structural variables and rates of return. For example, Salinger (1984) found, at best, weak support for the hypothesis that minimum efficient scale in concentrated industries is related to rates of return.[13] He found no statistical support that

[13] Large capital requirements do not constitute a long-run barrier to entry unless other conditions, such as imperfect capital markets or sunk costs, are present (see Chapter 3).

his other entry barrier proxy variables (such as advertising intensity) are related to rates of return.

Econometric studies linking profit to market structure often conclude that measured profitability is correlated with the advertising-to-sales ratio and with the ratio of research and development expenditures to sales. These studies also commonly find that high rates of return and industry growth are related.

Some researchers have studied how the speed of adjustment of capital (and hence profit) is related to concentration. Capital-output ratios appear to rise with concentration, though less so recently than in the past. (Table 8.6). The full explanation for the correlation between capital-output ratios and concentration is not known. One possible reason for this result is that the plant of minimally efficient scale (the smallest plant that can operate efficiently) is so large relative to industry size that when economies of scale are important, only a few of them can fit into the industry. However, for most industries, minimum efficient scale (Chapter 2) is a small fraction of total industry demand.

It is possible that the more capital-intensive, concentrated industries use relatively more specialized capital. If so, their rates of adjustment of output should be slower than those of less concentrated industries because it is usually more difficult to adjust specialized capital than it is to adjust less specialized capital. If highly concentrated industries adjust more slowly than unconcentrated industries, that explains why high (or low) profits take longer to fall back to (rise to) the industry average in these industries (Stigler 1963, Connolly and Schwartz 1985, Mueller 1985). See Chapter 3 for studies of entry.

TABLE 8.6 **Capital-Output Ratios and Concentration**

	Average Capital/Output Ratio (Percent)	
Four-Firm Concentration Ratio	1963	1997
0–10	26.5	38.8
11–20	26.9	32.8
21–30	32.7	37.1
31–40	34.5	39.9
41–50	37.7	36.8
51–60	37.9	39.4
61–70	44.2	46.6
71–80	49.8	49.0
81–90	51.8	35.6
91–100	57.7	42.5

Source: 1963 series from Collins and Preston (1969, 272); 1997 series is based on authors' calculations using the 1997 *Census of Manufactures, Industry Series* and *Concentration Ratios in Manufacturing*. The numbers in Table 8.6 are based on gross book value of capital and, because of data unavailability, do not exclude depreciation.

Similarly, if concentrated industries take a long time to react to demand changes, then, all else equal, good economic news should raise the value of a company in a concentrated industry more than the value of a company in an unconcentrated industry. Lustgarten and Thomadakis (1980) find that good economic news raises the stock market values of companies in concentrated industries much more than those in unconcentrated industries, and bad economic news lowers their values more.

Price-Cost Margins and Industry Structure. Following Collins and Preston (1969), many economists examine the relationship across industries between price-average variable cost margins based on Census data and various proxies for industry structure, such as the four-firm concentration ratio and the capital-output ratio. A typical regression based on data from 1958 (Domowitz, Hubbard, and Petersen 1986, 7) is

$$\frac{p - v}{p} = \underset{(.01)}{.16} + \underset{(.02)}{.10\ C4} + \underset{(.02)}{.08\ \frac{p_k K}{pQ}} + \text{other variables},$$

where $(p - v)/p$ is the price-average variable cost margin, v is a measure of average variable cost, $C4$ is the four-firm concentration ratio, and $p_k K/(pQ)$ is the ratio of the book value of capital to the value of output. The numbers in parentheses below each coefficient are standard errors, which are a measure of how precisely the coefficients are estimated.[14] The $p_k K/(pQ)$ term is necessary because price-average variable cost margins are used (see Equation 8.4).

The sensitivity of price to increases in concentration can be derived from this equation. According to the equation, if the value of capital to output, $p_k K/(pQ)$, is 40 percent (the average value across industries), the concentration ratio of the top four firms, $C4$, is 50 percent, and if other variables are zero, the predicted price-average variable cost margin is .24 ($\approx .16 + [.10 \times .5] + [.08 \times .4]$), or $p = 1.3v$. That is, price is 30 percent above average variable cost.

If this industry's four-firm concentration ratio doubles from 50 percent to 100 percent, the price-average variable cost margin rises to .29 or $p = 1.4v$. That is, price rises to approximately 1.4 times average variable cost, which is an increase in the price of only about 7 percent. Thus, even very large increases in concentration may raise price by relatively modest amounts.

Domowitz, Hubbard, and Petersen (1986b) found that, for the time period 1958–1981, the differential in the price-average variable cost margins between industries of high and low concentration fell substantially over time. When they estimated a price-average variable cost equation with more recent data, the coefficient associated

[14] A commonly used method to express the confidence one has in a coefficient is to construct an interval (called a "95 percent confidence interval") for a coefficient that includes all values within roughly two standard errors of the estimated coefficient. The 95 percent confidence interval for the coefficient on $C4$ covers .06 to .14.

with the concentration ratio is much lower than its value in 1958. That is, the already small effect of concentration on price in 1958 shrunk in later years. Further, in the later period, a statistical test of the hypothesis that the concentration measure does not affect the price-average variable cost margin could not be rejected. In general, they found that the relationship between price-cost margins and concentration is unstable, and, to the extent that any relationship exists, it is weak, especially in recent times.

Instead of using industry average variable cost Census data to study the relationship between the price-average variable cost margin and industry structure, other investigators, among them Kwoka and Ravenscraft (1985), used Federal Trade Commission (FTC) data to investigate price-average variable cost margins at the individual firm level.[15] The studies using individual firm data showed that the link between higher concentration and higher price-cost margins is ambiguous. Some studies find that the link, if it exists at all, is very weak, whereas others discern no link at all. They also find that the presence of a large second or third firm greatly reduces the price-cost margin that can be earned. This discovery indicates that it is a mistake to use only four-firm concentration ratios to measure market structure.

Various studies report significant effects from other explanatory variables. Kwoka and Ravenscraft (1985) showed that industry growth has a significant and positive effect on price-average variable cost margins. Lustgarten (1975b) concluded that increased buyer concentration sometimes lowers price-cost margins. Comanor and Wilson (1967) reported that a higher advertising-sales ratio may raise the price-cost margin. Freeman (1983) showed that unions lower the price-cost margin.[16]

International Studies of Performance and Structure. Because international trade is more important in many other countries than it is in U.S. markets, the bias from ignoring imports and exports may be more substantial in studies based on data from those countries than on U.S. data. Concentration ratios based only on domestic concentration may not be economically meaningful as measures of market power. The relevant competition may well be from firms located outside a given country.

Nonetheless, despite differences across countries in sizes of domestic markets, domestic concentration ratios are correlated across countries (Pryor 1972). That is, an industry that is concentrated in the United States is also likely to be concentrated in the United Kingdom. However, the correlation is not perfect, as illustrated by Sutton (1989, 1998) for the U.S. and U.K. frozen food industries.

[15] The advantage of using the firm rather than the industry as the unit of observation is that the researcher can disentangle the effect of industry concentration on a firm's price-cost margin from the effect of the efficiency of that firm alone. For example, one firm's price-cost margin may be high either because the firm is particularly efficient (low cost relative to all other firms) or because all firms in the industry enjoy a high price (lack of competition in the industry). See Benston (1985) for a critique of studies that rely on the FTC data.

[16] Salinger (1984) and Ruback and Zimmerman (1984) also found that unionism has a significant negative effect on the profits of highly concentrated industries. Voos and Mishel (1986) showed that, although unions may depress the price-cost margin, the price is not significantly above the one that would prevail in the absence of a union.

Regardless of which country's data are used, most studies have difficulty detecting an economically and statistically significant effect of concentration on performance (Hart and Morgan 1977, Geroski 1981). However, Encoau and Geroski (1984) found that the United States, the United Kingdom, and Japan tend to have slow rates of price adjustment in their most concentrated sectors.[17]

Performance and Structure in Individual Industries. Most studies of SCP are based on cross-sectional data rather than data on a particular industry over time. There are two serious shortcomings in cross-sectional studies of the relationship between structure and performance across different industries.

First, it is unrealistic to expect the same relationship between structure and performance to hold across all industries. Suppose that one monopolized industry has a high elasticity of demand, and another monopolized industry has a low elasticity of demand. As Equation 8.2 shows, the price-cost margin in the industry with the high elasticity of demand is lower than the price-cost margin in the industry with the low elasticity of demand. Most cross-sectional studies fail to control for differences in demand elasticities across industries, thereby implicitly assuming that the elasticities are identical across industries.

Second, it is unlikely that the four-firm concentration ratios published by the U.S. Census Bureau correspond to the concentration ratios for relevant economic markets. If concentration ratios are not defined for the proper markets, one should not expect to find any correlation between performance and concentration across different markets.

To remedy these two problems, some studies focus on a single industry over time or across different locations. One can, for example, examine how performance in the industry changes over time because of changes in government regulation of entry. Two industry studies are reviewed here.

Airlines. The airline industry would appear to have low costs of entry between city pairs for airlines already in operation. All that is needed is to fly a plane from wherever it is to the new origin and destination pair. That is, the airline industry appears to be a contestable market. Despite the apparent ease of entry, however, studies of the airline industry consistently show that concentration in a city-pair market does influence fares.[18] Actual entry, not potential entry, is critically important in influencing airline fares.

Call and Keeler (1985), Bailey, Graham, and Kaplan (1985), and Graham, Kaplan, and Sibley (1983) found that fares are higher where concentration is high. They typically concluded that fares rise by roughly 6 percent if the four-firm concentration ratio doubles from 50 to 100 percent between two cities (Bailey, Graham, Kaplan 1985,

[17] The industrial organization of Japan is discussed in Caves and Uekasa (1976) and Miwa (1996).

[18] One interpretation of this result is that it is not so easy to construct an optimal airline network that flies passengers from "spoke" cities to "hub" cities where they can interconnect other hubs or spokes. Only in very dense markets with heavy end-to-end travel between city pairs (for example, Chicago–New York) with no interconnecting passengers (and hence no need for feeder traffic) are markets likely to be contestable. Carlton and Klamer (1983) discuss the economics of such networks. The limited numbers of gates, landing slots, and take-off slots at congested airports also limit the ease of entry.

165). Again, there is a statistically significant effect of concentration on performance, but it is of modest magnitude. Borenstein (1989) presented evidence that concentration at an airport (rather than on a particular route between two cities) can also lead to modest increases in fares.[19] Bamberger and Carlton (2003) find that route and airport concentration influence fares, but that this effect is much smaller when one accounts for connecting passengers. (Moreover, they find that the creation of hubs leads to output expansion, a clear benefit to consumers.) Weiher, Sickles, and Perloff (2002) show that the markup of airline fares over marginal cost depends primarily on whether one or two firms dominate a route.

Railroads. In contrast to the apparent ease of entry by airlines, it is now so costly to build a railroad that no one is likely to enter with a new large rail system. Therefore, the number of competitors can be taken as a completely exogenous variable if one focuses on commodities that are shipped only by rail and for which truck (or other) transportation is uneconomical. Studies have estimated the relationship between railroad rates as a function of distance, tons shipped, and concentration after the railroads were deregulated under the Staggers Act of 1980 and were given greater freedom to set fares.

MacDonald (1987) estimated that a railroad facing no competition can charge rates for transporting wheat that are 18 percent higher than when there is a competing railroad. When three railroads compete, rates fall by another 2 percent. These results are statistically significant, yet they indicate that rates do not go up all that much even for dramatic increases in concentration.[20]

Measurement and Statistical Problems. In summary, there is at best weak evidence of a link between concentration and various proxies for barriers to entry and measures of market performance. Are the theories concerning the relationship between performance and structure wrong, or are these studies flawed?

Although many SCP studies are well done, others are seriously flawed. Many of the negative findings in these studies may be due to two important problems. First, these studies commonly suffer from substantial measurement problems or related statistical problems. Second, and more important, most of these studies are conceptually flawed. Most suffer from a variety of measurement errors and other statistical problems that are difficult to correct. Many of these problems were discussed above. We analyze three additional ones here.

First, concentration measures and performance measures are frequently biased due to improper aggregation across products. Because most firms sell more than one product, any estimate of profits or price-cost margins for a firm reflects averages across dif-

[19] See also Hurdle et al. (1989), Borenstein (1992), Brueckner, Dyer, and Spiller (1992), and Evans and Kessides (1993).

[20] Although 20 percent is not small, it is less than one might expect as the difference between monopoly and competition among these firms. A 20 percent price overcharge is about what a monopolist would charge if it faced a demand elasticity of -6. The demand elasticity for rail transport of grain is believed to be considerably less elastic than -6.

ferent products. For a firm that makes products in many different industries, aggregate statistics can be misleading. For example, the Census assigns firms to industry categories based on the primary products produced and includes their total value of production under that industry category. The Census also tabulates statistics at the product level, based on data from individual plants. Because a plant is less likely than a firm to produce several products, product-level data are preferable because such data are less likely to have an aggregation bias than industry-level data.

Second, as discussed in the sections on measuring performance and structure, the performance and structural variables tend to suffer from other measurement errors. Some researchers include variables in addition to concentration to control for such measurement problems in an attempt to reduce these biases. For example, because most price-cost margins ignore capital and advertising, some economists include those two variables in their regressions of price-cost margins on concentration. The inclusion of these additional *explanatory variables* (those used to explain the measure of performance) may not eliminate the bias if they are measured with error or determined by industry profitability. For example, researchers frequently mismeasure advertising, and advertising may be more heavily used in highly profitable industries. The proper interpretation of the coefficients of variables such as advertising is that they reflect, in part, measurement error in the performance measure and not fundamental economic forces influencing "true" price-cost margins (based on marginal costs).

Third, many studies inappropriately estimate linear relations between a measure of performance and concentration. For example, if an increase in the concentration ratio has a smaller effect on performance above a certain level of concentration, the relationship between performance and concentration will flatten and resemble an S-shaped curve. This S-shaped curve can be approximated reasonably by a straight line only if the observed levels of concentration lie in the relatively straight portion of the curve. If concentration ratios vary from very low levels to very high levels, an estimate based on a presumed linear relationship may lead to incorrect results.

White (1976) and Bradburd and Over (1982) searched for critical levels of concentration below which price is less likely to increase as concentration increases, and threshold levels of concentration above which price is more likely to increase as concentration increases. They were only partially successful in finding such a level: There appears to be some evidence of an increase in price at four-firm concentration ratios above roughly 50–60 percent.[21]

Conceptual Problems. Many SCP studies have such serious conceptual problems that it is difficult to use them to test our second key question about the relationship between performance and structure. The two most common conceptual problems concern whether long-run performance measures are used and whether the structural variables are exogenous.

[21] Bradburd and Over (1982) present evidence that the effect of concentration on an industry's performance depends on levels of past concentration. As a highly concentrated industry becomes less concentrated, price remains higher than it would if the industry had never been highly concentrated.

The theories summarized in Table 8.1 predict how long-run profits vary with market structure. They say nothing about the relationship of short-run profits and market structure. Thus, an SCP study based on short-run performance measures is not a proper test of the theories.

The length of time it takes to reach the long run differs by industry. At any moment, some industries are highly profitable while others are not. Over time, some firms exit from the low-profit industries and enter the high-profit industries, which drives rates of return toward a common level. Stigler (1963), Connolly and Schwartz (1985), and Mueller (1985) find that high profits often decline slowly in highly concentrated industries. Only by analyzing both the level of profits (or other measures of performance) and the rate at which they change can the analyst distinguish between a long-run barrier to entry and the speed with which entry occurs (see Chapter 3). Most analyses do not make this distinction. This issue may be regarded as a problem in accurately measuring performance.[22]

The more serious conceptual problem with many SCP studies is that the structural variables are not exogenous. Many researchers, after finding a link between high profits (or excessive rates of return, or large price-cost margins) and high concentration ratios, infer improperly that high concentration rates are bad because they "cause" high profits. Profit and concentration, however, influence each other. An alternative interpretation of a link between profits and concentration is that the largest firms are the most efficient or innovative (Demsetz 1973, Peltzman 1977). Only when a firm is efficient or innovative is it profitable to expand in a market and make the market concentrated. In this interpretation, a successful firm attracts consumers, either through lower prices or better products. A firm's success, as measured by both its profits and its market share, is an indicator of consumer satisfaction, not of poor industry performance. One implication of this hypothesis is that a firm's success is explained by its own market share and not just by industry concentration, as found by Kwoka and Ravenscraft (1985).

If concentration is not an exogenous measure, then an estimate of the relationship between profits and concentration, which assumes that concentration affects profits and not vice versa, leads to what is referred to as a simultaneity bias. Weiss (1974), however, estimated the relationship between performance measures and concentration using statistical techniques designed to eliminate the simultaneity bias problem and found that the different estimation procedures make little difference in the estimated relationship.

Although the regression results may not change, their interpretation does. Even a correctly estimated relationship between performance and concentration is uninformative regarding causation. Concentration does not cause high profits; long-run barriers to entry do. These barriers lead to both high profits and high concentration.[23]

[22] The various measurement problems with performance may not be as serious as they first appear. Schmalensee (1989) used 12 different accounting measures of profitability in a SCP study. Strikingly, although these 12 measures are not highly correlated, many of his key SCP results held over all measures.

[23] Research on SCP continues. Noteworthy work includes Marvel (1978), Lamm (1981), Cotterill (1986), Schmalensee (1987, 1989), Cubbin and Geroski (1987), and especially Sutton (1991, 1998).

Modern Structure-Conduct-Performance Analysis

The original structure-conduct-performance literature sought to establish a systematic relationship between price and concentration. As we have noted, the criticisms of this approach are many, but perhaps the most significant criticism is that concentration itself is determined by the economic conditions of the industry and hence is not an industry characteristic that can be used to explain pricing or other conduct. The barrage of criticism has caused most research in this area to cease. But Sutton and his coauthors have developed an approach that builds on the structure-conduct-performance idea of looking for systematic patterns of competitive behavior across industries, and that at the same time addresses the endogenous determination of entry (Sutton 1991, 1998).

Sutton's research examines what happens to competition as market size grows. Does the market become less concentrated? Do other dimensions of the product—such as quality, promotional activity, and research and development—change? What are the fundamental economic forces that provide the bases for systematic answers to these questions across different industries? In answering these questions, Sutton analyzes markets in which the product is either homogeneous or heterogeneous and considers the cost of entering the market or altering certain attributes of products.

Theory

We divide our discussion of Sutton's theory into two cases depending on whether a firm's cost of entry is an exogenous sunk cost or an endogenous sunk cost. In the former case, each firm must spend some fixed amount, F, to enter the industry. In the latter case, the amount a firm must spend to enter the industry is variable and is chosen by the firm in an effort to affect the desirability of its product by influencing certain dimensions of the product.

Exogenous Sunk Cost. To illustrate his theory, Sutton examines markets with homogeneous and heterogeneous products. We start by considering a market in which the firms produce a homogeneous product and the only variable firms can compete on is price, not quality. Each firm incurs a fixed cost F and has a constant marginal cost m. At low prices, the industry demand curve is $Q = s/p$, where Q is industry quantity, s is a measure of market size (total expenditure, which is assumed to be determined independently of price), and p is price. That is, for low prices and given s, the market elasticity of demand is -1. At some high price p_m, the demand curve is perfectly elastic. Thus, a monopoly would charge a price of p_m in this market (see Chapter 4).

The final equilibrium and the change in equilibrium as the market size grows are determined by the form that competition takes. To fix ideas, Sutton considered three types of competition, each "tougher" than the next. The level of competition is lowest in a cartel in which all firms explicitly collude to set the monopoly price p_m and divide up the total cartel profit or monopoly profit among the n firms. Regardless of the number of firms, n, the price remains at p_m. Thus, profit per firm declines as n grows because the total monopoly profit is divided among more and more firms. At the equilibrium n, the total cartel profit is driven to zero.[24]

A more competitive market is a Cournot oligopoly. For any number of firms, n, the equilibrium Cournot price is $p(n) = m[1 + 1/(n - 1)]$.[25] Thus, the Cournot price p falls to m as n increases. The output per firm, q, equals $(s/m)[(n - 1)/n^2]$, while profit per firm is $[p - m]q - F$, which equals $s/n^2 - F$. Hence with free entry, n equals $\sqrt{s/F}$, at which point profit per firm is zero.

Finally, consider the toughest form of competition, Bertrand, where price equals m for any given $n > 1$. Here, the only free-entry equilibrium has one firm with positive profit. If a second firm enters, price is driven to marginal cost, so that profit is negative (because of the fixed cost), which leads to one firm's demise.

For each model of competition, Figure 8.1 shows how price changes as n increases. As the figure illustrates, for any given $n > 1$, price is lower as competition becomes "tougher," with Bertrand being the toughest and cartel being the least tough model of competition.

Figure 8.2 relates a measure of equilibrium industry concentration, $1/n$, to market size s for each model of competition, where by equilibrium market concentration, we mean that n such that total profit equals zero (or more accurately, if one additional firm enters, it will earn a negative profit).

Figure 8.2 reveals two interesting results. First, as expected, concentration falls as market size increases for all but the most competitive game (Bertrand). The intuition for this result is that larger markets can accommodate more firms.

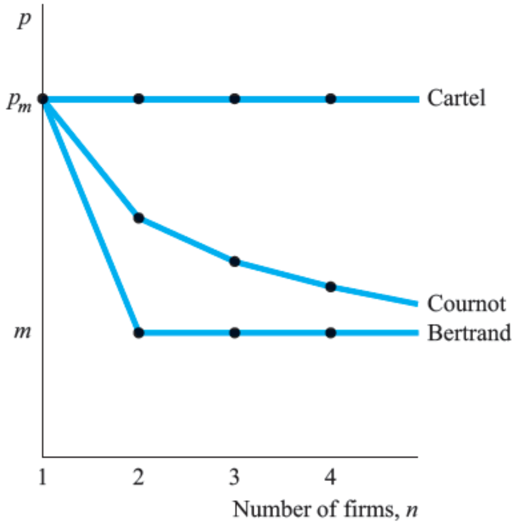
The second result is counterintuitive: For any given market size, equilibrium market concentration is *higher*, the tougher the competition. Concentration is lowest for the cartel model, even though the cartel model has the highest price. The reason for this result is that tough competition leads to a low price, which discourages entry. This result illustrates that relying on concentration alone to make inferences about price and competitiveness can lead to erroneous conclusions.

The case of exogenous fixed costs with heterogeneous products has much less crisp results than the case of exogenous fixed costs with a homogenous product. In a model with heterogeneous products (such as the models in Chapter 7), the concentration in the market depends on the nature of the game, such as how many different products one firm may produce and whether a firm has an advantage if it can choose its products before other firms choose.

[24] Let the cartel profit be $\boldsymbol{\pi} = [p - m]\ Q - nF$. The price that maximizes cartel profit is the same price that maximizes $[p - m]\ Q$. Define $\boldsymbol{\pi}_m$ as the maximum of $[p - m]\ Q$ (that is, it is the profit, ignoring fixed costs). Then, each firm's individual profit is $\boldsymbol{\pi}_m/n - F$. In equilibrium where $\boldsymbol{\pi} = 0$, the equilibrium n equals $\boldsymbol{\pi}_m/F$.

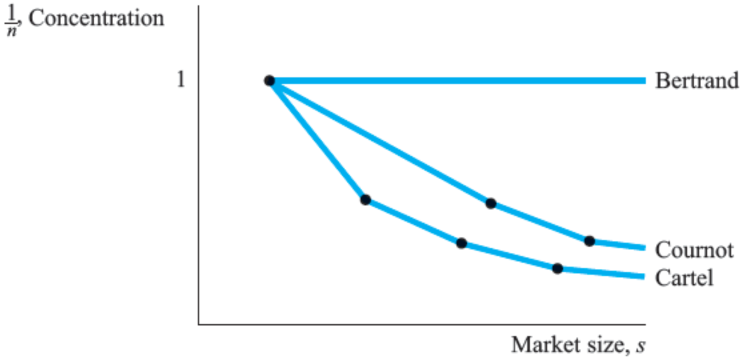
[25] Each firm selects its output q_i to maximize its profit, which can be written as $p_i(\Sigma q_j)q_i - mq_i - F$, where $p_i(\Sigma q_j)$ is the inverse demand curve. Differentiating this expression with respect to q_i yields the first-order condition for each firm's optimal output level given its rival's output levels. Setting $q_i = q$ for all i yields the symmetric Cournot equilibrium (assuming that the resulting price is less than p_m). See Problem 6 at the end of the chapter.

FIGURE 8.1 Relationship Between Prices and Number of Firms Under Three Market Structures



Sutton's main result for heterogeneous products is that the "toughness" of competition is, in general, diminished when one moves from a homogeneous to a heterogeneous product and so (analogous to the result that occurs in the case of a homogeneous product as competition weakens) the equilibrium concentration tends to fall

FIGURE 8.2 Relationship Between Concentration and Market Size Under Three Market Structures



for any given market size *s*. However, unlike the case of a homogeneous product, there are many possible equilibrium outcomes for any given market size, *s*, and the best that an economist can derive is a *lower bound* on concentration for any given *s*. When this lower bound is low, there are very few empirical predictions one can make about equilibrium concentration because any equilibrium concentration is possible as long as it exceeds the lower bound.

The property that equilibrium concentration (or its lower bound) decreases with market size *s* depends on the assumption that fixed costs are exogenous and that product quality is given. Given this property, all else equal, concentration should be lower in big countries than in small countries, when market size is determined by the size of country.

Although this result holds for many industries, there are some industries that are highly concentrated in both large and small countries (Pryor 1972). How can this fact be explained? In examining this question, Sutton and his coauthors have substantially increased our understanding of the competitive process. We now turn to their findings.

Endogenous Sunk Costs. In most markets, firms compete not just on price but also along many other product dimensions, such as quality, reliability, research and development, and promotional activity. To fix ideas, let W be an index of quality, which we will broadly interpret to include information about the product. The key new assumption is that a firm may spend money to improve its product's W. For example, firms increase their expenditures on advertising, research and development, and engineering to increase W by raising the quality of the product or by heightening consumers' perception of its quality. Firms can compete for customers by spending money to improve product quality, lowering price, or both. Here we say that the firm has *endogenous* sunk costs because the firm decides how large an investment to make.

Paying to improve quality has two important effects. First, it raises the firm's fixed cost and perhaps its marginal cost of production if a higher-quality good costs more to produce. Second, it attracts customers who were previously buying a lower-quality good. These two effects can combine to completely reverse the results in the previous section, in which an increase in market size is associated with a decrease in equilibrium concentration. As market size *s* increases, firms have an incentive to compete by improving the quality W of their product. To raise quality, a firm must incur larger sunk costs, a circumstance that reduces the incentive for additional firms to enter the industry that otherwise arises from the larger *s*. As a result, as market size increases, concentration no longer necessarily falls. A given industry in different-size markets can remain highly concentrated, but bigger markets will have higher-quality products.

For this reasoning to hold, several assumptions need to hold. Consumers must value improvements in quality sufficiently so that they switch from lower-quality products to higher-quality goods. To establish the condition under which this assumption is so, Sutton uses a model of *vertical differentiation*. In this model, every consumer agrees on a ranking of products by quality, W, with all consumers preferring a higher-quality product to a lower-quality product.

Suppose that a consumer's surplus from a good of quality W equals $U = \theta W - p(W)$, where θ is a parameter reflecting the weight that consumers place on quality and $p(W)$ is the price of a product with quality W. Because consumers differ in θ, even though all consumers prefer more W to less, some consumers place such a low value on extra W that they are willing to pay very little extra money for a high W product, while other consumers so enjoy extra quality that they will buy a higher-quality good even if the price is relatively high. The optimal W for any consumer will depend on the price function $p(W)$, which reveals how prices rise as W rises and on the consumer preference, θ, for quality.

Sutton proves that as long as $p(W)$ and the marginal cost of producing a high-quality product do not rise "too fast" as W increases, then the equilibrium has three striking properties. First, the firms that produce the highest quality available in the market are the largest firms.

Second, an increase in market size leads to an increase in the quality of the best products in the market, with higher-quality products being chosen by consumers at higher prices and some lower-quality products disappearing from the market. Thus, the equilibrium quality rises as the market expands. Third, with higher quality and its attendant costs, fewer firms can afford to remain in the industry and concentration will remain high. Consequently, the property that a market remains concentrated as s increases continues to hold even where there is both horizontal and vertical differentiation as long as there is sufficient substitution between the vertical dimension (quality W) and the horizontal dimension over which consumers can have different preferences.

For both the endogenous and the exogenous sunk cost cases, the key empirical predictions about concentration and market size depend on the validity of certain assumptions. The most important assumption is that the form of the game—Bertrand, Cournot, or cartel—remains unchanged as market size increases. In a given market, this assumption may or may not be plausible. Moreover, neither Sutton nor anyone else has made significant progress in defining the industry economic characteristics that predict the form of the game that describes the competitive process. Therefore, analogous to the criticism of the earlier literature that concentration need not be exogenous, here we have the criticism that the form of the competitive game need not be exogenous.

Empirical Research

Sutton has produced two voluminous books of studies using data from six countries—France, Germany, Italy, Japan, the United Kingdom, and the United States—to test his theories, especially those concerning the endogeneity of advertising and technology. Sutton's empirical work helps explain why concentration is similar across different-size countries for some industries but not others. See Example 8.1.

In Sutton (1991), he tests his theoretical predictions about the relationship between concentration and market size for several industries in the food and beverage sector. He separates the industries into two types, one in which there is little advertising and the other in which there is significant advertising. The first industry type corresponds roughly to the use of exogenous sunk cost, while the second corresponds roughly to the case of endogenous sunk cost. For each type of industry, Sutton runs a regression

EXAMPLE 8.1 *Supermarkets and Concentration*

Ellickson (2000) applies Sutton's theory to the supermarket industry. In contrast to Sutton's focus on sunk costs associated with advertising or technology, Ellickson examines the role of sunk cost at the store level of building a large store and at the firm level of having the expertise and distribution systems to provide a wide variety of brands. Ellickson explains that these costs are important in distinguishing high-quality firms from low-quality stores, where he uses store size, existence of a deli or bakery, and existence of scanners and ATM machines to measure quality.

Ellickson examines the four-firm concentration ratio for supermarkets across 320 different metropolitan statistical areas (MSAs). Regardless of the size of the MSA, four or five firms that typically own multiple stores account for 70 to 80 percent of sales within each MSA. Moreover, concentration at the metropolitan level has remained high both over time and as the markets grew. Ellickson explains these results using Sutton's endogenous sunk cost theory. According to that theory, increases in market size should lead to higher-quality supermarkets but not more firms, which is exactly what Ellickson finds.

Further, consistent with the theory, the largest firm in each MSA provides a higher quality product than do the smaller firms. In addition, the quality of these largest firms differs across MSAs in exactly the way that the theory would predict: The firms in the largest markets have the highest quality.

Ellickson also examines how the industry has changed over time as MSAs have grown. The trend in the supermarket industry has been one of increasing concentration over time. For example, the average four-firm concentration ratio across 154 MSAs has grown from 45 percent in 1954 to 75 percent in 1998. Consistent with the theory, the number of products offered by each store has increased from 14,145 in 1980 to 21,949 in 1994, while average store size has been growing at the rate of 1,000 square feet per year.

of the form $C4 = a + b\ln(s/\sigma)$, where $C4$ is the four-firm concentration ratio and s/σ is the market size divided by the size of an efficient plant.[26] An econometric test of the theory is that b is negative for the first type of industry but zero for the second type. For his sample, Sutton indeed finds this result, which provides impressive empirical support for his theories.

Thus, Sutton's work increases our theoretical and empirical understanding of the relationship between concentration and competition. Still, there are two important caveats to Sutton's results. First, as his detailed analysis of each industry in each country reveals, the assumption that the competitive game is the same across

[26] Sutton actually uses a more complicated method because his theory predicts a lower bound to the relationship between concentration and market size.

countries is not always a particularly good one. There is little research so far explaining why in some countries competition in a particular industry is more intense than in others.[27]

Sutton uses the difference across countries in the competitive game for an industry to his advantage. Sutton identified industries and countries where competition is unusually intense and found, consistent with his theory, that the industry is more concentrated in those countries. He identifies countries with lax attitudes toward cartels and again, consistent with his theory, finds that those industries tend to have lower concentration levels.

The second caveat is that Sutton's theory predicts a lower bound to the relationship between market size and concentration. The reason for the lower bound is that there can be a multiplicity of equilibria, with some having greater concentration levels than the lower bound. The theory therefore is unable to help us much in predicting concentration in a particular country when the lower bound is low.

Although Sutton explains that this theory of lower bounds is the most one can say under general conditions, the analyst is left in the uncomfortable position of having a theoretical structure that may not narrow the possible equilibria very much. Sutton's detailed history of each industry shows that many idiosyncratic factors often are critical in explaining an industry's evolution. Thus his work provides a sobering lesson because it reveals the limits of theory to explain industrial structures.

Modern Approaches to Measuring Performance

An economist's guess is liable to be as good as anybody else's.

—Will Rogers

The SCP studies focus on our second question, concerning the relationship between performance and structure, and pay relatively little attention to the first question—how to measure performance. In contrast, most modern empirical approaches focus on measuring performance or market power. These studies start by rejecting the traditional measures of performance on the grounds that they are significantly flawed due to accounting difficulties. These approaches estimate market power using models based on formal theories of profit-maximizing behavior described in earlier chapters.

Researchers use both static and multiperiod models to estimate market power. Some economists rely directly or indirectly on observations of marginal cost and price; others look at the behavior of output or price to see if it is consistent with the competitive model. The following sections discuss some of these methods.

[27] One curious finding is that concentration tends to be slightly higher in the United States than in European countries. One explanation is that the United States has more intense competition, which Sutton's theory suggests should lead to higher levels of concentration.

Static Studies

Most modern studies based on static models can be divided into those that estimate marginal cost directly, those that estimate entire models of a market (thereby obtaining estimates of marginal cost and of the markup), and those that observe the relationship between changes in price and factor costs to test whether an industry is competitive.

Estimate Marginal Cost Using Cost Data. The most direct way to answer our first key question about the degree of market power in an industry is to calculate the price-cost markup directly.[28] Although price data are available for most industries, unfortunately, marginal cost data are generally not.

If information on total cost is available, however, an economist can estimate the relationship between observed total cost and observed total output and then calculate marginal cost. A price-cost margin is then simply calculated. Weiher et al. (2002) estimated marginal cost using total cost information and then calculated Lerner measures of market power directly.

Even total cost data, however, are rarely available. Studies that estimate cost functions frequently examine regulated industries because regulators force the firms to provide cost data. For example, Keeler (1983, 71) and Friedlaender and Spady (1980, Ch. 4) found that price exceeded long-run marginal cost by about 22 percent for rail service for bulk commodities in the Northeast during the late 1960s and early 1970s. Genesave and Mullin (1998) use cost data from a court case and find small markups.

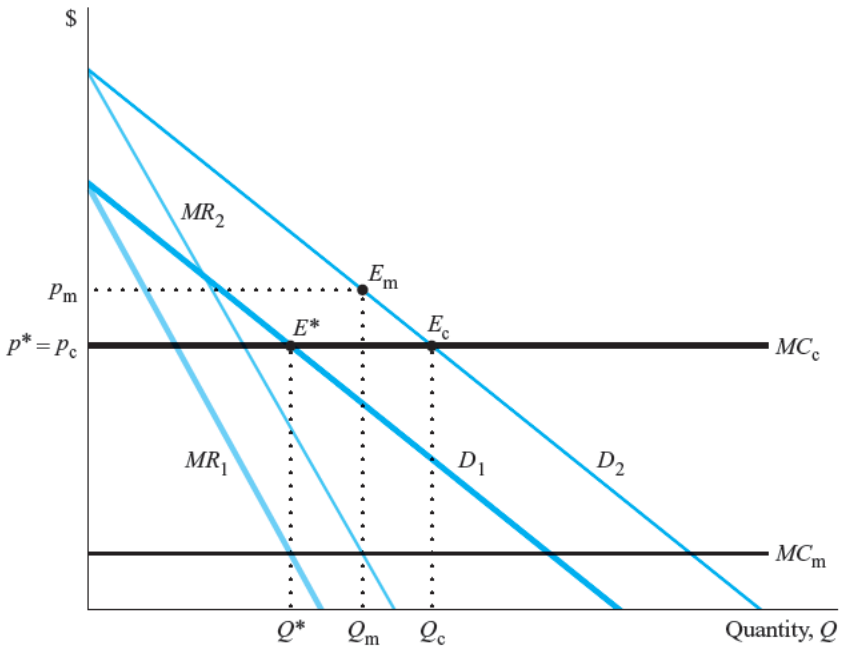
Estimate the Markups Using an Industry Model. If cost data are not available, so that we cannot directly estimate marginal cost, *MC,* how can we calculate the price-cost markup? One method is to use assumptions about the shape of the demand and *MC* curves to infer the markup from observations on how the equilibrium price and quantity change over time.[29] This approach is called the new empirical industrial organization.

For many markets, we have enough information to estimate a demand curve. Figure 8.3 shows the demand curve, D_1, in a particular market. Suppose that we believe that the industry's marginal cost, *MC*, is constant, although we do not know its level. Currently, the market equilibrium, point E^* in Figure 8.3, is at price p^* and quantity Q^*. That equilibrium could be produced by a competitive industry with a relatively high marginal cost, MC_c, or by a monopoly with a relatively low marginal cost, MC_m

[28] Hall and Hitch (1939) conducted a series of interviews with businesspeople regarding their firms' pricing practices. Most claimed that they set price above marginal cost.

[29] See Bresnahan (1989) for a more extensive discussion and Corts (1999) for a critique that explains that these empirical methods depend on the validity of the conjectural variations model that is theoretically an inadequate model. Apparently the first modern study was Rosse (1970). Six other influential early studies are Iwata (1974), Applebaum (1979, 1982), Gollop and Roberts (1979), Just and Chern (1980), and Bresnahan (1981). The other major early conceptual work was Rohlfs (1974); however, it did not contain an empirical application.

FIGURE 8.3 Identifying Market Power



(which intersects the monopoly's marginal revenue curve, MR_1, at Q^*). With only this information, we cannot *identify* (determine) the marginal cost and price-cost markup.

However, if in the next period the demand curve shifts to D_2, which is to the right and parallel to D_1 in Figure 8.3, we can determine whether the industry is competitive or monopolistic. If the industry is competitive, the new equilibrium is at point E_c, so the price remains constant, $p_c = p^*$, and output increases substantially to Q_c. That is, by noting that the shift in demand does not change the price, we know that the industry marginal cost is MC_c and that Lerner's price-cost margin, $(p - MC_c)/p$, is 0.

If, instead, the shift in demand leads to a new equilibrium at point E_m, the price increases from p^* to p_m and the quantity increases only to Q_m. This increase in price is consistent with noncompetitive behavior. Thus, if we know MC is constant, an outward shift of the demand curve reveals whether the market is purely competitive. If the price does not change, the market is competitive; if the price increases, there is market power.

Economists use generalizations of this approach to estimate the degree of market power, Lerner's price-cost margin, and the marginal cost curve. Typically, they make specific assumptions about the shapes of the demand and marginal cost curves, which allow them to identify the price-cost margin by observing shifts in equilibrium price and quan-

TABLE 8.7 Estimated Price-Cost Margins

Study	Industry	$(p - MC)/p$
Bresnahan (1981)	Autos	.10–.34
Appelbaum (1982)	Rubber	.05
	Electrical machinery	.20
	Tobacco	.65
Porter (1983a)	Railroads (with cartel)	.40
Lopez (1984)	Food processing	.50
Roberts (1984)	Coffee roasting (largest firm)	.06
Spiller and Favaro (1984)	Large banks before deregulation	.88
	Large banks after deregulation	.40
Suslow (1986)	Aluminum	.59
Slade (1987b)	Retail gasoline	.10
Karp and Perloff (1989a)	Rice exports (largest estimate)	.11
Karp and Perloff (1989b)	Small black-and-white TVs in Japan	.58
Buschena and Perloff (1991)	Philippines coconut oil (post-1974)	.89
Wann and Sexton (1992)	Fruit cocktail	1.41
Deodhar and Sheldon (1995)	German bananas	.26
Genesove and Mullin (1998)	Sugar refining 1880–1914	.05
Hyde and Perloff (1998)	Australian retail meats	≈0

Sources: Articles cited and Bresnahan (1989, Table 1)

tity over time. One method is described in Appendix 8B.[30] Table 8.7 reports the estimated price-cost margins for several industries using these approaches. See Example 8.2.

Indirect Approaches. Some economists use the changes in price associated with changes in costs to test whether an industry is competitive without having to make detailed assumptions about the shapes of both the demand and the supply curves. If marginal cost shifts up by a certain amount in a constant-marginal cost market, the competitive price rises by the same amount because price equals marginal cost. For ex-

[30] Bresnahan (1989) surveys many of these studies, including Iwata (1974), Gollop and Roberts (1979), Spiller and Favaro (1984), Roberts (1984), and Applebaum (1979, 1982). Analogous techniques can be used to estimate monopsony power as well (Just and Chern 1980, Azzam and Pagoulatos 1990). Separate price-cost margins can also be estimated for individual firms, as shown in Spiller and Favaro (1984), Baker and Bresnahan (1985, 1988), Slade (1986, 1987a, 1987b, 1992), Gelfand and Spiller (1987), and Karp and Perloff (1989b). Any shock that shifts the relevant demand curve or marginal cost curve can be used to identify market power. For example, changes in taxes (Kolstad and Wolak 1983, 1985, 1986; Wolak and Kolstad 1988) or changes in the supply of a fringe (Buschena and Perloff 1991) help identify market power. Some of the more interesting applications (Bresnahan 1981, 1987) estimate market power based on spatial competition models (Chapter 7), taking explicit account of product differentiation.

EXAMPLE 8.2 *How Sweet It Is*

If we have data on marginal cost, *MC*, and market price, *p*, we can calculate the Lerner Index (Chapter 4) of market power, $(p - MC)/p$, directly. Unfortunately, such cost data are usually not available.

As a result, most new empirical industrial organization studies of market power identify the Lerner Index by using assumptions about the shape of the demand and marginal cost curves. Consequently, these estimates of market power are only as reliable as the (untested) assumptions about the shapes of the curves. Typically, these studies estimate the degree of market power using a parameter λ so that the Lerner Index is (Appendix 8B)

$$(p - MC)/p = -\lambda/\epsilon,$$

where ϵ is the estimated market demand elasticity. If the market is competitive, $\lambda = 0$ and there is no gap between price and marginal cost. If the market is monopolized, $\lambda = 1$. If λ is between zero and one, the degree of market power is between that of a competitive and monopolized market.

Genesove and Mullin (1998) have data on cost for the sugar refining industry. Consequently, they are able to see how well the estimation approach works compared to directly calculating market power using cost data.

The sugar refining industry became a highly concentrated industry as the result of acquisitions by the American Sugar Refining Company in the late 1800s. The largest firm accounted for over 60 percent of sales. Detailed cost data for this industry for 1880–1914 are available as a result of antitrust litigation. Genesove and Mullin use these data to calculate marginal cost directly, which they then use to calculate the Lerner Index. According to these calculations, a typical value of the Lerner Index during the 1880–1914 period is .05 (nearly perfectly competitive), while a typical value for λ is .1.

Next, Genesove and Mullin ignore their detailed cost data and use a technique similar to that described in Appendix 8B to estimate the demand curve, the marginal cost curve, and λ. Using this method, they estimate that $\lambda = .04$ and that the implied Lerner Index is .02. Although their econometric approach leads to a lower estimate of λ than does their cost method, the econometric method succeeds in correctly telling the researchers that, despite the high industry concentration, a monopoly ($\lambda = 1$) model is less consistent with the data than is a competitive ($\lambda = 0$) model.

ample, in a competitive market, a per-unit tax of $1 raises price by $1. By observing the relationship between the change in price and the change in costs (or some element of costs), one can test whether the industry is competitive.

Sumner (1981) examined the effect of tax differences across states on the price of cigarettes. He argued that if the retail prices of cigarettes differ between states by the amount of the tax differences, the market is relatively competitive. Bulow and Pfleiderer (1983) pointed out that it is possible to construct demand curves for which a monopoly does pass on costs on a one-for-one basis. Sullivan (1985) used a different method to avoid this criticism and confirms Sumner's finding of a significant degree of competition in cigarettes. Similarly, Ashenfelter and Sullivan (1987) used changes in excise taxes to identify market structure.

Hall (1988a) demonstrated another method of determining market power without making specific assumptions about the demand curve. He showed that, with constant returns to scale, shifts in costs are sufficient to identify market power.[31] When such an industry expands output in response to a shift in demand, the total value of its output (revenues) increases by exactly the increase in its total cost if the industry is competitive. If value rises by more than the additional cost, then price is above marginal cost and the industry is not competitive.[32]

Hall estimates very large markups, but subsequent work by Domowitz et al. (1988) and Roeger (1995) find much lower markups.[33] Roeger (1995) obtains markups ranging from 5 to 23 percent.

Multiperiod Studies

Almost all real-world markets last for many periods. A multiperiod model should be used to estimate market power if firms, in setting strategies, take previous behavior into account; if adjustment costs are significant, so that costs in this period depend on decisions in previous periods; or if demand today depends on past consumption. Economists use at least two types of multiperiod models to estimate market power: models of collusive behavior and models of behavior with costs of adjustment.

Collusion and Repeated Static Games. Stigler (1964a) argued that the opportunity and desire by oligopolistic firms to collude (at least tacitly) provides a basis for explaining all oligopoly behavior (Chapter 5). In this theory, prices below the monopoly

[31] Rosse and Panzar (1977), Panzar and Rosse (1987), and Shaffer (1982) showed how to test whether a market is competitive, oligopolistic, or monopolistic using information on shifts in revenue in response to shifts in factor prices. To estimate the actual degree of market power, however, one must have additional information or make some strong assumptions such as Hall's constant returns to scale assumption.

[32] Suppose that the industry has a demand curve with a constant elasticity of ϵ and a constant marginal cost. A monopoly sets price equal to $1/(1 + 1/\epsilon)$ times the constant marginal cost (as can be shown by rearranging Equation 8.2). If ϵ is -2, then the price is twice the marginal cost. If, holding ϵ constant, demand shifts out so that one more unit is sold, revenues rise by *p*, but total cost increases by *MC*, which is only half of *p*.

[33] Domowitz et al. (1988) do not find that concentration plays an important and statistically significant role in explaining the deviation between price and marginal cost. However, Shapiro (1987), using a variant of Hall's method, does find a strong relation between margins and concentration.

level are due to failures to enforce the cartel fully. In this story, market structure matters. For example, the more firms in an industry, the harder it is to detect cheating by any one firm, so more cheating occurs, and the average price is lower.

Game theorists model Stigler's insight as a supergame over repeated static games. In one version, random fluctuations in price due to fluctuations in demand or supply costs could make "cheating" by cartel members hard to detect because the price fluctuations could be due to either cheating or shifts in economic conditions. To prevent firms from cheating, all cartel members agree that if the market price drops below a certain level—a "trigger price"—each firm will expand its output to the precartel level for a certain period of time and prices will fall as a result. If firms expect other firms to stick to this agreement, a firm that cut its price might gain in the extremely short run, but would lose in the end because of the destruction of the cartel by this predetermined punishment mechanism (Chapter 5).

Porter (1983a), Lee and Porter (1984), and Ellison (1994) used this theory to estimate a model of 1880s railroad cartel behavior. Comparing high and low price periods, Porter finds that the cartel increased its rate by over 60 percent during periods of successful collusion. See Example 5.6.

Dynamic Models with Adjustment Costs. If firms have substantial adjustment costs from training new workers, from storage of inputs or outputs (inventories), or in accumulating capital, they must plan their actions over many periods if they are to maximize long-run profits. For example, if the firm must pay compensation to laid-off workers (an adjustment cost), the firm hires fewer workers in period t if it believes demand will be lower in period $t + 1$. Similarly, firms' costs may fall over time if there is **learning by doing** (costs fall with production because workers become more skilled at their jobs due to experience or as better ways of producing are discovered); actions by a firm in this period affect its costs and profits in later periods.[34]

Pindyck (1985) showed that, in a dynamic setting, a mechanical application of the Lerner Index for each period can be misleading. In the intertemporal case, neither the short-run demand elasticity nor the Lerner Index provides a meaningful measure of monopoly power. One solution is to discuss the steady-state price-cost margin (the margin that eventually would be reached and that would persist if there were no further cost or demand shocks) or to compare the path of price or quantity with respect to the path under the price-taking assumption.

The game-theoretic literature abounds with dynamic models of oligopoly that are too general to be usable in estimation. To estimate these models practically, further restrictions have to be imposed. Roberts and Samuelson (1988) use a dynamic oligopoly model with reasonably general functional forms to reject the hypothesis that the ciga-

[34] Analogous to dynamic models with adjustment costs are those where demand today depends on quantities in previous periods. Some marketing studies attempt to estimate demand curves with this property, as do studies of durable goods such as aluminum (Suslow 1986b). Similarly, in pumping oil, the costs today depend on how much was pumped in the past and price is expected to rise at the rate of interest (according to the Hotelling formula), so empirical studies of oil reflect these dynamic issues as well (Matutes 1985).

rette market is competitive. With their general functional form, however, they cannot estimate the degree of market power. Karp and Perloff (1989a, 1993a) used a dynamic oligopoly model with a linear demand curve and quadratic costs of adjustment to estimate steady-state price-cost margins for the international coffee and the international rice export markets. For recent work on dynamic oligopoly, see the references cited in Chapter 6, especially Ericson and Pakes (1995, 1998), Fershtman and Pakes (2000), and McGuire and Pakes (1994).

Value of Modern Approaches to Measuring Performance

The modern approaches have three major advantages over the SCP approach. First, they estimate the market performance rather than use an accounting proxy. Second, they use changes in exogenous variables (wages, taxes, demand growth) to explain variations in performance rather than endogenous variables such as concentration ratios and advertising. Third, they are based on maximizing models for individual industries so that hypotheses about behavior can be tested. Their key disadvantage is that many of these models require making detailed assumptions about the shapes of the supply and demand curves and about oligopoly behavior. Moreover, none of the modern approaches that we have discussed focuses on the use of cross-sectional variation across industries to make any predictions as to what factors cause competition to differ across industries. It was the search for such factors that was at the heart of the SCP approach and central to Sutton's approach.

SUMMARY

The empirical relationship between measures of performance, such as price-cost margins, and market structure, such as concentration and entry barriers, is not clear. Serious measurement problems can plague such structure-conduct-performance (SCP) studies. Accounting measures of performance may fail to measure economic profits or costs accurately, especially when long-lived capital assets are present. Concentration ratios for individual industries can be measured accurately, but make sense only when the individual industries constitute a relevant economic market. Finally, the measurement of barriers to entry is often subjective and typically fails to distinguish between long-run barriers to entry and the speed with which entry can occur.

Studies relating measures of industry performance to concentration and barriers to entry across industries suffer from several conceptual problems. A statistically significant relationship between concentration and performance would not necessarily imply that concentration caused price to be above the competitive level. An alternative explanation is that firms become large (concentration rises) because they are efficient. If so, within an industry, profits of the largest firms are higher than those of the smallest. The empirical results indicate either no effect or a small positive effect of concentration and barriers to entry on performance, but this effect is often statistically insignificant. Sutton and his collaborators have produced research that addresses many of the criticisms of the SCP approach and simultaneously uses industry information to make predictions about industry concentration.

Studies of individual industries can avoid many, though not necessarily all, of the conceptual problems of older SCP cross-sectional studies. Such studies tend to find a small but statistically significant effect of concentration on industry measures of performance, such as price.

Modern studies statistically estimate the price-cost margin for a particular industry rather than rely on accounting proxies. These studies have their own disadvantages: Researchers typically have to make detailed assumptions about demand, cost functions, or oligopoly behavior. Many of these industry studies find substantial margins. These methods have not yet been used to explore in detail the relationship of industry structure to the degree of deviation from perfectly competitive behavior.

PROBLEMS

1. An investor decides to purchase a business. He hires a consultant to help him find a good one. The consultant advises to find a business that faces no competition because such a business can earn rates of return in excess of those businesses that do face competition. Is this good advice?
2. The supply of medical doctors cannot be expanded quickly because it takes years to train them. If a hospital wishes to enter a new market, does it face a barrier to entry?
3. Concentration ratios are typically a firm's share of *domestic* production. If the United States engages in more international trade, will such concentration measures lose meaning? Could this effect explain the vanishing of the price-concentration effect over time?
4. (*Difficult*) Evaluate the following argument: "There exist demand curves for which a monopoly would pass along cost increases in price on a one-for-one basis. Therefore, nothing can be inferred about the competitiveness of an industry by comparing price changes to cost changes." In your evaluation, see if you can derive a demand curve with the stated properties (Bulow and Pfleiderer 1983).
5. Distinguish between zero profits and a price-cost margin that equals zero.
6. Suppose that the demand function is $Q = s/p$, where Q is the total quantity demanded, s is a measure of the size of the market, and p is the price of the homogeneous good. Let F be a firm's fixed cost and m be its constant marginal cost. If n firms compete in a Cournot model, calculate the price, p, a typical firm's output, q, and a typical firm's profit, π.

   a. Prove that:

   i. $p = m\left[1 + \dfrac{1}{n-1}\right]$,

   ii. $q = \dfrac{s}{m}\dfrac{n-1}{n^2}$, and

   iii. $\pi = s/n^2 - F$.

   b. If entry is free, what does n equal?

   c. What happens to equilibrium concentration, $1/n$, as s increases?

   d. What happens to equilibrium firm size as s increases?

Answers to odd-numbered problems are given at the back of the book.

APPENDIX 8A

Relationship Between the Herfindahl-Hirschman Index (HHI) and the Price-Cost Margin

An oligopoly consists of n identical firms that produce a homogeneous product. Each Firm i chooses its output, q_i to maximize its profits,

$$\pi_i = p(Q)q_i - mq_i,$$

where m is the constant marginal (and average variable) cost for each firm, and p, the price, is a function of total industry output, $Q = nq_i$.

The firms play Cournot (see Chapter 6), so each firm's first-order condition—which is obtained by setting the derivative of profits with respect to q_i equal to zero—is that marginal revenue equals marginal cost:

$$MR = p + q_i p' = m = MC, \tag{8A.1}$$

where p' is the derivative of price with respect to Q. Rearranging the terms in Equation (8A.1), this expression can be expressed in terms of the Lerner Index:

$$L \equiv \frac{p - m}{p} = -\frac{p'Q}{p}\frac{q_i}{Q} = -\frac{s_i}{\epsilon} = -\frac{1}{n\epsilon}, \tag{8A.2}$$

where $s_i \equiv q_i/Q = 1/n$ is the output share of Firm i and $1/\epsilon = (p'Q)/p$ is the reciprocal of the elasticity of demand. Because all firms are identical, Equation (8A.2) holds for every firm in the industry.

As Cowling and Waterson (1976) show, the industry average of firms' price-cost margins using share weights is

$$\sum_i s_i \frac{p - m}{p} = -\frac{\sum_i s_i^2}{\epsilon} \equiv -\frac{HHI}{\epsilon},$$

where HHI is the Herfindahl-Hirschman Index. That is, the HHI divided by the absolute value of the market demand elasticity equals the weighted average of the firms' price-cost margins.

APPENDIX 8B

Identifying Market Power

Under what conditions can the price-cost margin be determined if we cannot observe marginal cost directly? One approach to answering this question involves estimating a complete model of the market where the shapes of the demand and marginal cost curves are specified and profit-maximizing behavior is assumed.[1]

To illustrate this approach, suppose that an industry consists of a number of identical firms that produce a homogeneous product. The demand curve is $p(Q; Z)$, where p is the single price in the market, Q is output, and Z is another variable that affects demand, such as income or the price of a substitute.

Because industry revenues are $R \equiv p(Q; Z)Q$, we define the effective (or perceived) marginal revenue as

$$MR(\lambda) = p + \lambda p_Q Q,$$

where λ is a parameter to be estimated and $p_Q \equiv \partial p/\partial Q$. If the industry is monopolized, $\lambda = 1$ and effective $MR(1)$ is the usual MR measure: $p + p_Q Q$. If the firms in the industry are price takers, $\lambda = 0$ and effective $MR(0)$ equals price. Various other oligopolistic and monopolistically competitive market structures produce a λ that lies strictly between 0 and 1.

The profit-maximization or optimality condition is that effective marginal revenue equals marginal cost: $MR(\lambda) = MC$. As a result, λ is a measure of the gap between price and marginal cost. That is, the Lerner's Index is

$$L \equiv \frac{p - MC}{p} = -\frac{\lambda p_Q Q}{p} = -\frac{\lambda}{\epsilon},$$

where ϵ is the market elasticity of demand. This expression is very similar to those derived in Appendix 8A that depend on the number of firms, the market share, or the Herfindahl-Hirschman Index.

As an example, suppose that the demand curve has the particular linear form

$$p = \alpha_0 + \alpha_1 Q + \alpha_2 Z + \alpha_3 ZQ + \epsilon_1, \tag{8B.1}$$

[1] The following discussion of the role of market demand shocks in identifying market power is based on Just and Chern (1980), Bresnahan (1982), and Lau (1982).

so that the effective marginal revenue is

$$MR(\lambda) = p + \lambda\, p_Q Q = p + \lambda(\alpha_1 + \alpha_3 Z)Q. \tag{8B.2}$$

A profit-maximizing firm sets its effective marginal revenue equal to its marginal cost. If its marginal cost curve is linear in Q and factor price W,

$$MC = \beta_0 + \beta_1 Q + \beta_2 W + \epsilon_2,$$

its optimality equation, $MR(\lambda) = MC$, can be written as

$$p = \beta_0 + (\beta_1 - \lambda\alpha_1)Q - \lambda\alpha_3 ZQ + \beta_2 W + \epsilon_2. \tag{8B.3}$$

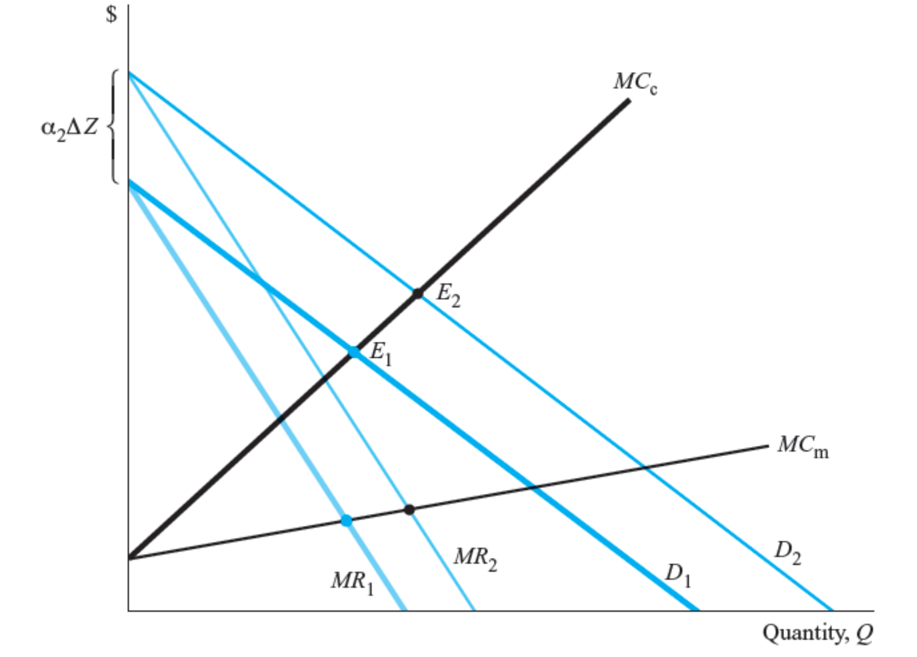
Using the appropriate statistical techniques, one can regress p on a constant, Q, ZQ, and W to obtain estimates of the coefficients in Equation 8B.3. By dividing the estimate of the coefficient on the ZQ term, $-\lambda\alpha_3$, from Equation 8B.3 by the estimate of α_3 from the demand Equation 8B.1, one obtains an estimate of the market structure parameter λ. The reason that one can identify λ is that the demand and MR curves rotate with Z due to the ZQ interaction term, which affects where the MR curve intersects the MC curve. Alternatively, if we know MC, we can use the information about price from the demand curve to determine λ. Rotating the demand curve leaves the level of demand unchanged at the rotation point, but changes the elasticity of demand. As the elasticity of demand changes, the price changes, which allows us to estimate λ.

If there is no ZQ term (that is, if $\alpha_3 = 0$) in the demand curve, λ may not be identified. The only remaining term with a λ in Equation 8B.3 is $(\beta_1 - \lambda\alpha_1)Q$. Although we know α_1 from the demand equation, that is not enough to identify λ because the estimated coefficient also depends on β_1 (the unknown slope of the MC curve).

The need for the demand curve to rotate is illustrated in Figure 8B.1.[2] Initially, the researcher observes the market equilibrium, E_1, price and quantity. The researcher estimates the demand curve D_1 (and, hence, can infer the marginal revenue curve, MR_1) but does not directly observe costs. The observed equilibrium, E_1, is consistent with a competitive industry structure and a marginal cost curve MC_c, where the equilibrium, E_1, is determined by the intersection of MC_c and D_1. It is also consistent with a cartelized market structure and a lower marginal cost curve, MC_m, where the quantity associated with E_1 is determined by the intersection of MC_m and MR_1.

[2] Lau (1982) shows that virtually any functional form for the demand curve leads to identification except the two most commonly used forms: linear or log-linear. If one wants to use a basically linear specification, one must add an interaction term, a squared term in output, or something else that adds some nonlinearity and allows the demand curve to rotate. Even if one does that, there is an additional serious problem with the linear specification: see Perloff and Shen (2001).

FIGURE 8B.1 Not Identified: Parallel Shift of the Demand Curve



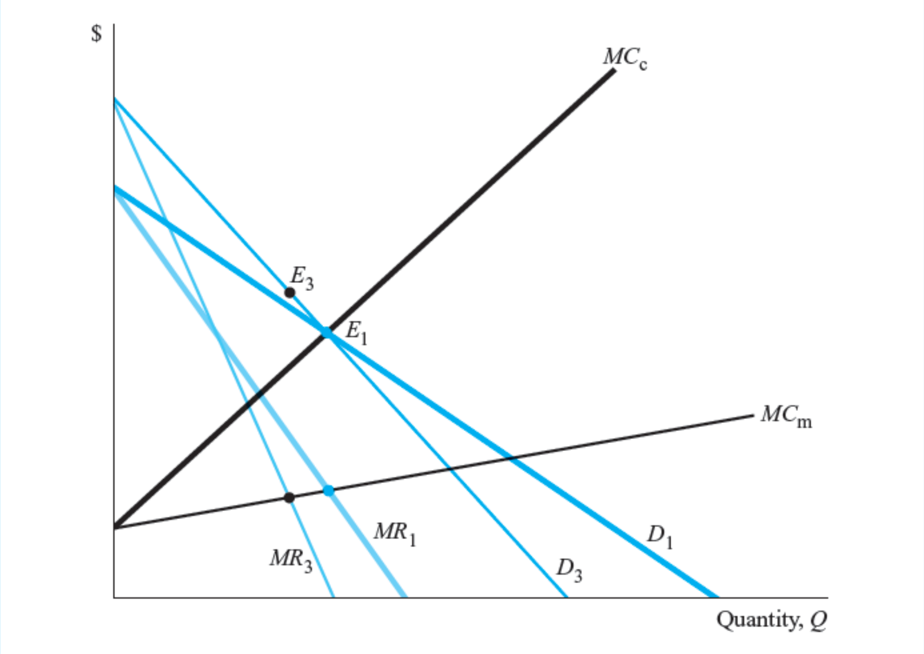
If $\alpha_3 = 0$, and Z increases by ΔZ, the intercept of the demand curve shifts up by $\alpha_2 \Delta Z$, as shown for the new demand curve, D_2. The new equilibrium, E_2, is still consistent with either of the two marginal cost curves. Thus, the researcher cannot determine from this shift in Z if the industry is competitive or cartelized.

In contrast, if $\alpha_3 \neq 0$, a shift in Z reveals λ. In Figure 8B.2, when Z increases, the new demand curve, D_3, rotates (for graphical simplicity, D_3 is rotated around the original equilibrium point). If the industry is competitive and the marginal cost curve is MC_c, the new equilibrium on D_3 remains E_1; whereas, if the industry is cartelized and the marginal cost curve is MC_m, the new equilibrium on D_3 is E_3. Thus, whether or not the equilibrium shifts reveals whether the market is competitive.

Anything (not just variables in the market demand curve) that causes the residual demand curve facing a firm to rotate can identify λ. For example, a dominant firm's residual demand curve is the market demand curve minus the supply of a competitive fringe. If the fringe supply curve rotates, the residual demand curve rotates even if the market demand curve does not. Similarly, a shift in an ad valorem tax rate, t, can identify the market structure.

As the chapter shows, information about the shape of the marginal cost curve also can help identify λ. It is possible to identify λ even if the demand curve does not rotate

FIGURE 8B.2 Identified: Rotation of the Demand Curve



($\alpha_3 = 0$), if the marginal cost curve is constant in Q ($\beta_1 = 0$). Because $MC = \beta_0 + \beta_2 W$, marginal cost is a constant in any given period, but that constant shifts with the exogenous factor W over time causing price to change, which allows one to estimate the demand curve. The coefficient on the Q term in Equation 8B.3 is $\beta_1 - \lambda\alpha_1 = -\lambda\alpha_1$, so by knowing α_1 from the demand curve, one can identify λ.

PART THREE

Business Practices: Strategies and Conduct

CHAPTER 9

Price Discrimination

All . . . men have their price. —*Sir Robert Walpole*

Firms in a perfectly competitive market have no discretion in their pricing policies; they must take the market price as given. Most markets, however, are not perfectly competitive, and firms have some discretion over their pricing policies. In order to maximize profits, these firms may use nonuniform pricing: charging customers different prices for the same product or charging a single customer a price that varies depending on how many units the customer buys. When the price paid depends on the amount purchased, the price schedule is nonlinear. Price discrimination refers to any nonuniform pricing policy used by a firm with market power to maximize its profits. One common type of nonuniform pricing is third-degree price discrimination: A firm charges different customers different unit prices for the identical good. Not all price differences are due to price discrimination. For example, some price differences reflect variations in product characteristics or differential costs in supplying the product to various customers.

This chapter examines three main questions:

1. What are the common types of nonuniform pricing?
2. What are the necessary conditions for price discrimination to occur?
3. What are the welfare effects of price discrimination?

The next chapter analyzes more complicated methods of price discrimination, such as two-part tariffs and tie-in sales.

Nonuniform Pricing

In the models of competition, oligopoly, and monopoly discussed so far, the price per unit is the same for all customers. Many firms, however, set nonuniform prices. Numerous magazines offer student discount subscriptions. Many movie theaters offer discounts to senior citizens. The American Economic Association's membership fees vary with a member's income. Products are often packaged with discount coupons that entitle the bearer to purchase the product for a lower price next time. In effect, these coupons allow a firm to charge first-time users a higher price than repeat users. See Example 9.1.

This chapter deals with some simple types of price discrimination while Chapter 10 examines more complicated ones, such as:

1. **Two-Part Tariff:** A firm charges a consumer a fee (the first part of the tariff) for the right to buy as many units of the product as the consumer wants at a specified price (the second part of the tariff). For example, a health club may charge members an annual fee to join the club and additional fees for using particular facilities. Similarly, some amusement parks charge visitors an admission fee and additional fees for each ride.
2. **Quantity Discount:** A firm's price varies with the number of units of the good that a customer buys. Price discounts for large purchases are quite common. Electricity bills are frequently computed according to a *declining-block* schedule, in which the first units of usage incur one charge, and subsequent units incur lower charges.
3. **Tie-in Sale:** A customer can buy one product only if another product is also purchased. A common example of a tie-in sale is the purchase of a durable machine under the condition that the consumer also purchase from the seller all repair services or all repair parts. Firms may sell copy machines under the condition that customers also purchase related supplies (for example, developing chemicals) from the seller. Cameras could be sold under the condition that purchasers buy their film from the seller. Sometimes buyers have no choice and must buy film from the seller; for example, Polaroid cameras use only Polaroid film.[1]
4. **Quality Discrimination:** A firm offers consumers a choice of different quality products at the same price or at prices that do not fully reflect the quality differential. By offering a high-quality, high-priced product that appeals to consumers who place a high value on the product, and a low-quality, low-priced product that appeals to other consumers, a firm can *separate* the two types of consumers and charge high prices to those most

[1] A tie-in sale allows a firm to effectively charge higher prices to consumers who use more of the tied product. Because certain tie-in sales are currently a violation of antitrust laws, they were more common before the antitrust laws were passed. However, as Chapter 19 discusses, the law is unclear as to exactly what constitutes an illegal tie-in sale.

EXAMPLE 9.1 *Coupons*

One common means of price discriminating is to use cents-off coupons or to provide rebates to some but not all consumers. As we show later in the chapter, consumers who increase their purchases the most in response to a special low price (those whose demands are relatively elastic) are the ones who should receive the coupons. Marketing studies find that consumers who have a low cost of transportation (own cars), who have the space to store items (own homes), and who place a low value on time or who have flexible time schedules (a nonworking spouse without small children) are the most likely to take advantage of special promotions. By using coupons, manufacturers can provide discounts to people who are price sensitive and relatively likely to clip and use coupons, without providing discounts to other, typically wealthier people.

The number of coupons distributed was 100 billion in 1981, 200 billion in 1985, and 310 billion in 1994, but only 269 billion in 1995. According to the Promotion Marketing Association (PMA), marketers distributed 336 billion coupons in 2002. Of these, only 3.8 billion were redeemed (for $3.1 billion), the lowest rate since 1980. The average face value of a coupon was 81¢.

One recent innovation is the use of online coupons. Consumers downloaded 242 million coupons in 2002, 111% more than in 2001 but still a tiny fraction of all coupons. (Retailers are concerned that counterfeit coupons are being produced or exchanged over the Internet.)

The PMA reported in 2003 that 79% of all people in the United States use coupons. Coupon use increases with age, going from 71% for 18- to 24-year-olds to 84% for those 65 and older. Richer people are less likely to use coupons, with usage rates falling from 82% for people earning under $25,000 to 76% for those earning over $75,000.

Coupons are most likely to be used for household cleaners, followed by prepared foods, detergents, medications and home remedies, paper products, condiments and gravies, personal soap and bath additives, frozen prepared foods, cereal, and skin care preparations. More than four out of five (80.4%) of coupons were redeemed at grocery stores, with the rest handled by mass merchandisers (9.4%), drug stores (3.7%), and convenience stores (2.4%). The frequency of coupon use while shopping was 76% at grocery stores and 54% at mass merchandisers and drug stores.

Sources: Blattberg et al. (1978); Narasimhan (1984); Philip H. Dougherty, "Advertising: Redemption of Coupons," *New York Times*, July 13, 1988:C19; Eben Shapiro, "Consumers' Use of Shopping Coupons Is Up," *New York Times*, September 30, 1992:C12; George Lazarus, "Coupons Cruising at a Record Clip," *San Francisco Examiner*, October 4, 1992: E5; "Coupon Redemption Rate Down," *Editor & Publisher Magazine*, January 20, 1996:21; M. A. Mariner, "Disappearing Coupons," *San Francisco Chronicle*, January 29, 1997: Food 2; **www.pmalink.org/about/press_releases/release55.asp; www.couponmonth.com/pages/news.htm;** "Coupon Use Is Down," *Beacon Journal*, September 6, 2003; Bob Tedeschi, "E-Commerce Report," *New York Times*, March 17, 2003:C6.

willing to pay them. Therefore, the problem of what range of qualities a monopoly should produce is closely related to the theory of price discrimination.

Not every seller who charges a nonuniform price is price discriminating. There are many other explanations for prices to vary across consumers. (See Lott and Roberts 1991.) For example, a quantity discount may reflect cost savings from dealing with large orders that a manufacturer is passing on to consumers. This chapter and the next, however, focus on explaining how nonuniform pricing can be profitable for a firm with market power.

Incentive and Conditions for Price Discrimination

A firm price discriminates to increase its profits; however, a firm can price discriminate only under certain conditions. We now explain why price discrimination increases profits and what conditions are necessary for it to occur.

Profit Motive for Price Discrimination

Price discrimination is profitable because consumers who value the good the most pay more than if prices were uniform. To show why there is an advantage to price discriminating, we return to a monopoly that charges all customers a single price. The monopoly sets that price so that its marginal revenue equals its marginal cost (see Chapter 4).

Its marginal revenue—the increased revenue that results from selling an additional unit—is the sum of two effects. The first is the increase in revenue from selling one more unit, which is the price, p, that it receives for the last unit. The second is the decrease in revenue on all existing output, $Q\Delta p$, where Δp is the fall in price needed to induce the sale of one more unit.[2] If the monopoly could lower the price on *only* the one additional unit, it would do so as long as the price exceeded marginal cost. It would then earn its current profit plus an additional amount on the last unit. The monopoly would earn additional profit from this price discrimination.

All methods of price discrimination can be viewed as attempts to minimize this second effect on marginal revenue from expanding sales. This chapter and the next identify a variety of pricing policies that are designed to minimize the cost to the monopoly of trying to expand output at a lower price to a particular customer without simultaneously offering the same lower price to all consumers.

[2] Because total revenue equals $p(Q)Q$, for a small change in quantity, marginal revenue equals $p(Q) + Q(dp/dQ)$. In the text, dp/dQ is called Δp.

Conditions for Price Discrimination

Even though all firms would like to price discriminate, many are not able to do so. Three conditions are needed for successful price discrimination.[3]

1. A firm must have some *market power* (the ability to set price above marginal cost profitably); otherwise, it can never succeed in charging any consumer more than the competitive price.
2. The firm must know or be able to infer consumers' willingness to pay for each unit, and this willingness to pay must vary across consumers or units. That is, the firm must be able to *identify* whom to charge the higher price. Similarly, if each individual's demand curve slopes down, the firm may be able to charge a different price for the different units any one consumer purchases (such as $10 for the first unit and $5 for the second unit).
3. A firm must be able to *prevent or limit resales* by customers who pay the lower price to those who pay the higher price. Any attempt to charge one group a higher price than another is doomed to fail if resales are easy. If the group charged the lower price can resell to the other group at a lower price than the monopoly charges them, no one in the latter group would buy directly from the monopoly. Limiting resales is necessary for all types of price discrimination.

Resales

If a firm charges nonuniform prices, consumers who buy at a relatively low price may resell to those facing a relatively high price and thereby render useless the attempt to charge different prices. Similarly, if a firm offers quantity discounts for a product, it must ensure that the discount is not so great as to encourage high-volume purchasers to buy the product and then resell it to those who demand fewer units. There are at least seven reasons why reselling the good may be difficult or impossible for consumers:

Services. Most services cannot be resold. For example, a dentist may charge Lisa a very high price and Jackie a very low price, but it is impossible for Lisa to gain by having Jackie purchase the dentist's services for her. For that reason, price discrimination in services is more likely than price discrimination in industries with tradeable products (Kessel 1958). See Example 9.2.

Similarly, having seen an art show, one cannot transfer the experience to others. In 2001, when Steve Martin's art collection went on display in Las Vegas, the gallery in the Bellagio Hotel charged art lovers a hefty $12 per ticket unless they were Nevada residents, who were charged only $6.

[3] Price discrimination can be practiced by a single firm or a group of firms, such as a cartel. To keep the exposition simple, we discuss the actions of a single firm.

EXAMPLE 9.2 *Thank You, Doctor*

Movies and television shows often portray as great heroes doctors who charge poorer patients lower rates. In very old movies, the country doctor accepts a chicken as payment instead of charging cash. Are doctors selfless creatures or profit maximizers who engage in price discrimination? Certainly some doctors see indigent patients for no fees or trivial fees as an act of charity. Others, however, may be price discriminating.

The Association for Behavioral and Cognitive Therapy publishes a directory of its members who provide therapy in the San Francisco area along with the rates each charges. In the 1990–91 edition, 3 therapists merely state that they use a sliding scale, 10 others show a range, and 31 list a single rate. Many in the latter group often cut their listed fee for some patients.

Of those who listed an explicit range, one stated that he charged between $0 and $120 per session. All the others set their minimum rate at $40 or more. Their maximum rate averaged 1.8 times their minimum.

Warranties. A manufacturer can void a warranty if a product is resold. For example, a manufacturer could say that the warranty on a product is valid only for the first-time purchaser, which imposes a cost on a buyer who purchases a product from a previous buyer.

Adulteration. A manufacturer can adulterate a product to make it unfit for other uses. For example, alcohol is used for drinking (alcoholic beverages) and for medicinal purposes (rubbing alcohol). Suppose alcohol were produced by a monopoly that wanted to charge a higher price to those who drink alcohol and a lower price to those who use alcohol for medicinal purposes. The monopoly could prevent medicinal users from reselling to drinkers by adulterating the medicinal alcohol (adding ingredients that make it unfit for internal consumption yet preserve its medicinal qualities). This particular approach to eliminating resales would not work if medicine consumers were willing to pay more than drinkers, and the manufacturer wanted to prevent the resale of drinking alcohol for medicinal purposes.

Transaction Cost. If consumers incur any large transaction costs to resell the product, resales are less likely. For example, suppose some consumers are mailed coupons that entitle them to purchase a product at a lower price than others. The transaction costs of finding consumers without coupons are too high for it to be worthwhile for consumers with coupons to purchase the product and then resell the product. In many markets, storage costs, search costs, or other transaction costs are too high for any resales to occur.

Two important examples of transaction costs are tariffs (a government tax on imported goods) and transportation costs. A manufacturer that wants to charge a high price in the United States and a low price in Europe would have to worry about resale from Europe to the United States. However, a large tariff or transportation cost that must be paid by anybody importing the product from Europe to the United States reduces or eliminates resales.

Laws sometimes allow a company to charge more for its product in one country than in another by preventing others from shipping the good from the low-cost country to the other. That is, these laws prevent price arbitrage (reselling to profit from differential prices). See Example 9.3.

Contractual Remedies. A firm may contractually forbid resale as part of its terms of sale. For example, many universities and colleges arrange for students and faculty members to purchase computers at lower than market rates. To buy at this reduced rate, one might have to sign a contract that forbids resale. If restrictions on resale are not legally binding or not easily enforceable, such contractual clauses may not prevent resales.

Vertical Integration. Suppose a manufacturer wants to sell aluminum ingots to producers of aluminum wire at a lower price than it charges producers of aluminum aircraft parts. If the manufacturer did charge two different prices, the wire producers would resell their ingot to the aircraft producers. The ingot manufacturer may choose to produce aluminum wire. A firm that produces at more than one stage of a production process is said to be *vertically integrated.* The vertically integrated firm can charge final consumers of aluminum wire a low price (that is, effectively charge and pass along a low price for aluminum ingot to its own aluminum wire division) and still charge the aircraft producers a high price for aluminum ingot with no fear of resale. Resale does not arise for two reasons. First, the monopoly controls the actions of its aluminum wire division and does not allow it to resell the aluminum ingot. Second, it is cheaper for aircraft producers to purchase aluminum ingot rather than purchase aluminum wire and transform the wire back into ingots. Vertical integration prevents resale in a way similar to the adulteration argument given above. See Example 9.4.

Government Intervention. The government can enact laws that allow firms in a competitive industry to act collectively to prevent resale. For example, government regulations control how much of an orange grower's crop can be sold as fresh fruit and how much as processed (Appendix 9A describes government programs in agriculture that foster price discrimination). The remainder of this chapter assumes that a firm can prevent or control resale of its product and investigates the ways in which firms price discriminate.

Types of Price Discrimination

There are many methods for charging nonuniform prices. This section examines some of the simplest ones. The more complicated ones are discussed in the next chapter. We first study *perfect* or *first-degree* price discrimination, in which consumers are left with no consumer surplus (the value to consumers in excess of the purchase price). Then we

EXAMPLE 9.3 *Halting Drug Resales from Canada*

Pharmaceutical companies price discriminate across countries. The prices of many popular drugs are substantially lower than in the United States in virtually every other country in the world. Zoloft, an antidepression drug, sells for one-third the U.S. price in Mexico and about half in Luxembourg and Austria. Many well-known brand-name drugs sell in Canada for one-third to one-half the lowest price available in the United States.

These price differences reflect price discrimination by pharmaceutical firms. Sometimes the lower prices in other countries are due to differences in incomes, patent laws, and legal liabilities. However, frequently, regulations in other countries are responsible for the relatively low prices.

U.S. pharmaceutical companies are horrified about the possibility that resales—where drugs they exported from the United States at relatively low prices are reimported—will drive down U.S. prices. In 2003, the U.S. House of Representatives passed a bill to permit imports, but because it has not become law to date, such imports remain illegal. Nonetheless, U.S. senior citizens have taken many well-publicized bus trips across the Canadian and Mexican borders to buy drugs at lower prices; and many Canadian, Mexican, and other Internet sites offer to ship drugs to the United States. According to various estimates, only about 1 percent to 3 percent of U.S. drug expenditures went to imported drugs, but the fraction is growing.

Some drug companies, among them GlaxoSmithKline and Pfizer, are trying to reduce imports by cutting off Canadian pharmacies that ship south of the border. Wyeth and AstraZeneca report that they watch Canadian pharmacies and wholesale customers for spikes in sales volume that could indicate imports, and then restrict supplies.

The drug companies have also pressured the U.S. Food and Drug Administration (FDA) to help prevent imports. To date, the FDA has not enforced restrictions on purchases by individuals. However starting in 2003, the FDA took several steps to reduce imports. The FDA raised the specter that imported drugs were not as safe as those purchased in the United States (although it has provided little evidence to date to support that claim). The agency sent threatening letters to various state attorneys general saying that state agencies that imported Canadian prescription drugs would be violating federal law. It completed a sting operation targeting the supplier of Canadian drugs to the employee insurance program of the City of Springfield, Massachusetts (which reported it could save $4 to $9 million a year by ordering drugs through Canada). It took actions to close a chain of Canadian drugstores that ship drugs to the United States (Rx Depot has 85 stores in 26 states and operates other stores in Canada under the name Rx of Canada).

Sources: Tim Harper, "Canada's Drugs 'Dangerous'," *Toronto Star*, August 28, 2003:A12; Christopher Rowland, "FDA Sting Targets Medicine Supplier; Springfield Uses Firm to Get Canadian Drugs," *Boston Globe*, August 28, 2003:C1; Tony Pugh, "Canadian Online Pharmacies Struggle to Find Suppliers," *San Diego Union-Tribune*, September 7, 2003:A-3; Gardiner Harris, "U.S. Moves to Halt Import of Drugs from Canada," *New York Times*, September 10, 2003:C2.

EXAMPLE 9.4

Vertical Integration as a Means of Price Discrimination: Alcoa Shows Its True Metal

Alcoa had considerable monopoly power in the production of primary aluminum ingot from 1888 to 1930 due to tariffs that protected it abroad and by its control of bauxite lands at home. Further, the disruptions of World War I slowed entry of new firms.

The traditional view of why Alcoa forward integrated into processing activities (bought firms in these industries) was to demonstrate the technical and commercial feasibility of new aluminum products. Recent research indicates, however, that Alcoa probably vertically integrated in order to price discriminate.

Explicit price discrimination was not possible, because aluminum ingots are easy to handle and hence easy to resell. Alcoa overcame this problem by vertically integrating into some industries that purchased aluminum ingots.

Suppose that there are only two downstream industries (or groups of industries) that buy aluminum, and that Industry 1's demand for the product is less price-elastic than Industry 2's demand. Alcoa wanted to charge a higher price to Industry 1 than Industry 2. If it did so, however, Industry 2 firms would resell the aluminum to the high-price industry.

If Alcoa vertically integrated into the low-price industry (that is, Alcoa buys Industry 2), it could prevent resales by its own subsidiary. Moreover, because Alcoa supplied its subsidiary with its product internally, the only industry Alcoa explicitly sold ingot to was the high-price industry.

Alcoa only forward integrated into some industries that used primary aluminum. As predicted by this theory, Alcoa integrated into the high-elasticity industries. The five uses of aluminum listed in the following table represented more than 90 percent of Alcoa's output during most of this period. Of these uses, iron and steel production and aircraft manufacturing had the most inelastic demands because of a lack of good substitutes for aluminum in their production process. Alcoa did not integrate into these industries. Alcoa did integrate into the other, relatively elastic industries. Because there were many substitutes for aluminum in the manufacture of cookware (such as tin, glass, steel, iron, and so forth), electric cable (copper), and automobile parts (various metals), their demand for aluminum was relatively elastic.

Major Industries Using Aluminum

Industry	Elasticity of Demand for Aluminum	Integrated by Alcoa?
Cookware	Elastic ($\epsilon \approx -1.6$)	Yes
Electric cable	Elastic (copper substitute)	Yes
Automobile parts	Elastic ($\epsilon \approx -1.5$)	Yes
Iron and steel	Inelastic (no substitutes)	No
Aircraft	Inelastic (no substitutes then)	No

Source: Perry (1980).

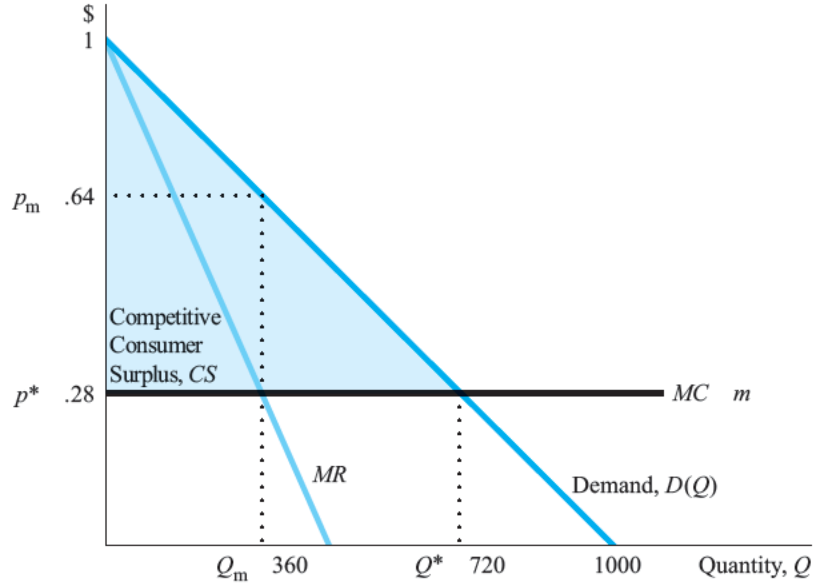
study *third-degree* price discrimination, in which each group of consumers faces its own price per unit. Chapter 10 examines *second-degree* discrimination, in which the price per unit depends on the number of units purchased. In second-degree and third-degree price discrimination, the firm fails to capture all of the consumer surplus.

Perfect Price Discrimination

The purpose behind all methods of price discrimination is to capture as much consumer surplus (Chapter 3) as possible. Perfect price discrimination or first-degree price discrimination occurs when a monopoly is able to charge the maximum each consumer is willing to pay for each unit of the product.

Each Consumer Buys One Unit. Suppose that each consumer wants one unit of a product, but consumers are willing to pay a different amount for it, so that the demand curve slopes downward as shown in Figure 9.1. Assume that the firm knows the maximum amount that each consumer is willing to pay. If it can prevent resales, the firm charges each customer the maximum that person is willing to pay so the customer is left with no consumer surplus. The firm sells to any consumer who will pay at least as much as the firm's (for simplicity, constant) marginal cost, $MC = m$. That is, the perfectly discriminating monopoly sells Q^* units and the marginal consumer pays p^* as shown in Figure 9.1.

FIGURE 9.1 Competitive, Nondiscriminating Monopoly, and Perfectly Discriminating Monopoly



A competitive industry would also sell Q^* units and charge everyone a single price, p^*, which equals the marginal cost. Thus, a competitive industry and a perfectly discriminating monopoly charge the marginal consumer the same price, p^*, and sell the same total quantity, Q^*.[4] The difference is that the perfectly discriminating monopoly charges all but the marginal customer more than p^* so that there is no consumer surplus. Consumer surplus is maximized under competition (the area under the demand curve and above p^* in Figure 9.1) and eliminated (and captured) by a perfectly discriminating monopoly. Therefore, perfect price discrimination entails no efficiency loss (the price on the last purchase still equals marginal cost), but does affect the distribution of income.[5]

A nondiscriminating monopoly charges a single price, p_{m}, and produces Q_{m}, where its marginal revenue, *MR,* equals its marginal cost, *MC,* as shown in Figure 9.1. Consumers have a small amount of consumer surplus (the area under the demand curve and above p_{m}), which is smaller than the consumer surplus under competition. The perfectly discriminating monopoly produces more than the nondiscriminating, single-price monopoly. The single-price monopoly produces too little; it is inefficient.

The perfectly discriminating monopoly sells more than the nondiscriminating monopoly because it makes an incremental profit on each additional sale. By charging each consumer a different price, the perfectly discriminating monopoly avoids the adverse second effect on marginal revenue that a nondiscriminating monopoly faces. That is, the discriminating monopoly does not decrease revenues on the first units sold when it sells additional units at a lower price. The effect on marginal revenue of eliminating the second effect is that the demand curve becomes the marginal revenue curve.[6] The monopoly lowers price to only the additional customer and so gains that price as an increase in its revenues from selling one more unit.

Each Consumer Buys More Than One Unit

So far, we have assumed that customers differ in their willingness to pay and that each customer demands only one unit no matter how low the price. Now consider how perfect price discrimination works when consumers are identical but demand more units

[4] We ignore the effects of redistributing income through price discrimination. A discriminating monopoly earns higher profits and consumers have less income than under competition.

[5] However, see Edlin, Epelbaum, and Heller (1998) for a general equilibrium analysis.

[6] The perfectly discriminating monopoly picks Q so that its profit is maximized. Its profit is the area (revenues) under its inverse demand curve, $p(Q)$, less its costs, $C(Q)$:

$$\pi(Q) = \int_0^Q p(q)\mathrm{d}q - C(Q).$$

Its first-order condition for an interior profit maximum is

$$p = p(Q) = C'(Q).$$

That is, profit is maximized at the quantity Q where price equals marginal cost. The second-order condition is that $p'(Q) - C''(Q) < 0$. That is, the slope of the marginal cost curve is greater than the slope of the demand curve.

as price falls. Suppose that each consumer is identical to all others and has the downward-sloping demand curve for the product. We now assume that the demand curve in Figure 9.1 reflects each consumer's curve rather than the market aggregate. Marginal cost is still assumed to be constant at *m.*

A perfectly discriminating monopoly charges a different price for each *unit* of the product that is sold and thus, by charging *quantity-dependent* prices, extracts all the consumer surplus from each customer. The monopoly charges a high price for the first unit consumed, a lower price for the next unit, and so on until it charges *m,* the marginal cost, for the last unit. That is, the monopoly sets its (marginal) price schedule equal to each customer's demand curve.

An alternative and equivalent method of perfect price discrimination would be to charge an optimal *two-part tariff,* where each customer pays a lump-sum fee for the right to purchase plus a per-unit charge of *m* for each unit consumed regardless of how many units each consumer purchases. If a customer's consumer surplus is *CS* (Figure 9.1) when price is *m,* then the monopoly sets the lump-sum fee equal to *CS.* The consumer is indifferent between buying or not because the monopoly captures all the consumer surplus. This pricing method yields the competitive output and generates the same profit for the monopoly as it would earn if it perfectly price discriminated. A similar approach used by unions is discussed in Example 9.5.

If each consumer has a downward-sloping demand curve but consumers differ, the monopoly charges each consumer *m* per unit consumed but charges each one a different lump-sum fee in order to extract all of the consumer surplus. Of course, a monopoly may not have detailed enough knowledge about each consumer's demand curve to design a pricing policy that captures all the consumer surplus of each consumer. If the monopoly lacks this detailed information, it may find it profitable to use the more complicated pricing policies described in the next chapter. However, sometimes it is possible to monitor customers to determine the values they place on products. For example, a firm that rents out copy machines may use a meter in the copy machine to keep track of the number of copies each customer makes and then set the rent depending on the number of copies made. This method of pricing maximizes profit if those who make the most copies value their machine the most.

Because perfect price discrimination requires detailed knowledge about individual buyers, it is more likely to occur (or be attempted) when one-on-one bargaining occurs. For example, a car salesperson may ask potential buyers about their jobs, where they live, and where else they have shopped in an effort to estimate the maximum they are likely to spend. Similarly, doctors may be able to successfully price discriminate if they can identify the wealthy people in their area (see Kessel 1958 and Example 9.2).

Different Prices to Different Groups

A firm that does not have enough information to identify each customer and determine what each one is willing to pay is unable to practice first-degree price discrimination and extract all consumer surplus. The firm may have, however, enough information to imperfectly price discriminate.

Suppose a firm can determine whether a particular customer belongs to one group rather than another where the demand elasticities for the aggregate demand curves of

EXAMPLE 9.5 *A Discriminating Labor Union*

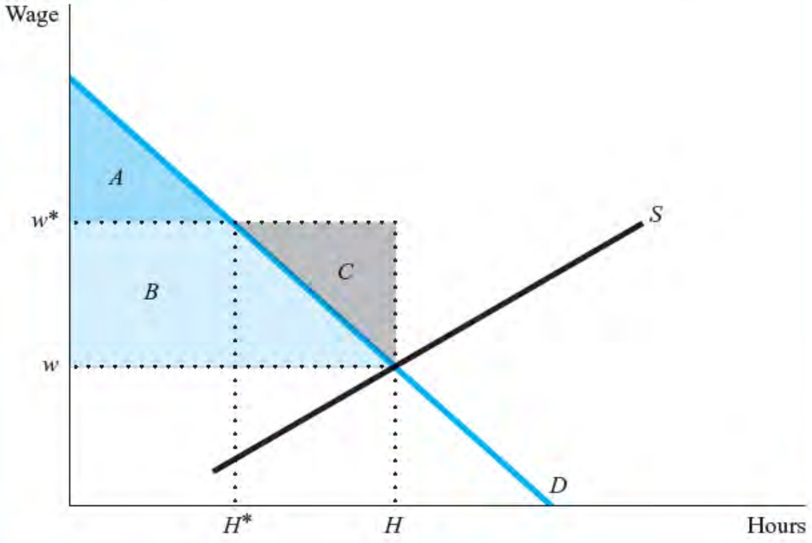
A powerful labor union may be able to act as a perfectly discriminating monopoly and capture all the consumer surplus. Because it is difficult to charge different prices for each hour of labor services, unions use an alternative approach. The union sets both a wage and a minimum number of hours (Leontief 1946).

As shown in the diagram, if the labor market were competitive, a wage of *w* would be charged, and *H* hours of labor services would be sold. Purchasers of labor services would have consumer surplus equal to areas *A* and *B*. If, in contrast, all workers belong to a union, and the union acts like a perfectly discriminating monopoly, it charges a wage equal to the demand curve for each hour of labor services it sells (so that the wage for the last hour it sells is *w*), and it captures all the consumer surplus.

Alternatively, the union could set a single wage, *w**, and a minimum number of hours, *H*, and receive the same total amount of compensation. The union offers the firms the following choice: You may buy *H* hours of labor at *w** (so the total wage bill is *Hw**) or you may buy no hours at all. As shown in the diagram, if the union only set the wage at *w** and did not set a minimum number of hours, firms would purchase fewer hours (*H**). The only reason that the firms agree to buy so many hours at this wage is that the alternative is to buy no labor services at all.

As the diagram shows, the firms receive consumer surplus equal to area *A* for the first *H** hours, and then have negative consumer surplus (equal to area *C*) for the next *H* − *H** hours. The union receives profits above the competitive level equal to areas *B* and *C*. If *w** is set appropriately so that area *A* equals area *C*, the union makes as much profit with this scheme as it would if it perfectly price discriminated.

The Longshoremen's union used this technique (U.S. Department of Labor 1975). Two-thirds of the union contracts in the transportation industry (excluding railroads and airplanes) had wage-employment guarantees in the early 1970s. In contrast, only 11 percent of union contracts in all industries had such guarantees.



the two groups differ. If it is possible to prevent (or limit) resale between the two groups, and if the firm knows the aggregate demand curve of each group, then it is profitable to set different prices for the two groups. The monopoly is practicing **third-degree price discrimination**: It charges consumers in different groups different unit prices. For example, if high transaction costs prevent resale, a firm could charge consumers in California higher prices than those in New York.

If the monopoly has a constant marginal and average cost of m, its profit, π, is

$$\pi = [p_1(Q_1) - m]Q_1 + [p_2(Q_2) - m]Q_2, \tag{9.1}$$

where $p_1(Q_1)$, the inverse demand curve, is the price that the monopoly must charge Group 1 if it is to sell it Q_1 units and $p_2(Q_2)$ is, similarly, the inverse demand curve for Group 2. That is, p_1 depends only on the number of units sold to that group, Q_1 (and not Q_2) and $p_2(Q_2)$ depends on only Q_2. Total profit is $\pi = \pi_1 + \pi_2$, where π_i, the profit from sales to Group i ($i = 1, 2$) is $[p_i - m]Q_i$. That is, π_i is the profit per unit sold to Group i, $[p_i - m]$, times the number of units sold to that group, Q_i.

The monopoly maximizes its total profit (Equation 9.1) by maximizing its profits from sales to each of the groups separately. The monopoly charges the same price to every member of a given group. Thus, we can determine how the monopoly sets its price to each group by using the same method that we used for a nondiscriminating monopoly in Chapter 4. That is, the monopoly maximizes its profit when its marginal revenue from sales to Group i, MR_i, equals its marginal cost of producing that last unit, m:

$$MR_1 \equiv p_1\left(1 + \frac{1}{\epsilon_1}\right) = m, \tag{9.2a}$$

$$MR_2 \equiv p_2\left(1 + \frac{1}{\epsilon_2}\right) = m, \tag{9.2b}$$

where ϵ_i is the elasticity of demand for Group i, so that the marginal revenue for Group i equals $p_i(1 + 1/\epsilon_i)$ as discussed in Chapter 4.[7]

Because the marginal cost, m, is the same in both Equations 9.2a and 9.2b, it follows that the profit-maximizing monopoly equates marginal revenue across the two markets: $MR_1 = MR_2$. In the optimal solution, if the monopoly sells one less unit in Market 1 and one more unit in Market 2 or vice versa, revenues must be unaffected. Otherwise it would pay to reallocate sales between the two markets, which implies that

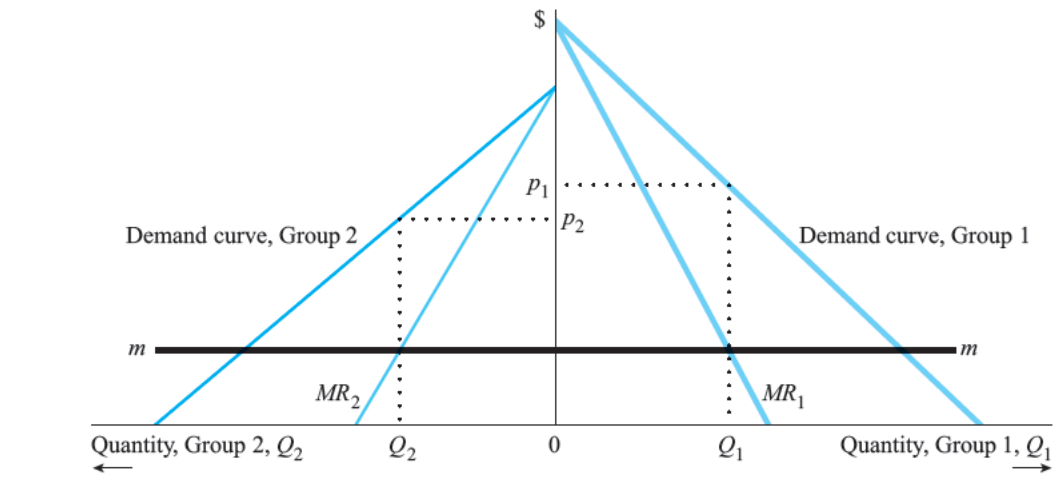
[7] The first-order conditions for a profit maximization are obtained by differentiating Equation 9.1 with respect to Q_1 and Q_2:

$$MR_i = p_i + Q_i\,p_i' = m = MC, \qquad i = 1, 2.$$

By multiplying and dividing by p_i, we obtain

$$MR_i = p_i\left(1 + p_i'\frac{Q_i}{p_i}\right) = p_i\left(1 + \frac{1}{\epsilon_i}\right).$$

FIGURE 9.2 Price Discrimination



profits were not maximized. Equating the common marginal revenue to marginal cost yields maximum profits.[8]

The pricing decision of the discriminating monopoly is illustrated in Figure 9.2. The figure shows the demands of two consumer groups. The demand curve for Group 2 on the left side of the diagram is "flipped" so that it is read in the opposite direction from the Group 1 demand on the right side of the diagram. Setting the marginal revenue of each demand curve equal to a constant marginal cost m yields the optimal pricing and output decision (p_1, Q_1) and (p_2, Q_2).

We can rewrite Equations 9.2a and 9.2b as

$$\frac{p_1 - m}{p_1} = -\frac{1}{\epsilon_1}, \tag{9.3a}$$

$$\frac{p_2 - m}{p_2} = -\frac{1}{\epsilon_2}. \tag{9.3b}$$

[8] If marginal cost is not constant, cost varies with total output, so that Equation 9.1 is

$$\pi = p_1(Q_1)Q_1 + p_2(Q_2)Q_2 - C(Q_1 + Q_2),$$

where marginal cost is $C'(Q_1 + Q_2) = MC$. The optimal price and output for one consumer group depend on the optimal price and output of the other consumer group. The optimal pricing and outputs still satisfy Equations 9.2a and 9.2b:

$$MR_i \equiv p_i + Q_i p_i' = C'(Q_1 + Q_2) \equiv MC \quad \text{for} \quad i = 1, 2.$$

That is, the percentage markup of each Group *i*'s price over its marginal cost, $[p_i - m]/p_i$, is inversely proportional to its elasticity of demand. The higher the group's elasticity of demand, the lower the price and the closer the price is to marginal cost. As a result, the group whose demand is relatively sensitive to price is charged a lower price. Equations 9.2a and 9.2b can be combined to show that the price ratio to the two groups depends on their relative elasticities:

$$\frac{p_1}{p_2} = \frac{1 + 1/\epsilon_2}{1 + 1/\epsilon_1}. \tag{9.4}$$

For example, if Group 1 has a nearly perfectly elastic demand ($\epsilon_1 \approx -\infty$) and Group 2 has a demand elasticity of -2, then $p_1/p_2 = 1/2$. Group 2, the group with the relatively inelastic demand, is charged twice as much as Group 1. Alternatively stated, a profit-maximizing discriminating monopoly provides a discount to the group that has the higher elasticity of demand. See Examples 9.1, 9.3, and 9.4.

Still another method of price discrimination is to make the price that the consumer pays depend on whether the consumer turns in a previous version of the product, as well as the identity of the manufacturer of that previous version (Fudenberg and Tirole 1998). For example, the price of the latest version of Microsoft's word-processing software, Word, may depend on whether the consumer's previous word-processing software was an earlier version of Word.

Other Methods of Third-Degree Price Discrimination

Firms can practice third-degree price discrimination in other, subtle ways. For example, in many markets some consumers are better informed than others about prices. One way a firm can charge different prices to consumers is to set a high list price (the price at which an item is marked or listed to sell). The firm charges the list price unless a customer complains that it exceeds the price of the product at other stores. In the event of a complaint, the store matches the lower price. This method of pricing causes uninformed consumers to pay higher prices than knowledgeable ones.[9]

Another example of third-degree price discrimination involves exploiting differences in the value customers place on time. High-wage, high-income people typically value their time more than low-wage, low-income people (and may have a more inelastic demand for certain goods). One clever way to price discriminate between these two groups is to make a special offer that requires consumers to spend time to take advantage of the offer. For example, suppose a store is willing to sell an item over the telephone at the regular price and mail the item to the consumer. The store runs a sale, but only gives the low price to consumers who take the time to come in and pick the item up at the store. This is an effective method of price discrimination in which consumers who place a high value on time receive the item by mail and pay the regular price, and consumers who place a low value on time pick the item up at the store and pay the low price.

[9] The moral of this story is don't be afraid to complain about high prices. Department stores often have a policy that they will not be underpriced by their rivals. Chapter 13 analyzes how a firm's behavior is affected by the presence of both informed and uninformed consumers.

A related method of price discrimination is to exploit differences in the willingness of consumers to wait to consume a new product. For example, some people insist on being among the first to see a new movie or own the latest electronic gadget. Early purchasers pay more than later purchasers if prices fall over time. Not all firms with market power can profitably price discriminate over time, however. If consumers know that prices will fall in the future, some postpone buying. Price discrimination over time will be profitable provided the number who are willing to wait for lower prices is not too large (Stokey 1979—see **www.aw-bc.com/carlton_perloff** "Discrimination Over Time").

Welfare Effects of Price Discrimination

There is no ambiguity about the welfare effects of perfect price discrimination. Output is at the efficient, competitive level, but consumers are poorer than they are under competition; therefore, perfect price discrimination does not distort efficiency but does affect the distribution of income.

The welfare effects of third-degree discrimination are more difficult to analyze. We do know that, as with first-degree discrimination, consumers wind up with less surplus than under competition. Moreover, from Equations 9.3a and 9.3b, we know that third-degree price discrimination prices exceed marginal costs, so they are not as efficient as perfect competition or perfect price discrimination.

Third-degree price discrimination, however, may be better or worse than nondiscriminating monopoly pricing from an efficiency viewpoint, depending on the shapes of the demand and cost curves. The closer imperfect price discrimination is to perfect price discrimination, the more likely it is that the price discrimination leads to a more efficient outcome than nondiscriminating monopoly pricing.

Three sources of inefficiency are present in third-degree discrimination. The first is the usual one associated with monopoly: Price exceeds marginal cost, which results in an output restriction and hence an output inefficiency.

The second is a consumption inefficiency. Because different consumers pay different per-unit prices for a product, each consumer's marginal willingness to pay is not the same, which results in an inefficiency because of unexploited opportunities for further trade. For example, suppose that resale is impossible and there are two consumers. Larry is willing to pay $10 to consume the first unit and $9 to consume the second unit for a total of $19 to consume 2 units. If Larry is charged $10 per unit, he consumes only 1 unit. Andrew is willing to pay $7 to consume the first unit and $4 to consume the second unit for a total of $11 to consume 2 units. If Andrew is charged $5 per unit, he consumes only 1 unit. At the margin, Larry values the product more than Andrew. Larry values an additional unit at $9, and Andrew values the unit that he is consuming at $7. In such a case it is more efficient for Larry to consume 2 units and Andrew none. For example, if Larry paid Andrew $8 for his unit, both Larry and Andrew would be better off. Because resale is impossible, this trade does not occur, so there is inefficiency in consumption due to the price discrimination. Thus, if the discriminating monopoly produces the same (or less) output as the nondiscriminating monopoly, welfare is lower because there is no consumption inefficiency with the nondiscriminating monopoly since the monopoly charges all consumers the same price.[10]

A third source of inefficiency is that consumers may have to expend resources that do not benefit the firm to obtain a low price. For example, the consumer may have to wait in line or travel to a distant location to obtain the low price. One way to view this means of discriminating between groups of consumers is that the monopoly forces the consumer to buy a *bad* (such as the time waiting in line) in order to buy the good at a low price (Chiang and Spatt 1982).

Welfare may be higher with third-degree price discrimination than with a nondiscriminating monopoly if output is higher with discrimination. For example, suppose there are two groups of consumers and a nondiscriminating monopoly finds it optimal to set a price so high that one group buys no units. Then, because a discriminating monopoly serves both groups, output expands and consumers benefit in aggregate. In general, however, which type of monopoly leads to greater welfare is theoretically ambiguous and is an empirical question.[11] Competition can sometimes have unexpected price effects in cases where consumers differ. See Example 9.6.

The antitrust laws (Chapter 19) ban certain types of price discrimination. Apparently, it is not an antitrust violation to price discriminate among final consumers, but it is a violation to price discriminate among firms so as to affect their "competition" under the Robinson-Patman Act. Moreover, tie-in sales, which are closely related to a form of price discrimination, are illegal under certain circumstances. Given the ambiguous welfare effects of certain types of price discrimination, some economists question the desirability of a flat antitrust prohibition against these forms of price discrimination.

[10] Suppose Group 1 has an aggregate demand curve $Q_1 = a_1 - b_1 p_1$ and that Group 2 has an aggregate demand curve $Q_2 = a_2 - b_2 p_2$, where a_i and b_i (i = 1, 2) are numbers, and marginal cost is a constant, m. A discriminating monopoly chooses the profit-maximizing outputs Q_1* and Q_2*, so that

$$Q_i^* = \frac{a_i}{2} - \frac{b_i m}{2}, \quad i = 1,2.$$

A nondiscriminating monopoly picks a single price, p, and faces the demand curve for total quantity demanded, Q, given by $Q = (Q_1 + Q_2) = (a_1 + a_2) - (b_1 + b_2)p$ in the relevant range. If it is optimal to sell to both groups, the nondiscriminating monopoly chooses the profit-maximizing quantity

$$Q^* = \frac{a_1 + a_2}{2} - \frac{b_1 + b_2}{2} m,$$

so that $Q^* = Q_1^* + Q_2^*$. Hence, it follows that, in this case, output is the same with third-degree price discrimination and with simple monopoly. Welfare is lower with third-degree price discrimination than with a nondiscriminating monopoly because discriminating monopoly has a consumption inefficiency that simple monopoly does not.

[11] Schmalensee (1981b), Varian (1985), Katz (1987), and Ireland (1992) show that under special circumstances it is possible to make unambiguous welfare and output comparisons between simple monopoly and price discrimination, as we did in the previous footnote. Gale and Holmes (1993) examine welfare for airlines that use peak and off-peak pricing.

EXAMPLE 9.6 *Does Competition Always Lower Price?*

Grabowski and Vernon (1992) analyzed how a firm responds in its pricing to the retail sector when one of its brand name drugs comes off patent and faces competition from chemically equivalent drugs (generics). Before the patent expires, the firm takes advantage of its market power and sets a price above its marginal cost. After the drug comes off patent, the price will fall if all consumers regard generics as equivalent to the brand name drug. However, in their study of 18 major drugs whose patents had expired, Vernon and Grabowski noticed that, within two years of patent expiration, the market share of the firm with the previously patented product fell to about 50 percent, yet the price of the brand name drug *rose* by about 10 percent.

According to Grabowski and Vernon, there are two types of consumers, brand loyal and price sensitive. The brand-loyal consumers don't want to risk switching to another drug (even though it is chemically equivalent), while the price-sensitive consumers will switch if the generic price is lower than the price for the brand name drug. Prior to competition from generics, the firm sets one price to attract both the brand-loyal and price-sensitive consumers. When generics come into the market, the firm chooses not to meet the generic price (which is typically only 40 to 60 percent of the price for the brand name drug) and thereby forgoes sales to the price-sensitive segment. Once the firm does not have to attract the price-sensitive segment, it raises price to the remaining segment, the brand-loyal customers.

SUMMARY

Price discrimination occurs when a firm with market power uses nonuniform pricing to maximize its profits. Not all nonuniform pricing is due to price discrimination; some is due to cost differences.

For price discrimination to succeed, a firm must have some market power, know or be able to infer consumers' willingness to pay, and be able to prevent or control resales. In order to practice the types of price discrimination described in this chapter, firms must know quite a bit about individual consumers. For perfect, or first-degree, price discrimination, a firm must know each consumer's demand curve. For third-degree price discrimination, a firm must be able to identify the group that a consumer belongs to and must know the group's demand curve. Sometimes a firm does not have enough information to practice either first-degree or third-degree price discrimination. It must then use more complicated pricing methods to maximize its profits. These other pricing methods are examined in the next chapter.

Perfect price discrimination is efficient: The same quantity is sold as would be sold by a competitive industry. It leads to a redistribution of income where the monopoly obtains all the potential consumer surplus. Third-degree price discrimination is not as efficient as competition or perfect price discrimination. Compared to a nondiscriminating monopoly, it may be more or less efficient and have higher or lower welfare.

PROBLEMS

1. Disneyland Paris price discriminates by charging lower entry fees for children than adults. Why does it not have a resale problem?
2. A (natural) monopoly has a flat marginal cost curve. Its average cost curve is downward sloping (because it has a fixed cost). The firm can price discriminate perfectly.
   a. In a graph, show how much the monopoly produces, Q^*. Will it produce to where price equals its marginal cost?
   b. Show graphically (and explain) what its profit is.
3. Many retail stores provide discounts for regular customers, if those customers sign up for a rewards card. Why do such stores not worry that regular customers will resell these goods?
4. Suppose there are two groups of consumers and that it is optimal for a nondiscriminating monopoly to set $p = \$10$. At that price, no one from the first group chooses to purchase. Now, suppose the monopoly can price discriminate. Will total output expand? Why or why not?
5. Suppose a consumer wants just one unit of a good and is willing to pay at most \$10. Draw the demand curve and calculate the maximum consumer surplus that can be extracted. Suppose that there is a second consumer who also demands just one unit and is willing to pay at most \$9. A perfectly discriminating monopoly charges the first consumer \$10 and the second consumer \$9. Why is there no consumption inefficiency as occurs in third-degree price discrimination?
6. Would a price-discriminating monopoly ever produce less than a nondiscriminating monopoly?

Answers to odd-numbered problems are given at the back of the book.

SUGGESTED READINGS

Stole (forthcoming), Wilson (1993), and Varian (1989) provide an excellent survey of price discrimination. Borenstein (1985), Holmes (1989), Katz (1984), and Lederer and Hurter (1986) extend the analysis of third-degree price discrimination to market structures other than pure monopoly (such as monopolistic competition). Phlips (1983) covers a number of extensions to the basic theory and has some empirical applications.

APPENDIX 9A

An Example of Price Discrimination: Agricultural Marketing Orders

Federal and state governments mandate price discrimination in what would otherwise be competitive agricultural markets through the use of *marketing orders*. We first discuss how marketing orders permit farmers to price discriminate and then discuss the efficiency and welfare effects of these programs.[1]

Marketing Order Rules

Many marketing orders require farmers to participate in a *classified pricing* scheme, in which consumers in different markets are charged different prices. Typically, commodities are sold in at least two markets. In most marketing orders, the primary market is the fresh food (or domestic) market, in which the demand elasticity is relatively low and hence price is relatively high. The secondary market is the processed food (or export) market, in which the demand elasticity is relatively high and hence price is relatively low. Because processed foods cannot be converted back into fresh foods and it is costly to reimport exports, resales between the markets do not occur, so price discrimination is possible. How these market division schemes work varies across marketing orders.

One common scheme is a quantity restriction that dictates the share of a farmer's output that can be sold in the primary market. These *quantity share restrictions* increase prices in the primary market and lower them in the secondary market, where the extra output is sold. Examples include grade A milk, California almonds, Oregon-Washington filberts, Pacific Coast walnuts, California dates, and California raisins (Jesse and Johnson 1981). States that permitted quantity share restrictions in their marketing orders include California, Colorado, Georgia, South Carolina, and Utah (Garoyan and Youde 1975), though many of these programs have been dropped in recent years.

If there are no output restrictions, when classified pricing schemes are first introduced, they cause farms' profits to rise, which eventually induces entry and additional output. Output expands until the marginal farmers earn zero profits despite the price discrimination. Thus, in the absence of output restrictions, marketing orders produce a different equilibrium than would a price-discriminating monopoly or a competitive industry.

[1] We concentrate on the price discrimination aspects of marketing orders. Allegedly, Congress adopted marketing orders to help farmers act collectively to offset the monopsony power of milk processors (Novakovic and Boynton 1984). Modern defenders of marketing orders claim that they are "necessary" to stabilize prices or quantities, a view disputed by many agricultural economists (Jesse and Johnson 1981, Gardner 1984).

Efficiency and Welfare Effects of Marketing Orders

There are both gainers and losers under a price classification scheme, as is shown by a simplified model with fixed supply.[2] The marketing order allocates part of the total output to each of two markets: Class 1 (fresh) and Class 2 (processed), as shown in Figure 9A.1. By restricting output in the Class 1 market to Q_1, which is less than the competitive level, Q_1^c, the marketing order drives the price in the Class 1 market, p_1, above the competitive price, p_c. The excess output is sold in the Class 2 market, so Q_2 is greater than the competitive output Q_2^c, and the price in the Class 2 market, p_2, is below the competitive price, p_c.

Because the price in the fresh market is above the competitive price, $p_1 > p_c$, consumers of the fresh product lose consumer surplus equal to areas $A + B$. Farmers' profits on the Q_1 units they sell increase by area A ($= [p_1 - p_c]Q_1$), so the net loss (the loss to consumers not offset by a gain to producers) in the fresh market is B.

Consumers of the processed product gain consumer surplus equal to area C due to the lower price, $p_2 < p_c$. Farmers' profits are lower on the Q_2 units of output they sell in the Class 2 market than if a competitive price were charged by areas $C + D$. The net loss in the processed market is D.

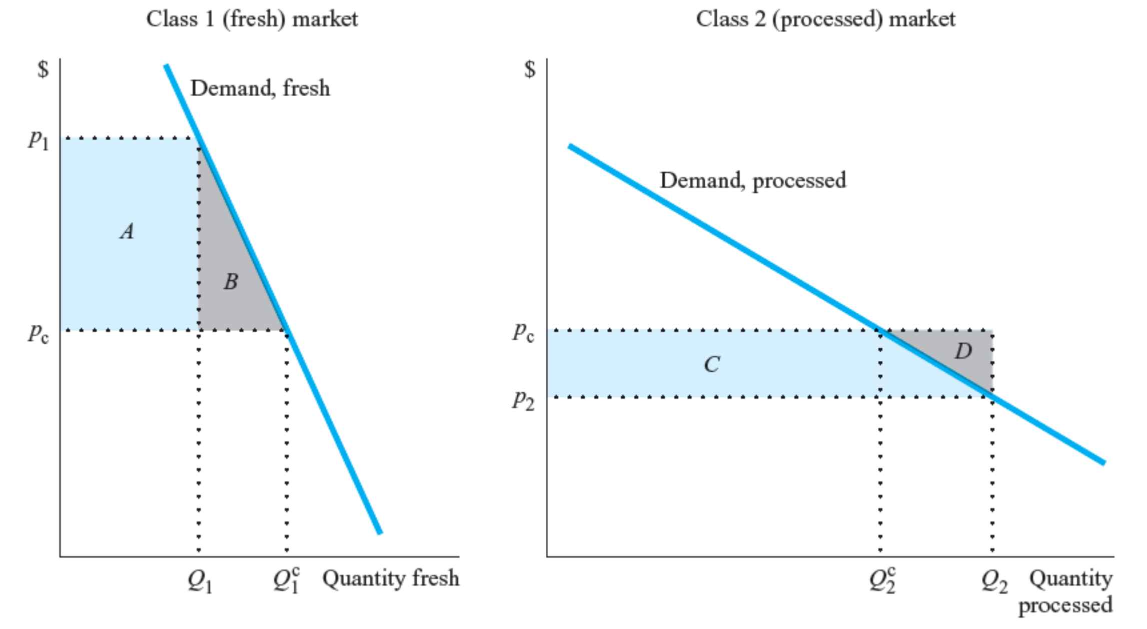
Consumers lose areas $A + B$ in the fresh market and gain area C in the processed market, for a net total loss of $A + B - C$. Farmers' profits increase. Farmers receive a *blend* (average) price $p_b = (p_1Q_1 + p_2Q_2)/(Q_1 + Q_2)$. The blend price is higher than the competitive price, or else there would be no point to engaging in such price discrimination. Farmers gain profits of A in the fresh market and lose profits of $C + D$ in the processed market, so their total net gains are areas $A - (C + D)$. Consumers lose more ($A + B - C$) than producers gain ($A - C - D$), for a net total loss of areas $B + D$. Thus, welfare is lower under a classified pricing scheme than under competition.

For simplicity, we assumed that supply was fixed and the only effect of the market allocation program was to redirect the product from the fresh to the processed market. More generally, where supply is not fixed, marketing orders, by increasing the effective price farmers receive (the weighted average of the fresh and processed prices), increase the amount supplied. Much of the extra supply is directed to the secondary market (to keep the price high in the primary market). Because the price in the secondary market is less than the competitive price (and hence marginal cost), the cost of this extra output exceeds its value to consumers. Thus, social loss can be even greater when supply curves are not vertical.

The social loss from most classified pricing programs is relatively small because the industries are small. There are, however, large social losses in dairy markets. Based on data from the early 1970s, Ippolito and Masson (1978) estimate that the effect of regulation was to raise the price of fresh (Class 1) milk 9.3 percent (at the farm level), to

[2] Ippolito and Masson (1978) and Berck and Perloff (1985) discuss static models. Berck and Perloff (1985) also show how the analysis changes in a dynamic model, where entry into the industry is slow. Cave and Salant (1987) model the voting behavior in agricultural marketing boards, which determine how marketing orders are run.

FIGURE 9A.1 Price Discrimination in Agricultural Marketing Orders



decrease the price of processed milk (used in products) (Class 2) by approximately 5.6 percent, and to increase the blend price facing regulated farmers by 3.7 percent. As a result, Class 1 consumption was 1.9 percent lower than it would have been without regulation, and Class 2 consumption was 9.6 percent higher. They calculate that the classified price regulation was equivalent to a tax on consumers of fresh milk of $333.8 million per year; and Class 2 users received a subsidy of $120.9 million. The producer profits on fresh milk increased by $210.6 million, and producer profits on processed milk fell by $105.2 million per year. Including the administration costs of running the program and the induced inefficiencies in transportation, they calculate that the total social cost was approximately $60 million per year.

Most other researchers, using other approaches, estimate larger costs. Kwoka (1977) estimates that the 1970 classified pricing and pooling schemes had an annual efficiency loss of $179 million. Heien (1977) calculates a total social cost of $175 million. LaFrance and de Gorter (1985) point out that these estimates are based on static analyses that ignore the time that it takes to build up a dairy herd. Using a dynamic model, they estimate that the social cost of the program is three times larger than the static estimates.

CHAPTER 10

Advanced Topics in Pricing

> *A fool and his money are soon parted. What I want to know is how they got together in the first place.* —*Cyril Fletcher*

This chapter analyzes more complicated methods of price discrimination than were discussed in Chapter 9, such as nonlinear pricing, two-part tariffs, quantity discounts, tie-in sales, and quality choice. It shows how these common pricing methods increase profits of firms with market power that can control or prevent resale. The pricing methods in this chapter do not require that the firm have as much knowledge about consumers as was required to practice the methods described in Chapter 9.

The key points of this chapter are

1. Nonlinear pricing schemes can be used by a firm to price discriminate when the firm does not know the demands of individual consumers.
2. A firm can induce consumers to reveal which group they belong to through their choice from among several two-part tariffs.
3. The effectiveness of tie-in sales (where multiple goods must be purchased at the same time) in increasing profits depends on whether demands for these goods are related.
4. A firm with market power may use a variety of other policies to increase its profit above what it would earn if it charged everyone the same price.

Nonlinear Pricing

Nonlinear pricing occurs when a consumer's total expenditure on an item does not rise linearly (proportionately) with the amount purchased. That is, the price per unit varies with the number of units the customer buys. Methods of nonlinear pricing are used to practice *second-degree price discrimination*, where a firm

EXAMPLE 10.1

Football Tariffs

When the Raiders moved back to Oakland in 1995, they changed how they sold tickets to their football games. Under the new system, a fan paid a fee of between $250 and $4,000 for a personal seat license (PSL), which gave the fan the right to buy season tickets for the next 11 years at a ticket price per game ranging between $40 and $60. The Carolina Panthers introduced the PSL in 1993, and at least 11 NFL teams used a PSL by 2002. By one estimate, more than $700 million has been raised by the PSL portion of this two-part tariff.

that can prevent or at least control resale between individuals charges different consumers different prices, but the firm does not know the demands of each individual.[1] Rather, the firm makes use of its knowledge about the underlying distribution of demand in the population. This section first presents a simple type of nonlinear pricing schedule—a single two-part tariff—and then discusses the more general problem of nonlinear pricing, illustrating it with an example of a multiple two-part tariff.

A Single Two-Part Tariff

A firm that uses a *two-part tariff* charges consumers a lump-sum fee for the right to purchase goods and a usage charge per unit (see Chapter 9; also Oi 1971 and Schmalensee 1981a). For example, tennis clubs commonly charge a membership fee plus a usage fee that depends on how many hours one plays tennis. Many firms that rent copy machines pay a minimum rental fee plus a fee that depends on their usage of the machine. As a final example, suppose that a firm sells cameras whose use requires a special type of film (for example, Polaroid's instant-picture cameras). One can think of the purchase of the camera as the payment of a lump-sum fee and the film purchases as the payment of a usage-sensitive fee. Also see Example 10.1.

When a two-part tariff is used, a firm must somehow prevent resale. Otherwise, it would make sense for one customer to pay one fixed fee and purchase all the goods, and then resell them to everyone else, so that only one fixed fee is collected. For example, suppose that a firm requires $100 from each buyer plus a per-unit charge of $1. If Lisa and Daniel each buy 50 units, they each pay $150, for a total expenditure of $300. If, however, Lisa buys for both, the total expenditure is $200. Thus, Lisa and Daniel could each pay $100 and be better off. To prevent this, the firm could try to prevent Lisa from reselling to Daniel by making it costly to divide shipped orders. The remainder of this section assumes that firms use one of the methods discussed in the

[1] As Chapter 9 shows, nonlinear pricing can also be used to practice first-degree price discrimination when a firm knows each consumer's demand curve. A monopoly that uses second-degree discrimination cannot extract all consumer surplus, as it could if it were able to implement first-degree discrimination.

previous chapter to prevent resale. This section analyzes the case in which a firm can use only one two-part tariff; subsequent sections relax this restriction.

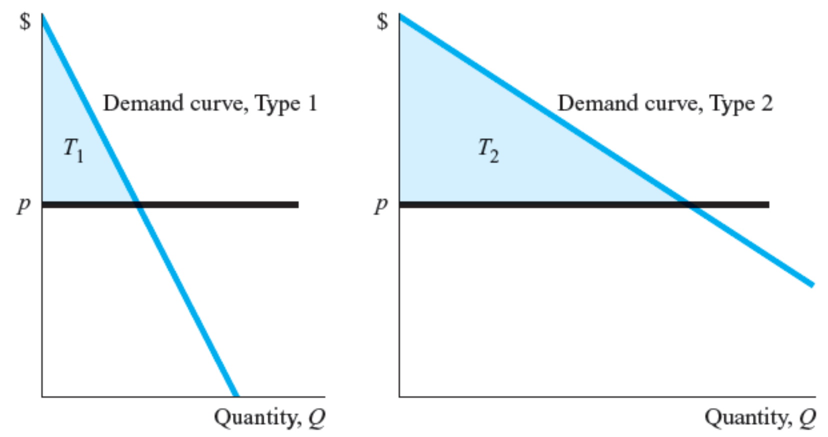
When consumers are identical, a two-part tariff can be used to extract all consumer surplus. This point was illustrated in Chapter 9 in the section "Each Consumer Buys More Than One Unit." Typically, however, there is more than one type of consumer, and the firm cannot distinguish among consumers. We assume that the firm knows that demands differ within the population but lacks specific knowledge of each individual consumer's demand. For example, from marketing surveys, the firm may be aware that 50 percent of its customers value its services greatly, whereas another 50 percent could easily switch to another product. Even though the firm knows the general distribution of demand, it may be unable to determine the group to which a particular customer belongs.

Suppose that there are only two types of consumers, and they have the demand curves of Figure 10.1. A Type 2 customer is willing to buy more at price p than a Type 1 customer and enjoys more consumer surplus ($T_2 > T_1$) than does a Type 1 customer. If a firm charging a price p per unit could identify each customer's type, it could also charge a Type 1 customer a fee of T_1 and a Type 2 customer a fee of T_2.

Suppose the firm must choose a single two-part tariff. It chooses a lump-sum fee, T, and a per-unit charge, p, so as to maximize profits. If p exceeds average variable costs, the firm earns positive net revenues from additional sales. If the firm is unable to distinguish consumer types and charges a single two-part tariff involving a per-unit charge of p, the lump-sum fee it charges cannot exceed T_1 if Type 1 consumers are to participate. For example, if the firm charges a lump-sum fee of T_2, Type 1 consumers refuse to purchase the product.

The firm faces a dilemma. If it charges a low price, it sells more of its product and can charge a higher lump-sum fee (as can be seen from Figure 10.1). On the other

FIGURE 10.1 Two Different Demand Curves



hand, its ability to charge a high lump-sum fee to extract the consumer surplus of Type 2 consumers is constrained by the Type 1 consumers' low willingness to pay. In many cases, the firm may make higher profits by concentrating on Type 2 consumers, letting Type 1 consumers choose not to purchase the product. The less similar Type 1 consumers are to Type 2 consumers, the more difficult it is for the firm to extract consumer surplus from Type 2 consumers with a single two-part tariff (see Appendices 10A and 10B).

The optimal two-part tariff typically generates more profits than a single price, because a single price is a special type of a two-part tariff: a two-part tariff with a zero lump-sum fee. The optimal two-part tariff generates less profit than perfect (first-degree) price discrimination, but may or may not generate less profit than third-degree price discrimination (where the firm charges a different price to each consumer group). However, unlike with third-degree price discrimination, a firm need not be able to identify which type a consumer is to use a two-part tariff.

One can think of a two-part tariff as consisting of a fixed charge for one product and a marginal charge for another. For example, the fixed charge could be the price of a camera, and the marginal charge that depends on usage could be the price of the film. Appendix 10A (and **www.aw-bc.com/carlton_perloff** "Two-Part Tariff for Two Products") shows that the usage-sensitive price (for example, the price of film) tends to exceed its marginal cost, but the fixed charge may well be below the marginal cost of the item (for example, the camera). In general, the fixed charge increases as the difference between the average quantity purchased and the quantity purchased by the marginal customer decreases, and as the elasticity of demand increases. The usage-sensitive price increases as the elasticity of demand decreases and as the difference between the quantities purchased by the average customer and the marginal customer increases.

Two Two-Part Tariffs

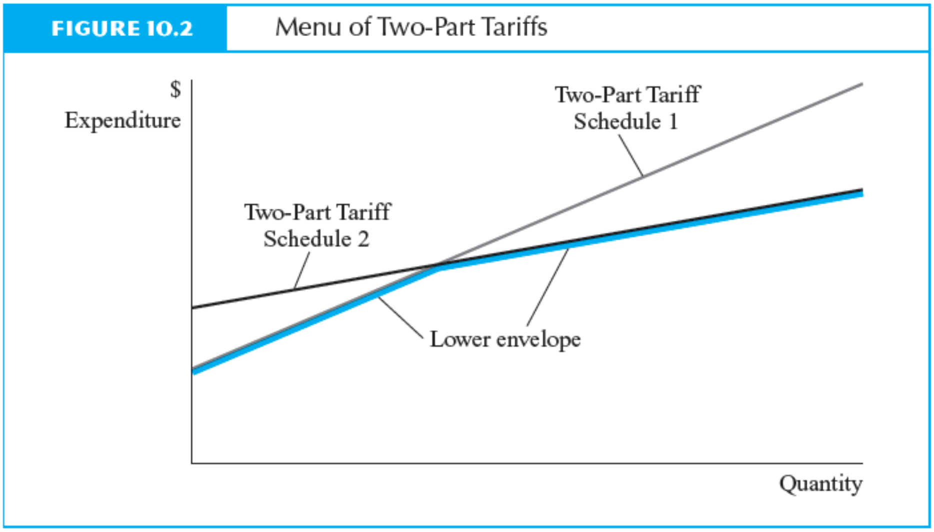
The two-part tariff just described is one of the simplest examples of a pricing structure in which the average price varies with output—which is a characteristic of any nonlinear pricing scheme. In general, the amount paid can vary with the amount purchased in any prespecified way: The price paid is a function of quantity, and the firm is allowed to choose any function it desires.

Finding the general nonlinear pricing policy that maximizes a monopoly's profits is complicated.[2] This section presents a simplified example to illustrate the key ideas.

Suppose a firm knows the demand curves of two types of consumers (Type 1 and Type 2) and the prevalence of different types of consumers in the population, but it does not know the type of any individual consumer. The firm can offer consumers a choice of two different two-part tariff schedules. Each consumer chooses or *self-selects* that schedule that corresponds to a higher level of utility. The two schedules are shown in Figure 10.2 as straight, black lines. The intercepts on the vertical axis are the fixed,

[2] See Katz (1983), Spence (1977b), Tirole (1988, Ch. 3), Wilson (1993), and Appendix 10B for more details.

FIGURE 10.2 Menu of Two-Part Tariffs



lump-sum fees, and the slopes of the curves are the constant marginal costs. From inspection, the consumer can purchase a small number of units for less money by choosing Two-Part Tariff Schedule 1 and a large number of units for less money by choosing Two-Part Tariff Schedule 2. Following this reasoning, consumers choose the lower "envelope" of the two curves, which is shown by a thick blue kinked line.

The firm chooses its two two-part tariff schedules to maximize its profits. The inability of the firm to identify the willingness of any individual customers to pay constrains its pricing policy. The firm provides a choice of two two-part tariffs in order to separate consumers into groups, so it can lower the price to one group without having to pass along the same low price to the other group. This motive is the same as in any price discrimination scheme (Chapter 9).

If the firm knew which consumers belonged to each group (and could prevent resale), the firm could design a two-part tariff for each group. From our earlier discussion of perfect price discrimination in Chapter 9, the optimal policy when the firm is knowledgeable is for the firm to charge each consumer a price equal to its marginal cost, *m,* and to extract the consumer surplus of each customer by charging a lump-sum fee. Thus, if in Figure 10.1, $p = m$, the firm charges a Type 1 consumer T_1 and a Type 2 consumer T_2.

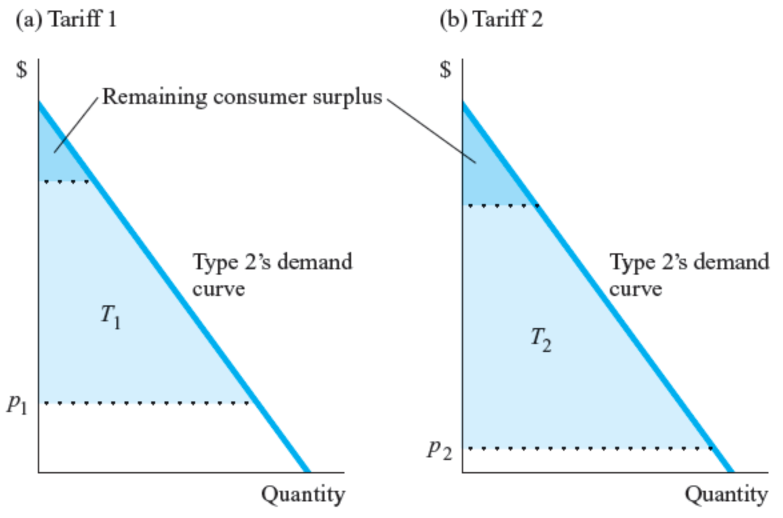
Suppose the firm simply announced that it had two two-part tariffs: one of (T_1, m) and the other of (T_2, m), where the first number (T_1) in parentheses is the fixed fee, and the second (m) is the marginal price. If the two types of consumers have demand curves as shown in Figure 10.1, no consumer would ever choose the second two-part tariff, because T_2 exceeds T_1. That is, all consumers would choose the first two-part tariff. Because consumers always choose (self-select) the pricing structure that is best for them, the firm's ability to price discriminate is constrained. In this example, consumer unwillingness to choose higher price schedules rules out the possibility of perfect price discrimination. The firm designs its pricing structure to maximize profits subject to a **self-selection constraint**: a restriction on a firm's pricing structure such that consumers in any group do not prefer another group's two-part tariff schedule. We focus on an optimal solution in which the monopoly serves both types of consumers.

For example, suppose that Type 2 consumers demand more units than Type 1 consumers at every price, as shown in Figure 10.1. Then the firm's optimal policy is for the Type 2 consumers' fixed fee, T_2, to exceed the fee for Type 1 consumers, T_1; the marginal price facing Type 2 consumers, p_2, to be below that for Type 1 consumers, p_1, and to equal marginal cost (see Appendix 10B). By offering a low price to the large demanders, customers derive a large consumer surplus, which the firm captures through T_2. The high T_2 discourages the small-volume buyers (Type 1), who prefer to pay a higher marginal price on the smaller amounts they purchase. In other words, the high-volume purchasers (Type 2) value low prices much more than the low-volume purchasers (Type 1), which enables the firm to separate the two groups.

Restaurants provide one extreme example of the use of two two-part tariffs to separate consumers into groups. Many restaurants offer consumers choices of all-you-can-eat buffets (high T, $p = 0$) and an á la carte menu ($T = 0$, high p). Large eaters choose the buffet.

Figure 10.3 illustrates this example for Type 2 consumers. Even though the fixed fee T_2 is greater than T_1, a Type 2 consumer prefers (T_2, p_2) to (T_1, p_1) because the price is lower ($p_2 < p_1$) so that the remaining consumer surplus is higher under the second tariff. Similarly, Type 1 consumers prefer (T_1, p_1) to (T_2, p_2); their remaining consumer surplus is higher under the tariff (T_1, p_1) because they can take advantage of the low fixed fee.

FIGURE 10.3 How a Type 2 Consumer Fares Under the Two-Part Tariffs



Type 2 consumers benefit from the presence of the Type 1 consumers. Without the Type 1 consumers, Type 2 consumers would receive zero utility. Consumer diversity helps consumers who have large demands. (See Appendix 10B.)

Tie-in Sales

A *tie-in sale* is one in which a consumer can buy one good only by purchasing another good as well. For example, if a supermarket sells you a pound of coffee on the condition that you also buy sugar, that would be a tie-in sale. Due to litigation growing out of legal restrictions governing the use of tie-in sales (see Chapter 19), there are many, well-documented examples of firms using tie-in sales.

Tie-in sales may be used to price discriminate. We use the term *price discrimination* quite broadly, in the sense that a tie-in sale enables a monopoly to increase its profit over and above what it would earn if the two goods were offered for sale individually at constant prices. As with all price discrimination schemes, the reason the tie-in can increase profits is that it enables a firm to charge more to consumers who value the good the most. Although tie-in sales can be used to price discriminate, it is important to recognize that there are many other reasons for tie-in sales that are unrelated to price discrimination.

General Justifications for Tie-in Sales

Tie-in sales may be used to increase efficiency, avoid price regulations, give secret price discounts, and to assure quality. After examining these motives in this section, we examine how tie-in sales can be used to price discriminate.[3]

Efficiency. Tie-in sales may be used to increase efficiency. For example, laced shoes are typically sold with laces; everyone who buys laced shoes needs shoe laces. As long as people's tastes for shoe laces do not differ dramatically, it is more efficient (that is, it lowers transaction costs) to sell laced shoes with standard shoe laces than to sell the shoe laces and shoes separately. In the extreme, every product can be thought of as composed of multiple products. For example, a radio consists of many individual components. The same is true of an automobile, which could be regarded as a package including an engine, tires, and a car body. Obviously, each of these products could be sold separately, but because consumers desire assembled products, they come tied together. See Example 10.2.

Another efficiency justification for tie-in sales is that they economize on the cost of grading individual units of a product. Instead of grading each unit separately, total search cost can be reduced if the buyer must purchase several items together. See Example 10.3.

Evade Regulations. Another common reason for tie-in sales is to evade price controls. Imagine that the government sets price controls on steel. Suppose the controlled

[3] In the next chapter, we analyze how tie-in sales can be used strategically to harm rivals.

EXAMPLE 10.2

You Auto Save from Tie-in Sales

Instead of purchasing an automobile already assembled—a tie-in sale—you could buy all the parts separately and assemble and paint the car yourself. Doing so would cost you substantially more, however. According to the *Journal of American Insurance*, the cost of buying an assembled 1988 Buick Skylark from a dealer was $12,568. The cost of purchasing the replacement parts for the car and the paint service was $40,280.

Source: Brendan Boyd, "By the Numbers," *San Francisco Chronicle*, March 3, 1989:B5.

price is below the market-clearing price (the price at which supply equals demand) by $5.00. One method of circumventing price controls is to sell steel at the controlled price but only on the condition that the consumer also pay $5.25 for a pencil that costs 25¢ to produce. In this way, the market-clearing price for steel is maintained, and the price controls are still met.

A variant of the preceding example is to use a tie-in sale to circumvent regulation. Some public utilities, such as electric utilities, are subject to rate regulation. If electric utilities were allowed to sell light bulbs and were also allowed to force consumers to buy light bulbs as a condition of receiving electric service, the electric utility could completely circumvent the rate regulation by charging a high price for light bulbs, unless the regulators also regulated the price for light bulbs.

Secret Price Discounts. Another motivation for tie-ins is to give a secret price discount. For example, a member of an oligopoly may want to give price discounts without rivals knowing about them (see Chapter 5). It may be able to keep its discounts se-

EXAMPLE 10.3

Stuck Holding the Bag

A large fraction of the world's diamonds is marketed by deBeers Consolidated Mines. A buyer is allowed to specify the average quality of diamond. The buyer is then given a bag containing several diamonds. The buyer has the right to reject or accept the bag in its entirety. A buyer who rejects a bag is not invited back. One rationale for this marketing procedure is that if buyers were allowed to investigate each stone in detail and reject any one stone, deBeers would have to spend resources to sort and grade the diamonds more carefully. The cost is avoided by the "take the bag or leave it" selling policy.

Source: Kenney and Klein (1983).

cret by selling a product at the oligopoly price but tying that sale to another product with a very low price. For example, a firm can give a 10 percent discount on a $100 price by giving as a gift to purchasers of the product another product that is worth $10. Alternatively, the firm could charge customers $10 less than they would have to pay if they purchased the tied product in the competitive market.

Assure Quality. Tie-in sales can assure quality. For example, Kodak claimed that it tied the development of its film to its film sales because it did not believe that independent developers could develop Kodak film as skillfully as could Kodak.[4] Kodak could have reasoned that if an independent developer made a mistake and produced poor pictures, the consumer would be unable to distinguish whether the film was bad or the developer was bad. The consumer might then be reluctant to purchase film from Kodak in the future.

Generally, a firm may assure quality by forcing customers to buy another of its products or services or to not use substitutes. When Searle introduced NutraSweet, a nonsugar sweetener, it claimed that it was natural and better tasting than other less expensive sugar substitutes, such as saccharin. Beverage manufacturers began using a blend of saccharin and NutraSweet in their diet sodas. Searle felt that the taste of the blend was not as good as NutraSweet alone, and, fearing that NutraSweet would be improperly judged, required that users of NutraSweet not use it as a blend.[5] Of course, both Kodak and NutraSweet may have had other motivations for their actions.

Tie-in Sales as a Method of Price Discrimination

A final reason for tie-in sales—the focus of the rest of this chapter—is to increase monopoly profit. That is, if a firm has a monopoly in a product, it may be able to increase its profits by tying another good to the sale of the monopolized good. Tie-in sales can be used to price discriminate in a variety of circumstances, and the way they work is analyzed differently depending on the circumstances. Therefore, after reviewing some general reasons for using tie-in sales as a method of price discrimination, we discuss their use in a variety of different circumstances.

There are two common types of tie-in sales. One is **bundling** (Adams and Yellen 1976) or a **package tie-in sale** and occurs when two or more products are sold only in fixed proportions. For example, a store requires that, if you buy one jar of coffee, you must buy one bag of sugar. Everyone who buys these products consumes them in these fixed proportions, or else they must dispose of part of one or the other product.

The other common type is a **requirements tie-in sale**, where customers who purchase one product from a firm are required to make all their purchases of another product from that firm. For example, IBM used to require that purchasers of its machines that used tabulating cards buy all of their tabulating cards, no matter how

[4] Kodak film was once sold only with development included. Purchasers simply mailed the film to Kodak for developing at no additional charge.
[5] *Newsweek,* January 28, 1985:57.

many, from IBM.[6] In such a requirements tie, different consumers might consume different relative amounts of the two products. For example, in the IBM case, a large accounting firm might consume many more tabulating cards than a small manufacturing firm. In some cases a requirements tie automatically occurs when the related product is only produced by the firm selling the other product. For example, Polaroid is the only firm that can sell film to fit the cameras it manufactures.

As with all methods of nonlinear pricing in which different consumers pay different prices for the same product, a firm can use a tie-in sale to price discriminate only if trades between consumers are prevented. For example, in the case of the (fixed proportions) package tie-in, the tie-in fails as a method of price discrimination if customers can break apart the package and resell the various products on the open market. Similarly, in the case of the requirements tie-in, the consumer must be unable to purchase the tied good elsewhere at the competitive price.

We now examine each of the two types of tie-in sales and the circumstances under which they increase profits, beginning with an analysis of a package tie-in where product demands are independent. Products have independent demands if the value a consumer places on one product does not depend on the consumption of the other product. We then turn to an analysis of products whose demands are related.

Package Tie-in Sales of Independent Products

To examine package tie-in sales of independent products, we first suppose that a firm has a monopoly in both products. Then we examine a firm that has a monopoly in only one of the two products.

Package Tie-in with Both Products Monopolized. Suppose a firm has a monopoly in both Product *A* and Product *B*. For example, a movie company, which sells movies to theaters, has two movies, *A* and *B*, which are in great demand. Would this monopoly earn higher profits if it sold *A* and *B* separately or as a package? The answer depends on the value that various consumers place on each of the two movies separately versus the value they place on the package (Stigler 1968c).

The monopoly sells to two types of consumers. Type 1 consumers are willing to pay at most $9,000 to purchase *A* separately and $3,000 to purchase *B* separately (see Table 10.1). Type 2 consumers are willing to pay at most $10,000 for *A* separately and $2,000 for *B*. The amount each group is willing to pay for *A* is independent of whether *B* is also purchased, and vice versa.

Suppose the cost of producing the products is zero, and the monopoly wants to maximize the revenues that it receives from selling to these two types of consumers. The monopoly has two choices: It can sell *A* and *B* separately or as a package. If it sells Product *A* separately, it maximizes revenue by charging a price of $9,000. At that

[6] *IBM v. United States,* 298 U.S. 131 (1936). The machine performed numerical calculations mechanically based upon the holes in the cards that were placed into the machine. Customers who bought cards through IBM presumably paid a higher price than they would have if they bought from others.

TABLE 10.1 **Example of a Profitable Package Tie-in**

	Type 1 Consumers	Type 2 Consumers
Amount ($) willing to pay for *A*	9,000	10,000
Amount ($) willing to pay for *B*	3,000	2,000
Amount ($) willing to pay for *A* and *B* together	12,000	12,000

price, both types of consumers purchase Product *A*, and the monopoly receives $18,000. Similarly, in order to maximize the revenue from selling *B* separately, the monopoly sets a price of $2,000 and receives revenue of $4,000. Therefore, the total revenues received from selling *A* and *B* separately are $22,000.

Now, suppose that the monopoly decides to sell *A* and *B* as a package. Both Type 1 and Type 2 consumers are willing to pay $12,000 for the package. If the monopoly sells the package for $12,000, it sells to both consumers and receives $24,000. Therefore, in this example, profit (revenues) is maximized by tying the two goods together and selling them as a package instead of selling them separately. By selling the products as a package, the monopoly effectively is able to charge Type 1 consumers a higher price for *B* ($3,000) compared to that charged Type 2 consumers ($2,000) and a lower price for *A* ($9,000) compared to that charged Type 2 consumers ($10,000). In other words, when both Type 1 and Type 2 consumers buy the same package, they are placing different relative values on the components of the package. Thus, this tie-in sale is an example of price discrimination: The monopoly charges different prices to different customers for the same product.

By changing the numbers in Table 10.1, it is possible to show that a tie-in is not always the most profitable strategy. In Table 10.2, it is more profitable for the monopoly to sell the products separately than as a package. The maximum profit from selling the products separately is $20,000 ($18,000 for *A*, which is sold to both types of consumers, and $2,000 for *B*, which is sold to only Type 2 consumers), whereas the profit from selling *A* and *B* together (to both types of consumers) is only $19,000.

Let us return to the example in Table 10.1. Suppose someone purchases the packages of *A* and *B* together, breaks the package apart, and sells *A* and *B* separately in a resale market. The market-clearing price for *A* that induces the Type 1 and Type 2

TABLE 10.2 **Example of an Unprofitable Package Tie-in**

	Type 1 Consumers	Type 2 Consumers
Amount ($) willing to pay for *A*	9,000	10,000
Amount ($) willing to pay for *B*	500	2,000
Amount ($) willing to pay for *A* and *B* together	9,500	12,000

consumers to hold *A* is $9,000. Similarly, the market-clearing price for *B* is $2,000. Now if the Type 1 and Type 2 consumers realize that a resale market will develop after they purchase their packages and that the price of *A* will be $9,000 and the price of *B* will be $2,000, they will not be willing to purchase the package. Instead, they will wait until the resale market develops and purchase *A* and *B* separately for a combined price of $11,000 rather than for the $12,000 in the example. Thus, with a resale market, nobody purchases the package, and the attempt to practice price discrimination through a package tie-in does not work. That is, a resale market destroys the ability of a monopoly to charge effectively different prices for the same product and thereby destroys the ability to price discriminate through tie-in sales.

In the absence of a resale market, in the example in Table 10.1, a package tie-in resulted in a successful price discrimination because there is a negative relationship between what each type consumer is willing to pay for the two items. For example, Type 1 consumers put a relatively high value on *B* and place a relatively low value on *A* (see McAfee, McMillan, and Whinston 1989). As a result, there is relative heterogeneity between the two consumer types in their valuation of the individual products, but relative homogeneity in their valuation of the package.[7] See Example 10.4.

Mixed Bundling with Both Products Monopolized. Some firms give consumers a choice between buying a bundle and buying goods separately, which is called *mixed bundling* (Adams and Yellen 1976). You can buy a computer with bundled software or buy the hardware and software separately. Baseball teams sell season tickets and tickets to individual games.

If you ran a restaurant, you would want to know which of the following pricing methods would maximize your profit:

- *Individual pricing:* Customers order each item separately off an à la carte menu. A customer may order any appetizer, main dish, or dessert and may skip any course.
- *Pure bundling:* Customers can only purchase a fixed-price meal—a bundle—that includes all courses and offers limited choices (possibly only one) for each course.
- *Mixed bundling:* Customers may either order a fixed-price meal or order from an à la carte menu.

You'd first have to determine whether you can bundle at all. If you could bundle, then you'd have to determine which pricing method gives you the highest profit.

[7] Even when the valuations for *A* and *B* are not negatively correlated, bundling can be profitable. For example, suppose that there are two goods, *A* and *B*, each of which is valued by consumers at either $7 or $13. There are then four possible value combinations: (7, 7), (7, 13), (13, 7), (13, 13). Assuming that the products can be produced at no cost, if the goods are sold unbundled, the profit-maximizing price of each good is $7, and four units of each good are sold for a total profit (revenue) of $56. If the products are bundled, the profit-maximizing solution is to sell three bundles at $20 each so that the total profit is $60. See McAfee et al. (1989).

EXAMPLE 10.4 *Tied to TV*

In most areas of the United States, a monopoly provider of cable television offers consumers various service packages. Typically, the basic bundle of service includes local programming, TV network programming, and about seven additional cable network channels.

How should a cable provider decide which networks to include in its bundle? One answer is to choose them so as to be better able to price discriminate through bundling. Consumers differ in their valuations of individual networks. However, the cable provider may be able to bundle various networks so that consumers with varying tastes value the overall bundle relatively similarly. For example, a family's teenagers may watch MTV but rarely tune in to CNN News, while their parents may do the opposite. Thus, each household could value the bundle of MTV plus CNN News at the same price, even though they value each network differently. If so, the cable provider can charge a single price for the bundle that extracts most of the consumers' value, even though consumers differ greatly in their valuations of individual networks.

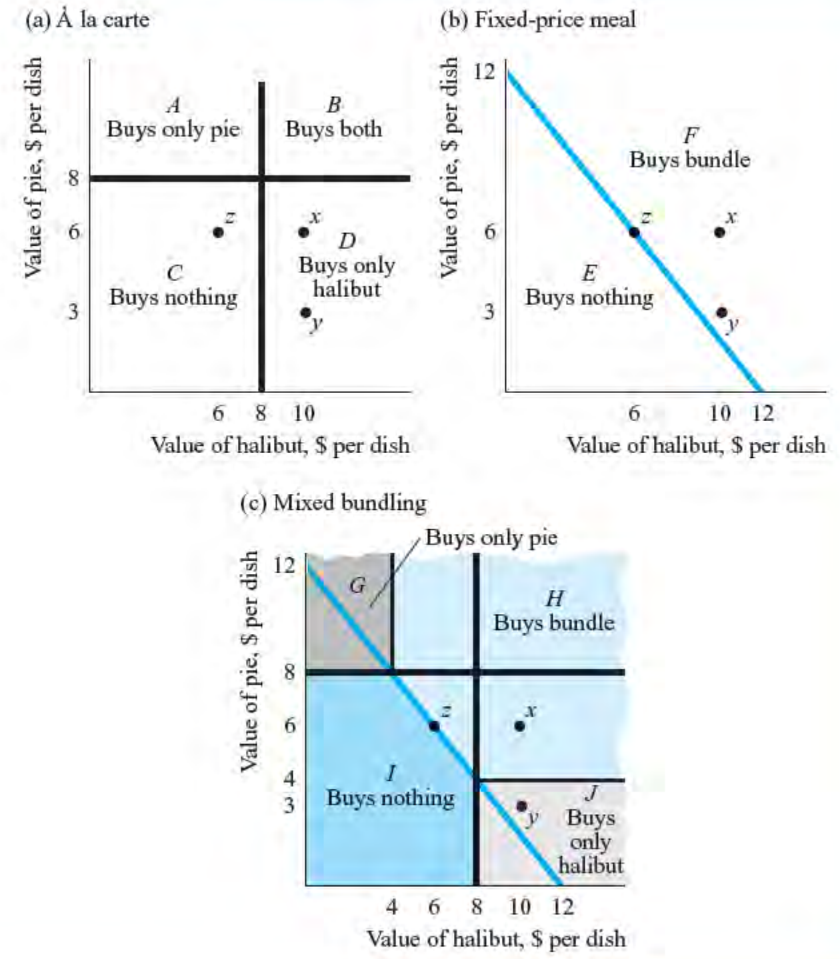
As more networks are added to the bundle, consumers who otherwise would have not purchased a particular network now purchase it as a part of the bundle. Moreover, the more networks in a bundle, the more likely it is that households value the bundle similarly, a preference that results in a more elastic demand curve. Crawford's (2001) empirical study of the bundling decisions of cable providers across the United States confirms these results. By combining several networks into a basic bundle service, a cable provider is able to increase its profit on average above unbundled sales by about 14 percent, according to Crawford's simulations. Consumers lose about 13 percent of the surplus they would have obtained in unbundled sales. Moreover, as price discrimination theory predicts, bundling together similar networks is less profitable than bundling dissimilar ones.

You can profitably bundle only if conditions on market power, resale, and tastes are met. We'll assume that restaurants are monopolistically competitive because their products are not perfect substitutes, so that you have some market power. Certainly you can ignore resale problems. A customer is unlikely to order the fixed-price dinner and then lean over to the person at the next table and offer to sell the appetizer.

It depends on your customers' tastes whether the condition is met that customers who put a relatively high value on one good place a relatively low value on another good. We illustrate how your decision depends on your customers' tastes in Figure 10.4.

For simplicity, suppose your restaurant only sells a main dish of halibut and a piece of pie for dessert. On the axes of the diagrams in the figure are the value or reservation price customers put on each good. In Figure 10.4a, your à la carte menu lists the price of both dishes at $8 each. Any customer in area *D*—who values the fish at more than

FIGURE 10.4 Customers and Mixed Bundling



$8 and the pie at less than $8—comes to your restaurant just for the halibut. An example of such a set of tastes is point *x*, where a customer values the fish at $10 and the pie at $6. People whose reservation prices are in area *A* buy only pie, those in area *B* buy both, and those in area *C* buy neither (point *z* at $6 for halibut and $6 for pie).

Figure 10.4b shows how customers make decisions if you offer only a fixed-price meal of halibut and pie. The customer who values fish at $10 and pie at $6, point *x* in area *F*, buys the bundle at $12, because the customer was willing to pay $16 for the meal. As we saw in Figure 10.4a, this customer would not have been willing to buy the pie if each had been priced separately at $8, where buying both dishes cost $16. Any consumer in *F*, who values the bundle at more than $12, buys the bundle. A consumer who places a value on the bundle of less than \$12, area *E*, would not buy the fixed-price meal. The consumer at *z*, who bought neither good from the à la carte menu buys both from the fixed-price menu.

Figure 10.4c shows mixed bundling. A customer may buy either dish at \$8 each or buy the fixed-price meal bundle at \$12. The customer whose reservation prices are \$10 for fish and \$6 for pie, point *x* in area *H*, would still buy the fixed-price meal. The customer reaps \$4 of surplus from the fixed-price bundle (\$16 – \$12) but only \$2(= \$10 − \$8) of surplus from consuming only halibut.

If, however, this person valued pie at only \$3, point *y* in area *J*, the customer would buy only halibut at \$8. Because the customer places a value of \$13 on consuming both dishes, the consumer surplus from buying the bundle for \$12 is \$1(= \$13 − \$12). If the customer orders only the fish, the consumer surplus is \$2(= \$10 − \$8). Thus, the customer is better off ordering only the fish.[8]

By similar reasoning, consumers in area *G* buy only pie. Finally, consumers in area *I* do not patronize your restaurant because the price of the bundle exceeds the value to them and the price of each dish exceeds the consumers' reservation price.

Which pricing scheme gives you the highest profit depends on both how much your various customers are willing to pay for the two dishes and your costs of producing these dishes. For pure or mixed bundling to pay, your restaurant must sell more food than with individual pricing. Selling more, however, will increase profits only if revenue rises by more than costs.

Suppose you have only three types of customers—*a, b,* and *c*, with valuations of the two dishes in Figure 10.5—and that your cost of producing a dish is \$3 for halibut and \$2 for pie. If you price each dish separately, you maximize your profit by charging \$11 for the halibut and \$8 for pie. At these prices, Customers *a* and *b* buy only pie and Customer *c* buys only halibut. You earn \$6 = \$8 (price of pie) − \$2 (cost of pie) on each of the two servings of pie you sell and \$8 = \$11 − \$3 on the halibut, for a total profit of \$20. Because Customer *a* was willing to pay \$10 for a piece of pie and the price is only \$8, you are not capturing all the consumer surplus.

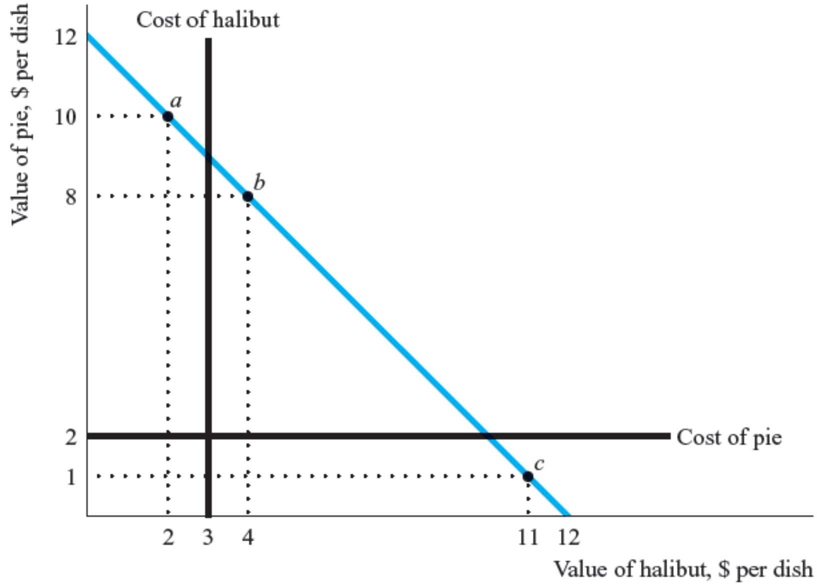
If you only sell a pure bundle, you charge \$12 and earn a profit of \$21 = (\$12 − \$5) × 3, where \$5 is the combined cost of producing both dishes. You make more by bundling than selling separately because customers who value one dish greatly put a lower valuation on the other. As a result, bundling allows you to sell more dishes. You sell two servings of pie and one of halibut with separate prices, whereas you sell three servings of pies and three of halibut with pure bundling.

Using mixed bundling, however, you can do even better. You set the bundle price at \$12, the price of halibut at \$10.99, and the price of pie at \$9.99. Customer *a* buys

[8] Let the reservation price for halibut be r_h, and that of pie be r_p. The price of halibut is p_h, the price of pie is p_p, and the bundle price is p_b. A consumer buys the bundle if

$r_h + r_p \geq p_b$ (the value of the bundle is greater than its price),
$r_p \geq p_b - p_h$ (the reservation value of pie is greater than the difference between the bundle price and the price of halibut), and
$r_h \geq p_b - p_p$ (the reservation value of halibut is greater than the difference between the bundle price and the price of pie).

FIGURE 10.5 Profitability of Mixed Bundling



only the pie (the bundle costs $2.01 more and that customer only values halibut at $2), Customer *b* buys the bundle, and Customer *c* buys only the halibut (you sell two servings of pie and two of halibut). You make $7.99 each from Customers *a* and *c* and $7 from Customer *b* for a total of $22.98.

You sell more dishes with pure bundling—three of each—than with mixed bundling—two of each. The reason your profit is lower with pure bundling is that you're selling dishes to customers who value those dishes at less than your cost of production. With the pure bundle, Customer *a* values the halibut at $2, which is less than your cost of $3, and Customer *c* values the pie at $1, which is less than your cost of $2. As a result, selling those extra dishes doesn't benefit you. If you sell Customer *a* the bundle, you make $7 profit. On the other hand, if you sell Customer *a* the pie at $9.99, you make $7.99 profit. Thus, you benefit from discouraging that customer from buying the halibut. With mixed bundling you are capturing essentially all the consumer surplus.

With these same customers, if you had no cost of production, you would maximize profit by using pure bundling. With no cost of production, you want to sell lots of dishes. In general, depending on your customers' tastes and your costs of production, any of the three pricing methods could maximize your profit. See Example 10.5.

Package Tie-ins with Only One Product Monopolized. Now suppose that the firm is a monopoly of only Product *A* (with no marginal cost of production). Product *B* is competitively produced and sold at *m* (its constant marginal cost of production).

EXAMPLE 10.5 *Not Too Suite—Mixed Bundling*

Word processors and spreadsheets are separate computer-software products. During the 1990s, software producers shifted from selling word-processor and spreadsheet programs separately to selling them as part of a *suite* in which the two software products were bundled together. Consumers could still buy each component separately, so producers were engaged in mixed bundling.

Why did this mixed bundling occur? One answer is that it is efficient to buy a bundle to ensure that the component parts will work together: that is, to ensure quality. Another explanation uses price discrimination theory, saying specifically that mixed bundling would be profit maximizing if consumers who placed a high value on a word processing program (such as Microsoft Word) placed a low value on a spreadsheet program (such as Microsoft Excel), and vice versa.

Gandal (2003) finds that the empirical evidence supports the discrimination theory. In a survey of home PC users, 43% said that they used both programs, 50% used only one, and 7% used neither. Among business PC users, 63% said they used both programs, while 37% used only one. That is, a relatively large fraction of users use only one (but not both) pieces of software. Gandal's sophisticated estimation of consumer demand preferences using a discrete random choice model confirms that, in general, consumers with a high value for spreadsheets had a low value for word processors and vice versa, so that there is indeed a sizable negative correlation in demand for the two.

We continue to assume that the demands for Product *A* and Product *B* are independent, in the sense that consumers' valuation of Product *A* is unrelated to whether or not they also consume Product *B* and vice versa. Does it pay for the monopoly of A to tie Products *A* and *B* together in fixed proportions?

Before analyzing this example in detail, let us use common sense to guess the answer. Suppose that it were profitable for the monopoly to tie a competitive good *B* to the purchase of Product *A*. Because *A* and *B* are independent products, any competitive product could be tied to Product *A* to increase the monopoly's profit. In other words, a monopoly of, say, cars, might be expected to tie all sorts of unrelated products to the sale of cars. We rarely, if ever, see a monopoly tie a completely unrelated product to the sale of its own product. Because we do not observe that, it must generally not be profitable for a monopoly to tie an unrelated, competitively available product to the sale of its monopolized product.

Let us now analyze the example formally. Suppose that a monopoly of Product *A* uses a package tie-in and requires that for every unit of *A* purchased, one unit of Product *B* must also be purchased. The monopoly purchases *B* at the competitive price, m, inserts it into a package with *A*, and charges p^* for the package. The profit for every package sold is therefore $p^* - m$. This tie is not profitable if the monopoly would make more money if it sold Product *A* separately and charged $p^* - m$.

Two types of consumers consider purchasing the product. One type likes Product B. If those consumers do not obtain Product B in the package with Product A, they buy it elsewhere at price m. For these consumers, it is as if they are buying B at price m and paying $p^* - m$ for A. They are completely indifferent whether they pay $p^* - m$ for A and m for B separately or whether they buy A and B together in a package for p^*.

The second type of consumer values a unit of B at less than m (that is, this type of consumer is unwilling to buy B at the competitive price). If they buy the package, they are forced to consume more of Product B than they would have if they had been allowed to buy B separately at the market price, m. For example, a particular consumer may well end up getting B when, in fact, this consumer has absolutely no use for Product B and values it at zero. These consumers will purchase the package consisting of A and B at p^* only if they value A at p^* or above.

If A were sold by itself for $p^* - m$, more of this second type of consumer would purchase Product A than are willing to purchase the package at p^*. For example, consumers who place no value on B but a value of $p^* - m$ on A would buy A separately at $p^* - m$, but would refuse to buy the package at p^*. If the monopoly makes a per-unit profit of $p^* - m$ when the package is sold, and the same $p^* - m$ when A is sold separately (at price $p^* - m$), the monopoly's profits are higher if it sells A separately, because more units of A are sold. In other words, for any price for a package, p^*, a monopoly can always do better by selling A separately for $p^* - m$ than by selling the package for p^*. By packaging A and B together, the monopoly is throwing away sales by forcing some consumers to buy a package that includes a product that they do not value highly. As a result, some consumers who value A reasonably highly do not buy the package. Thus, a monopoly does not have an incentive to package its product in fixed proportions with a good that is competitively produced if the goods are independently demanded.[9]

Interrelated Demands

Very often the demands for goods are interrelated. For example, the value of a camera depends on the availability of film. The price of film influences the demand for cameras, and vice versa. This interrelationship of demand creates incentives to price discriminate through package tie-ins and requirements tie-ins. Before illustrating this point, let us first examine profit maximization with interrelated demands without tie-in sales.

[9] The case of a requirements tie with independent demand is more complicated than that of a package tie. Mathewson and Winter (1977) show that the use of a tie of a competitive and monopoly product with independent demands could be profitable. They give an example of a gas supplier to gas stations that tied the sale of gasoline to batteries and other accessories (which the gasoline supplier did not produce but sold with a markup) so as to "spread out the distortion" of the markup above the efficient transfer price across several products instead of just one.

Profit Maximization with Interrelated Demands. Suppose a firm has a monopoly in two products, *A* and *B*. If the demands are independent, the demand for *A* depends only on the price of *A*, and the demand for *B* depends only on the price of *B*. With interrelated demands, the demand for *A* depends on both the price of *A* and the price of *B*. Similarly, the demand for *B* depends upon the price of both *A* and *B*.

The constant marginal costs of production for Products *A* and *B* are m_A and m_B, the corresponding prices are p_A and p_B, and the corresponding demand curves are $D_A(p_A, p_B)$ and $D_B(p_A, p_B)$. The profit from selling *A* is:

$$\pi_A(p_A, p_B) = (p_A - m_A)D_A(p_A, p_B),$$

where $p_A - m_A$ is its profit per unit of *A* sold. Similarly, the profit from selling *B* is

$$\pi_B(p_A, p_B) = (p_B - m_B)D_B(p_A, p_B).$$

The monopoly's problem is to maximize profits from its sales of the two products, π, which depend on the two prices:

$$\begin{aligned}\pi(p_A, p_B) &= \pi_A(p_A, p_B) + \pi_B(p_A, p_B)\\ &= (p_A - m_A)D_A(p_A, p_B) + (p_B - m_B)\, D_B(p_A, p_B). \qquad (10.1)\end{aligned}$$

In choosing the optimal prices to charge, the monopoly not only considers its profits from the production and sale of *A*, π_A, but also takes into account how the price of *A* affects the profit from *B*, π_B, and vice versa. That is, a monopoly with interrelated products must take the interrelationship into account in determining its optimal prices.

Figure 10.6 illustrates the monopoly's problem. The demand curve for *A* shifts out as the price of *B* falls from $5 to $4. By altering price p_B, the monopoly may be able to shift out its demand curve for *A* in such a way that it can extract a large enough profit from the extra sales of *A* to more than offset any decline in its profit from its sales of *B*.[10] (If p_A changes, D_B would also shift.) Thus, a monopoly of two complementary products may set at least one price higher than would separate monopolies and one price lower.[11]

[10] To choose the profit maximizing p_A and p_B, the monopoly sets the derivatives of profits, π, in Equation 10.1 with respect to each price equal to zero:

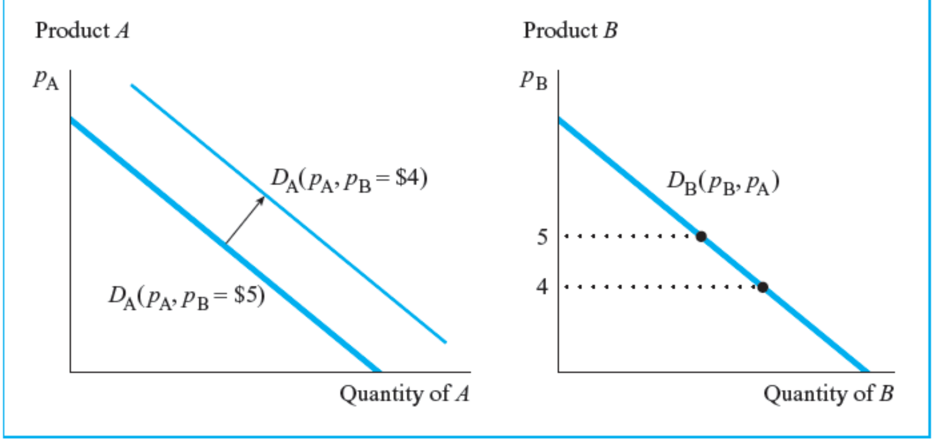
$$\frac{\partial \pi}{\partial p_A} = \frac{\partial \pi_A}{\partial p_A} + \frac{\partial \pi_B}{\partial p_A} = 0,$$

$$\frac{\partial \pi}{\partial p_B} = \frac{\partial \pi_A}{\partial p_B} + \frac{\partial \pi_B}{\partial p_B} = 0.$$

These conditions are different from the conditions that would result if, instead of a single monopoly that controls both p_A and p_B, there were two monopolies, one setting p_A and one setting p_B. Those conditions are $\partial\pi_A/\partial p_A = 0$ and $\partial\pi_B/\partial p_B = 0$, assuming Bertrand behavior.

[11] With substitutes, a single monopoly will generally charge higher prices than separate monopolies that ignore their (negative) price effect on each other's profit.

FIGURE 10.6 Interrelated Demands



Indeed, it may pay to set p_B below its production cost in order to sell A at a higher price. This result is similar to that from a two-part tariff. For example, one could interpret the sale of a camera and film as a two-part tariff in which the lump-sum fee is paid for the camera and the usage fee is paid for the film. As we explained above with respect to the two-part tariff, it could be profitable to charge a price below cost for the camera. We now investigate both package tie-ins and requirements tie-ins when demands are interrelated.

Package Tie-ins with Interrelated Demands. If demands are interrelated, package tie-ins are one method that a monopoly can use to avoid inefficient behavior by consumers and hence increase its profit. For example, automobiles are made from aluminum and steel. The automobile manufacturers' willingness to pay for aluminum depends on the price of steel. Therefore, if output can be produced using variable proportions of the two inputs, the demands for the two inputs are interrelated.

Suppose that the automobile and steel industries are competitive, but that aluminum is provided by a monopoly. The automobile manufacturers choose a combination of aluminum and steel based on the ratio of a monopoly aluminum price to a competitive steel price. Because the price of aluminum is relatively high (higher than the competitive price), they use relatively too much steel and too little aluminum, so that automobile production is inefficient. More disturbing to the aluminum monopoly, relatively little aluminum is purchased.

The aluminum monopoly could force the car manufacturers to sign contracts that require them to use relatively more aluminum in the manufacture of cars. For example, it could force them to use the efficient proportion of aluminum to steel (the ratio that would be chosen if all industries were competitive). The monopoly could impose this restriction by requiring that car manufacturers purchase the efficient amount of steel from the aluminum monopoly, which could purchase the steel on the competitive market. Of course, as with all tie-in sales, if the tie-in is to work, it must be impossible for the consumer, in this case the car manufacturer, to purchase steel secretly on the open market; that is, the consumer must not be able to undo the package tie-in.[12]

Requirements Tie-ins with Interrelated Demands. Perhaps the most common type of tie-in is a requirements tie-in in which consumers buy one good and are then required to make all their purchases of some other related good from the same manufacturer. Chapter 19 examines several examples that have arisen in litigation. One famous case involved the A. B. Dick Company, which had a patent monopoly to sell mimeograph machines.[13] A. B. Dick required that customers who bought mimeograph machines also buy all their ink from A. B. Dick, which did not have a monopoly in ink. Another famous tie-in case was mentioned earlier: the IBM tabulating card case, in which IBM required purchasers of its machines to buy all their tabulating cards from IBM.

In the typical requirements tie, the firm sets a price for the first good and charges a high price (above the competitive price) for the related good. Consumers with large demands effectively pay more for the first good than consumers with small demands. For example, a person who bought a tabulating machine and 100 tabulating cards effectively paid a higher price for the machine than someone who bought only 10 cards. Therefore, a critical element for a requirements tie to maximize profit is that consumers differ in their demand for the related good. We now examine in more detail why requirements ties may be profitable.

Suppose that a firm develops a new machine that automatically sews buttons on shirts. Prior to the development of the machine, buttons were sewn by hand onto shirts, and the labor cost was 1¢ per button. There are many shirt manufacturers. Suppose a large manufacturer sews on 10,000 buttons per year. That manufacturer is willing to pay $100 per year for the machine because its saves $100 from reduced labor costs. Another manufacturer that uses only 1,000 buttons would pay at most $10 for the machine per year.

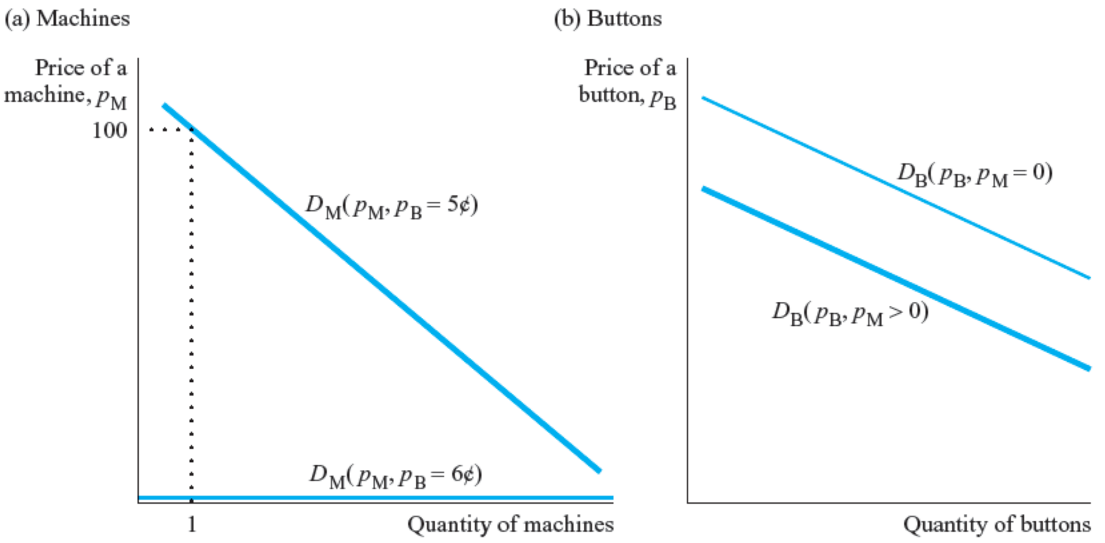
To keep the example simple, suppose that a machine lasts for only a year and that the total number of buttons that each manufacturer sews on shirts during a year is unchanged by this invention. The demand curve for the machine, $D_M(p_M, p_B)$, depends on the price of the machine, p_M, and the price of buttons, p_B. The price of buttons is 5¢ for the solid demand curve in Figure 10.7a.

Suppose that the monopoly of the machine decides to allow firms to use the machine for free provided they purchase all their buttons from the machine monopoly for

[12] An alternative approach is for the aluminum monopoly to take over the automobile industry (vertically integrate) so as to eliminate this inefficiency in production and increase the demand for aluminum and its profit (see Chapter 12). With interrelated demands, a package tie-in sale enables a monopoly to achieve the same increase in profit that it could achieve through vertical integration. Note also that the monopoly could achieve the same result as with the package tie-in by using a requirements tie-in in which the relative prices charged for aluminum and steel were in the same ratio as their marginal costs. Such prices will lead to an efficient use of aluminum and steel.

[13] *Henry v. A. B. Dick Co.,* 224 U.S. 1 (1912). A mimeograph machine produced copies of an original stencil using ink.

FIGURE 10.7 Shifts in Demand for Machines and Buttons as a Result of Tie-in



p_B = 6¢ (or perhaps a shade less), which is 1¢ above the competitive price. In other words, the monopoly ties the sale of buttons to the sale of the machine and charges a 1¢ premium (or a shade less) on each button. Any firm that has a use for the machine agrees to these conditions because the machine saves 1¢ per button.

As a result of this tie-in, the largest users of buttons pay a higher effective price for the machine. For example, the firm that uses 10,000 buttons effectively pays $100 for the machine; however, the firm that uses only 1,000 buttons pays only $10 for it. Thus, a tie-in between buttons and machines enables the monopoly of the machine to charge customers effectively different prices for the machine, charging the most to those who value the machine the most. This tie-in enables the monopoly to extract all consumer surplus from under the demand curve. That is, it allows the monopoly to achieve perfect price discrimination.

We can relate this example of a tie-in between a machine and buttons to the earlier examination of interrelated demand curves. If the monopoly charges 6¢ per button and captures all the consumer surplus, customers are unwilling to pay anything for a machine (the demand curve for the machine becomes the black line that lies on top of the horizontal axis in Figure 10.7a).

The adoption of the machine makes the demand curve for buttons shift out. As Figure 10.7b illustrates, if the machine is given to the user for free, the demand curve for buttons shifts up by one penny above where it was before the machine existed. (The thick blue line in Figure 10.7b can be thought of as the demand for buttons when the price of machines is set so high that no one purchases a machine, while the thin blue line is the demand for buttons when the price of the machine is zero.) As the price of the machine rises, the demand curve for buttons eventually falls to the initial demand curve.[14] Thus, the tie-in, which sets $p_M = \$0$ and $p_B = 6¢$, shifts down the demand curve for the machine and shifts up the demand curve for buttons relative to where they would be if $p_M > 0$ and $p_B = 5¢$. The tie-in allows the firm to perfectly price discriminate, so it is more profitable than setting any single positive price for the machine and not selling buttons (or selling them at the competitive price).

The preceding example is special in the sense that each firm has a pre-established demand for buttons and that it costs each firm the same to sew buttons on by hand. It is more likely that each firm has some flexibility over how many buttons it purchases and that the firms differ in the values they attach to sewing on buttons by machine. In response to a tie-in sale of a machine and buttons in which users are overcharged for buttons, some users might reduce their use of buttons. If so, the tie-in does not allow perfect price discrimination as in the simpler example. However, the tie-in may still be the most profitable approach even if all consumer surplus cannot be extracted. The tie-in can be regarded as a two-part tariff in which the machine's price is a lump-sum payment and the price of buttons is a per-unit charge. The tie-in sale cannot achieve perfect price discrimination for the reason discussed earlier that a two-part tariff cannot generally achieve perfect price discrimination,[15] that is, the firm is unable to separately identify and charge each consumer the most that consumer is willing to pay.

In the examples in this section, the tie-in serves to meter usage. Alternatively, the firm may meter usage explicitly. For example, IBM could have installed a meter to measure the number of cards each of its customers punched on the IBM machines. Customers could purchase cards anywhere, but the price they paid IBM for the machine would depend on their measured card usage. The choice between using a tie-in or an explicit meter depends on the relative costs of the two methods. Meters may be costly, and they may be easy to unhook, trick, or break. Tie-ins, on the other hand, may be hard to police (for example, customers could buy cards elsewhere), and they could distort efficient use. Photocopiers, telephones, and electric utilities often use explicit meters to monitor usage.

We have seen how a tie-in sale between two goods can raise the profit of a monopoly. More generally, a monopoly might require not just one good but several to be used in conjunction with its monopolized good. A monopoly could also specify that particular inputs must *not* be used with its monopolized goods.

[14] The decisions of whether to purchase the machine and how many buttons to buy are made simultaneously. No one would pay to purchase the machine and pay an extra 1¢ per button. The higher the price of the machine, the less buyers are willing to pay per button.

[15] If the number of buttons used per shirt is variable, and if, in response to an increase in the price of buttons, some shirt manufacturers use fewer buttons, the machine owner may wish to specify as a condition of purchase the minimum number of buttons that can be sewn on per shirt.

Quality Choice

Imagine a monopoly that produces several goods of differing quality—for example, a monopoly of cars of high, medium, and low quality. Goods of different qualities are typically related on the demand side because consumers can substitute among them. If the qualities of the goods are prespecified, the monopoly's decision about pricing them is the same as the one already examined in the discussion of interrelated demands. However, if the monopoly must also decide on the qualities of the goods to be produced, then it takes the demand interrelationship into account not only when it decides how to price the various quality goods, but also when it decides on the qualities of the goods it offers for sale (Mussa and Rosen 1978). Thus the monopoly's decisions about levels of quality are influenced by the same forces that influence the choice of nonlinear pricing schemes.

For example, the automobile monopoly may choose to produce only very high-quality and very low-quality cars. By not providing close substitutes for the high-quality cars, the monopoly may be able to extract more profit from selling them than it could if it also produced medium-quality, moderate-priced cars that were good substitutes for the high-priced, high-quality cars. The reason is that the firm can charge a high price for the high-quality car and not worry about consumers substituting to the low-priced, low-quality car because it is not a good substitute.

In summary, when consumers prefer different levels of quality, a monopoly manipulates the qualities of goods produced in the market in order to extract consumer surplus. The monopoly follows the principles already studied regarding price discrimination and chooses the quality spectrum so as to charge a high price to those who value the good the most, and a low price to those who value it the least, without having to pass along the low price to those who value the good the most (see Mussa and Rosen 1978; see also **www.aw-bc.com/carlton_perloff** "Quality Choice of a Monopoly").

Other Methods of Nonlinear Pricing

This chapter discusses only some of the many possible pricing schemes that monopolies can adopt as they attempt to maximize profits. A few other pricing schemes are quite common and deserve mention.

Minimum Quantities and Quantity Discounts

Many sellers specify that their product can be bought only in certain minimum amounts. Such a restriction causes pricing to be nonlinear. The average price per unit consumed is very high for small quantities and is lower after consumption reaches the minimum purchase level. A similar effect is achieved by granting quantity discounts.

Selection of Price Schedules

Sometimes consumers must choose the pricing schedule that will govern their purchases *before* they know how much they will purchase. For example, some telephone companies require consumers to select a pricing schedule at the beginning of the month. Some consumers elect to pay a large fixed fee and have unlimited calling; others elect to pay a modest lump-sum fee that entitles them to make calls and pay extra for calls in excess of a certain amount. At the end of the month, a consumer may discover that the pricing schedule not chosen would result in a lower bill. By requiring customers to specify in advance the pricing schedule they will face, monopolies can discriminate between those who can accurately predict their demands and those who cannot. Those who cannot accurately predict may overpay relative to those who can predict accurately.

In contrast, electricity companies generally do not require customers to choose a pricing schedule in advance. Instead, consumers of electricity typically face a declining block schedule where high prices are charged for the initial usage and lower prices are charged thereafter. Because this schedule applies to all consumers, the bill at the end of the month is independent of the customers' ability to predict their demands.

Another related example involves the purchase in advance of a fixed amount of a product for a lower price than for smaller, as-needed purchases. For example, many commuter railroads sell individual tickets at much higher prices per ride than monthly passes. If consumers misestimate how frequently they will travel, the railroad profits from their mistake.

Premium for Priority

If consumers differ in their desires to obtain a good quickly, a firm can charge more for rapid delivery. For example, a common pricing strategy for new goods is to price high initially and then to lower price over time. Airlines often charge more for tickets ordered one day in advance than for those ordered several weeks in advance. One possible reason for this pricing behavior is that business people, who often travel on short notice, have a less elastic demand than tourists, who do not travel on short notice. In general, when obtaining a good is uncertain, it may be possible to price discriminate by charging different prices for different probabilities of obtaining the good (Harris and Raviv 1981, Maskin and Riley 1984).[16] If, however, customers impose different costs on the firm (as is likely for customers who order in advance), labeling the price differences as price discrimination may be misleading.

[16] The welfare implications of such pricing schemes are very complex. For example, Gale and Holmes (1992) show that advance-purchase discounts offered by an airline can lead to an efficient allocation of capacity between peak and off-peak flights when it is not possible to operate a spot market on the day of a flight. Moreover, the socially optimum discount may be either larger or smaller than that offered by the monopoly.

Auctions

Some firms use auctions to sell valuable assets, such as art, antiques, off-shore oil leases, and Treasury bills. The purpose of an auction is to obtain the maximum revenue from buyers when the seller does not know which buyers value the goods the most. The objective is to design a pricing mechanism that induces the consumers with the greatest willingness to pay to bid high prices.

What is the best way to conduct an auction to obtain the maximum revenue? Should it be an auction in which bids start low and rise until there is no one willing to bid any higher (**English auction**)? Should it be an auction in which the price starts out very high and is slowly lowered until one person agrees to buy at that price (**Dutch auction**)? Should a minimum bid be specified? The answer to these questions is, under plausible assumptions, surprisingly simple. If buyers maximize expected consumer surplus and have independent valuations of the item in the auction, (such as idiosyncratic tastes for a painting) Dutch and English auctions yield the same expected revenues, and it is optimal to set minimum bids.[17]

This result does not hold when the buyers share a common value for an item, as in a case where the item will be resold to consumers, such as an auction for wholesale oil that is eventually resold to final consumers. Here bidders must guard against placing too high a bid because they overestimate the common value (eventual resale price). Such excessive bidding is called the winner's curse because the auction winners are likely to have overpaid owing to their belief that they are better able to estimate the expected market value of the item than other bidders. See Example 10.6.

SUMMARY

If a firm with market power lacks detailed knowledge about the demands of individual consumers, it cannot charge different prices to different consumers so as to maximize its profit. Instead, the firm must offer the same pricing policy to all consumers and let them choose (self-select) how much to pay and consume. However, a firm can earn a higher profit than if it set a single price by using nonlinear pricing policies. Many nonlinear pricing policies—such as a menu of two-part tariffs—induce different consumers to behave differently from each other and to pay different prices.

Tie-in sales work similarly to two-part tariffs and other nonlinear price schemes. They cause different consumers to pay different prices. Both package and requirements tie-in sales increase a firm's profit under appropriate circumstances.

Other pricing policies in addition to nonlinear price schedules are widely used. These policies allow a firm with market power to earn a higher profit than if it charged a single price to everyone. These policies include quality choice, auctions, priority of delivery, and minimum purchase orders.

[17] McAfee and McMillan (1987) and Klemperer (1999, 2001) survey the results from the large literature on auctions.

EXAMPLE 10.6 *Price Discriminating on eBay*

The online firm eBay conducts auctions on millions of items. In a typical eBay auction, the seller sets a minimum bid and then buyers submit their bids up to their maximum willingness to pay. If a new bidder places a bid that exceeds the current highest bid, then the new highest bid listed on eBay's computer is the previous highest bid plus an increment (50¢ for low-price items). Thus, the auction allocates the good to the highest bidder at a price equal to the maximum willingness to pay of the second-highest bidder, plus an increment.

A seller in an auction is a monopoly and wants to obtain the highest price. How should the monopoly set the minimum bid? Bulow and Roberts (1989) answer this question using the principles of price discrimination.

In an ascending (English) auction, the high bidder wins at a price just above that of the next highest bidder. That means that if no one else bids, the high bidder obtains the good for the minimum bid. Thus, the seller has an incentive not to set an extremely low minimum bid. However, the disadvantage of a relatively high minimum bid is that it dissuades some bidders from entering the auction (forcing the eventual winner to pay a higher price). So, the seller should set the minimum bid taking account of this trade-off.

Bajari and Hortacsu (2003) use sophisticated econometric techniques to analyze eBay auctions for collectible coins. Because many collectors resell their coins to other collectors, they should all place a common value on any given coin. In a common value auction, the winning bidders should worry that they have overbid—suffered from the *winner's curse*—because they overestimated this common value. Sophisticated bidders should therefore reduce their bids in common value auctions as the number of bidders increases.

Bajari and Hortacsu find that in a typical auction, the minimum bid is set on average at about 70 percent of the current retail market value of the collectible coin. They also find that more bidders enter the auction when the minimum bid is low. For example, they estimate that about four to five bidders show up when the reserve price is zero, but only two enter if the minimum bid is set at 80 percent of the retail price. Bidders do recognize the winner's curse and reduce their bids by 3.2 percent for each additional bidder in the auction. Finally, they estimate that the optimal minimum bid that maximizes expected seller revenue is about 10–20 percent below the current retail price. They conclude that sellers on eBay are doing a pretty good job of setting minimum bids in order to capture the consumer surplus of the high bidder.

PROBLEMS

1. If a firm faces identical consumers and uses a two-part tariff, will its marginal price be above its marginal cost? If not, how does it make a profit?
2. In an English auction with only two bidders, can you describe a situation where a reserve price is desirable?
3. A person who consumes X units of Good 1 and Y units of Good 2 derives utility of $Y + 10X$. Suppose the person has \$100, the price of Y is \$1, and the nonlinear expenditure for purchasing X units of Good 1 is X^2. What X maximizes that person's utility?
4. Suppose a manufacturer sells a button-fastening machine that saves a firm the labor cost of 1¢ per button sewn on shirts. Suppose firms differ in the total number of buttons they sew on. The manufacturer sells its machine with a requirements tie-in that requires a purchaser to buy all its buttons from the manufacturer. Suppose the manufacturer can install a meter that measures how many buttons each machine sews. If the manufacturer can charge according to the use measured on the meter, is there any advantage to the tie-in? Would it be sensible to outlaw tie-in sales but allow the manufacturer to charge according to the metered use?
5. Let the demand for Products 1 and 2 be $q_1 = 10 - 2p_1 + p_2$ and $q_2 = 10 + p_1 - 2p_2$, where q_i is the quantity of Good i and p_i is the price of Good i. Assume production costs are zero. Calculate the prices that two separate monopolies would charge when each regards the other's price as beyond its control. Calculate the prices that a single monopoly of both goods would charge.
6. A monopoly produces and delivers goods to consumers who are located at varying distances from the factory. It costs m per unit to produce the good and \$1 per mile to transport a unit of the good. Resales are impossible. Calculate the price that a monopoly charges consumers at location t if demand is $q_t = a - bp_t$, where q_t and p_t are the quantity and price at location t. How does p_t change as t increases? Who bears the freight cost?
7. In Figure 10.5, if your costs of production are \$1 each for halibut and pie, which pricing scheme—individual pricing, pure bundling, or mixed bundling—maximizes your profit?

Answers to odd-numbered problems are given at the back of the book.

APPENDIX 10A

The Optimal Two-Part Tariff

The following problem illustrates the forces that determine the optimal two-part tariff. Let p be the per-unit usage price, T the lump-sum fee, N the number of demanders, and Q the total amount of the product demanded. The quantity demanded, Q, is a function not only of p (price) but also of the lump-sum fee (T). Imagine that consumers can be indexed by the parameter α. The higher is α, the more consumers are willing to pay for the product. Let $f(\alpha)$ be the number of consumers of type α.

The parameter α varies between $\underline{\alpha}$ and $\overline{\alpha}$. For any choice of p and T, there is a critical level, α^*, such that consumers whose α exceeds α^* (who value the product more than α^* type consumers) purchase the product. Consumers whose α is below α^* choose not to purchase the product. The consumer of type α^* is called the *marginal* consumer. Let $S(p, \alpha)$ be an α-type individual's consumer surplus at price p in the absence of a fixed fee, T. Then the marginal consumer (the consumer who is indifferent between buying and not buying) is one whose surplus equals the lump-sum fee:

$$S(p, \alpha^*) = T. \tag{10A.1}$$

For the marginal consumer, the surplus obtained from paying price p is exactly equal to the lump-sum fee, T, so that the total surplus from the purchase is zero. The number of consumers who purchase the product (the number of consumers whose α is greater than α^*, which from Equation 10A.1 depends on p and T) is

$$N(p, T) = \int_{\alpha^*}^{\overline{\alpha}} f(\alpha)\mathrm{d}\alpha. \tag{10A.2}$$

If $q(p, \alpha)$ is the demand curve of an α-type consumer, then the total amount demanded as a function of p and T equals the sum of all the demands of consumers whose α exceeds α^*:

$$Q(p, T) = \int_{\alpha^*}^{\overline{\alpha}} q(p, \alpha)f(\alpha)\mathrm{d}\alpha. \tag{10A.3}$$

If marginal cost is constant and equals m, then the firm's profit is

$$\pi = N(p, T)T + (p - m)Q(p, T), \tag{10A.4}$$

where $(p - m)$ is the profit per unit and NT is the total of the lump-sum fees collected. The firm maximizes its profit by choosing p, the price, and T, the lump-sum fee, optimally. The discussion at **www.aw-bc.com/carlton_perloff** "Derivation of the Optimal Two-Part Tariff," shows that the first-order condition for the determination of price is

$$\frac{p - m}{p} = -\frac{1}{\epsilon}\left(1 - \frac{q^*}{\bar{q}}\right), \tag{10A.5}$$

where $q^* = q(p, \alpha^*)$ is the demand of the marginal consumer, $\bar{q} = Q(p, T)/N$ is the average quantity demanded across all consumers who purchase (those with a $\alpha \geq \alpha^*$), and ϵ is the price elasticity of demand of consumers who purchase the good:

$$\epsilon = \int_{\alpha^*}^{\bar{\alpha}} \frac{q}{Q}\frac{p}{q}\frac{\partial q(p, \alpha)}{\partial p} f(\alpha)\mathrm{d}\alpha. \tag{10A.6}$$

Notice that ϵ differs slightly from the usual price elasticity of demand, which accounts for the change in α^* as p rises.

Equation 10A.5 would be the same as the optimal first-order condition of a simple monopoly ($[p - m]/p = -1/\epsilon$) were it not for the last term in the parentheses on the right side. That term is the ratio of the purchases of the marginal user (that is, the demand of the α^* consumer) to the purchases of the average user in the marketplace times $1/\epsilon$. The ratio $q^*/\bar{q}$ is less than 1 if, as in the usual case, the marginal purchaser buys less of the good than does the average purchaser.

Suppose that all consumers are identical, so that q^* equals $\bar{q}$. Equation 10A.5 becomes

$$(p - m)/p = 0, \tag{10A.7}$$

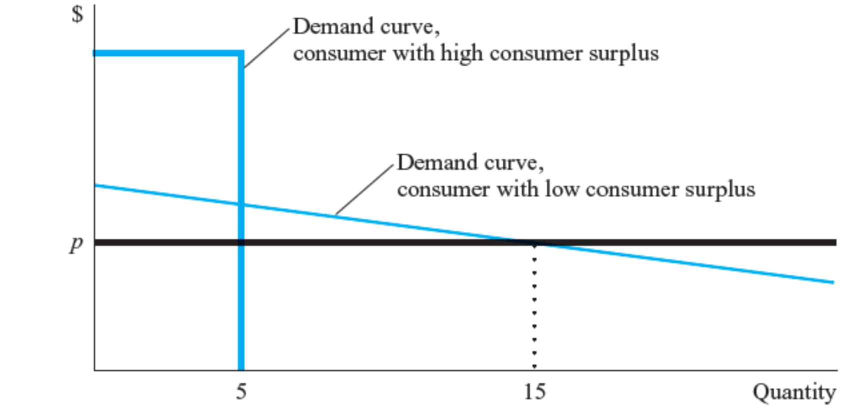
which implies that price should equal m. That is, if all consumers are identical, it is optimal to charge each consumer marginal cost. All of the profits then come from the lump-sum fee, T, as discussed in the chapter.

In the usual case, the marginal consumer (who is indifferent between buying and not buying) demands a lower quantity than other consumers so that q^* is less than $\bar{q}$. (The demand curve of the marginal consumer lies below that of other consumers.) Therefore, the term in parentheses on the right side of Equation 10A.5 is positive. Thus, in the usual case, price exceeds m. For the usual case, the usage-sensitive price is closer to m as ϵ increases in absolute value and as consumer diversity, as measured by the difference between 1 and the ratio of the marginal to average purchase, declines.

It is possible, however, that it is profitable for the usage charge to be below unit cost if consumers who get extensive surplus from the product at a given price buy only small amounts (e.g., 5 units), whereas the marginal consumers who get very little surplus buy large amounts (e.g., 15 units). This case is illustrated in Figure 10A.1. The intuition in this unusual case is that it is profitable for the firm to lower the price below m in order to raise the lump-sum charge to both types of consumers.

The optimal policy cannot involve a negative value for T, which would mean that people are *paid* a lump-sum amount for the right to consume the good whether or not

FIGURE 10A.1 Unusual Configuration of Demand Curves



they consume it. Obviously, if any manufacturer offered to pay people whether or not they consumed the good, everybody would sign up and bankrupt the manufacturer.[1] Hence, in the optimal solution, the lump sum, *T*, is positive or zero. A two-part tariff generally produces higher profits than a single-price policy because a single price is a special case of a two-part tariff in which the lump-sum fee is zero.

[1] This conclusion is overstated if the costs associated with collecting a gift (for example, time) are very large.

APPENDIX 10B

Nonlinear Pricing with an Example

With nonlinear pricing, Consumer *i* faces the following problem:

$$\max_{q_i, y_i} u_i(q_i, y_i) \qquad \text{s.t.} \quad E(q_i) + y_i = I_i, \tag{10B.1}$$

where $E(q_i)$ is the total expense when q_i units are consumed of Good 1, y_i represents all other goods whose per-unit price is normalized to 1, I_i is the consumer's income, and u_i is the consumer's utility function.

The maximization problem facing a firm that wishes to offer a nonlinear pricing schedule that maximizes its profit is to

$$\text{choose } E(\cdot) \text{ to maximize } \sum_i [E(q_i) - mq_i], \tag{10B.2}$$

subject to the constraint that each consumer *i* maximizes as in Equation 10B.1, $i = 1, \ldots, N$, where m is the constant marginal cost, and q_i is the amount consumed by consumer *i*.

The firm chooses the $E(q)$ to maximize profit, which equals the sum of the profits made by selling to each consumer, subject to the condition that each consumer maximizes his or her utility. (Each consumer who purchases Good 1 must be better off by doing so than by forgoing consumption of the good entirely.)

The solution to the problem posed in Equation 10B.2 is complicated (Katz 1983; Spence 1977b; Tirole 1988, Ch. 3). Rather than present the general solution, we illustrate some of the key ideas that arise in nonlinear pricing through an example where the firm can only use two-part tariffs.

Consider the problem discussed in the chapter of a firm that faces two consumers, Consumer 1 and Consumer 2, and offers two two-part tariffs (T_1, p_1) and (T_2, p_2). If it costs m to produce one unit of the good, and if Consumer 1 chooses (T_1, p_1) and Consumer 2 chooses (T_2, p_2), then profit equals

$$T_1 + (p_1 - m)q_1(p_1) + T_2 + (p_2 - m)q_2(p_2), \tag{10B.3}$$

where $q_i(p_i)$ is the demand curve of Consumer *i.* The key additional requirement of the equilibrium in this problem is that Consumer 1 prefers (T_1, p_1) to (T_2, p_2) and vice versa for Consumer 2. Let $U_i(T, p)$ be the utility of Consumer *i,* which depends on T and p. The self-selection constraints are

$$\begin{aligned} U_1(T_1, p_1) &\geq U_1(T_2, p_2), \\ U_2(T_1, p_1) &\leq U_2(T_2, p_2). \end{aligned} \tag{10B.4}$$

Let $S_i(p)$ be the consumer surplus of Consumer i at price p in the absence of a lump-sum fee. Then utility can be written as $U_i(T, p) = S_i(p) - T$. The pricing problem facing the firm is

$$\begin{aligned} \max_{T_1, p_1, T_2, p_2} \quad & T_1 + (p_1 - m)q_1(p_1) + T_2 + (p_2 - m)q_2(p_2) \\ \text{s.t.} \quad & S_1(p_1) - T_1 \geq S_1(p_2) - T_2 \\ & S_2(p_1) - T_1 \leq S_2(p_2) - T_2 \\ & S_1(p_1) - T_1 \geq 0 \\ & S_2(p_2) - T_2 \geq 0. \end{aligned} \tag{10B.5}$$

The objective function in Equation 10B.5 is the total profit that the firm earns from charging a lump sum, T_1 and T_2, and charging price p_1 when quantity q_1 is consumed, and price p_2 when quantity q_2 is consumed. The constraints in Equation 10B.5 are the consumers' self-selection constraints. The first constraint guarantees that Consumer 1 has more utility with a tariff (T_1, p_1) than with the tariff (T_2, p_2). The second constraint guarantees that the utility of Consumer 2 at (T_2, p_2) is greater than at (T_1, p_1). These two constraints guarantee that Consumer 1 chooses (T_1, p_1). and Consumer 2 chooses (T_2, p_2). The last two constraints in Equation 10B.5 guarantee that both consumers have positive utility.[1]

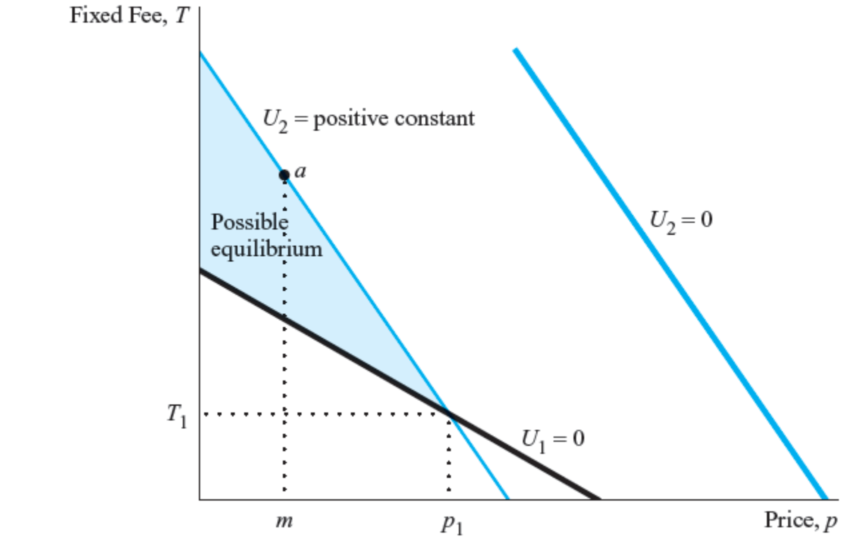
To illustrate the principles involved in nonlinear pricing, suppose that the demand of Consumer 2 is $\lambda(> 1)$ times larger than the demand of Consumer 1: $q_2(p) = \lambda q_1(p)$. That is, the demand curve of Consumer 2 lies strictly to the right of the demand curve of Consumer 1.

Figure 10B.1 shows the indifference curves of consumers in (T, p) space—that is, the combinations of T and p that leave a consumer indifferent. As T falls, p rises along an indifference curve; consumers trade off a higher T for a lower p. They receive higher utility as they move toward indifference curves closer to the origin. Along an indifference curve, when p falls, the amount by which T rises depends on the amount that consumers purchase. Those who purchase a large amount of the good are willing to pay a much higher fixed fee as the per-unit price of the good falls. Therefore, as Figure 10B.1 shows, the indifference curve for Consumer 2 is steeper than the indifference curve for Consumer 1 when the curves cross.

The equation of the indifference curve for $U_2 = 0$ is $T_2 = S_2(p)$, and that for $U_1 = 0$ is $T_1 = S_1(p)$. Along its zero utility curve, a consumer must pay a fixed fee exactly equal to surplus. Along the indifference curve at which Consumer 2 is just indifferent $(U_2 = 0)$ between purchasing the good or not, Consumer 1 does not

[1] It is possible that the optimal solution involves satisfying only one consumer type and that the other consumer type does not consume the good. We consider only the possibility in which it is profit maximizing for the firm to serve both types of consumers because this chapter has already examined the profit-maximizing two-part tariff when only one homogeneous group is involved.

FIGURE 10B.1 Indifference Curves in (*T*, *p*) Space for Type 1 and Type 2 Consumers



purchase the good. Consumer 1 does not purchase because, at any price p, the surplus enjoyed by Consumer 2 is higher than it is for Consumer 1 and Consumer 2 is just indifferent between buying and not. Therefore, Consumer 1 would rather forgo consumption of the good than pay a fixed fee.

The monopoly's optimal solution to Equation 10B.5 involves driving the utility of at least one of the two consumer types down to zero. If both types have positive utilities, the monopoly can raise the fixed fees, continue to sell its product, and still make more money. Thus, the monopoly raises the fixed fees until at least one type's utility is driven to zero.

Which type will that be? Suppose it were Consumer 2 whose utility is driven to zero. That means that Consumer 2 is on its zero utility curve in Figure 10B.1. Where, then, could Consumer 1 be? Consumer 1's utility is negative along the $U_2 = 0$ curve; therefore, the only (T, p) combinations that will keep Consumer 1 in the market are those that lie below the $U_2 = 0$ curve. But if there were a two-part tariff (T, p) that lay below the $U_2 = 0$ curve, Consumer 2 would prefer that point to the one on the $U_2 = 0$ curve. Thus, Consumer 2 cannot have zero utility in the optimal solution provided that both Consumer 1 and Consumer 2 are purchasing. By this reasoning, in the optimal solution involving both consumer types, the utility of Consumer 1 must be zero.

Thus, (T_1, p_1) lies on the curve $U_1 = 0$. Where is (T_2, p_2)? The answer is that (T_2, p_2) cannot lie below the curve $U_1 = 0$. Otherwise, Consumer 1 would prefer that two-part tariff to its own. The self-selection constraints guarantee that the utility of Consumer 2 at its two-part tariff cannot be less than the utility at (T_1, p_1). Hence, (T_2, p_2) can only lie within the shaded region of Figure 10B.1. The monopoly wants to extract as much surplus as possible from Consumer 2 and still satisfy the self-selection constraints. The monopoly achieves the goal by moving as far upward as it can for any p_2 and still remain in the shaded area in Figure 10B.1. Therefore, (T_2, p_2) lies along the upper part of Consumer 2's indifference curve that passes through (T_1, p_1). Consumer 1 would always prefer to remain at (T_1, p_1) rather than at any (T_2, p_2) point that lies above the $U_1 = 0$ curve and along the U_2 curve through (T_1, p_1) and Consumer 2 is indifferent between (T_1, p_1) and points on its indifference curve through (T_1, p_1), and, for simplicity, we assume that Consumer 2 chooses (T_2, p_2) if indifferent between (T_1, p_1) and (T_2, p_2).

These insights help solve the problem in Equation 10B.5. We have established two results. First, the utility of Consumer 1 is zero in the optimal solution so that $T_1 = S_1(p)$. Second, the utility of Consumer 2 at (T_1, p_1) must equal its utility at (T_2, p_2). Based on these two results,

$$S_1(p_1) - T_1 = 0, \tag{10B.6a}$$

$$S_2(p_1) - T_1 = S_2(p_2) - T_2. \tag{10B.6b}$$

We can solve for T_1 and T_2 in terms of p_1 and p_2 from Equations 10B.6a and b and substitute into Equation 10B.5 to reexpress the problem facing the firm as

$$\begin{aligned}\max_{p_1, p_2} \; & S_1(p_1) + (p_1 - m)q_1(p_1) + S_2(p_2) - S_2(p_1) \\ & + S_1(p_1) + (p_2 - m)q_2(p_2).\end{aligned} \tag{10B.7}$$

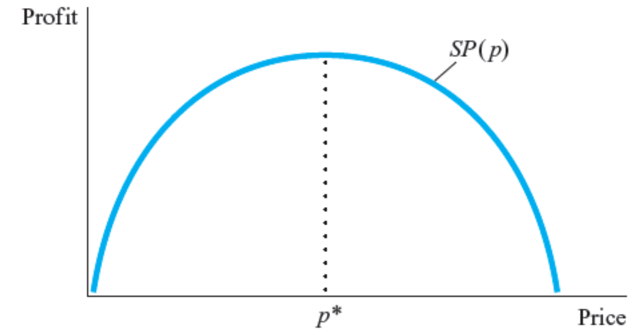
By assumption, for any p, $\lambda q_1(p) = q_2(p)$, and therefore $\lambda S_1(p) = S_2(p)$. As a result, we can rewrite Equation 10B.7 as

$$\begin{aligned}\max_{p_1, p_2} \; & S_1(p_1) + (p_1 - m)q_1(p_1) + \lambda S_1(p_2) - \lambda S_1(p_1) \\ & + S_1(p_1) + \lambda(p_2 - m)q_1(p_2) \\ \equiv \; & (2 - \lambda)S_1(p_1) + SP(p_1) + \lambda S_1(p_2) + \lambda SP(p_2),\end{aligned} \tag{10B.8}$$

where $SP(p_i) = (p_i - m)q_1(p_i)$ is the "standard profit" that a single price monopoly facing demand curve $q_1(p_i)$ earns at price p_i. The function *SP(p)* reaches a maximum at p^*, as shown in Figure 10B.2, which is the price that a standard, single-price monopoly would charge. To the left of p^*, the slope of *SP(p)* is positive, and to the right of p^*, the slope of *SP(p)* is negative.

We are now ready to determine the p_1 and p_2 that maximize Equation 10B.8. The first-order conditions are

FIGURE 10B.2 Monopoly Profit



$$(2-\lambda)\frac{dS_1(p_1)}{dp_1}+\frac{dSP(p_1)}{dp_1}=0, \qquad (10B.9a)$$

$$\frac{dS_1(p_2)}{dp_2}+\frac{dSP(p_2)}{dp_2}=0. \qquad (10B.9b)$$

Because $dS_1/dp_1 = -q_1(p_1)$ and $dS_1/dp_2 = -q_1(p_2)$,[2] we can write Equation 10B.9a as

$$\frac{dSP(p_1)}{dp_1}=(2-\lambda)q_1(p_1), \qquad (10B.10a)$$

$$\frac{dSP(p_2)}{dp_2}=q_1(p_2). \qquad (10B.10b)$$

From Equation 10B.10a, if $\lambda > 2$, the optimal p_1 is such that the slope of $SP(p_1)$ is negative; that is, p_1 exceeds p^*. According to Equation 10B.10b, p_2 should be chosen where the slope of $SP(p)$ is positive; hence p_2 is less than p^*. In other words, when the demand of Consumer 2 is large ($\lambda > 2$) relative to that of Consumer 1, Consumer 1 is charged a very high price—indeed, a price above the simple profit-maximizing price—so that Consumer 2 can be charged a low price and a high lump sum. Although profit is forgone by charging Consumer 1 a price above p^*, the extra profit from Consumer 2 more than compensates. Consumer 1's price is high so that the

[2] As price falls by $1, the additional surplus equals the quantity consumed (ignoring income effects).

(T_1, p_1) tariff is not attractive to Consumer 2, so Consumer 2 is willing to pay a high T_2. By using two two-part tariffs, it is possible to separate Consumers 1 and 2 and limit the problem of having to pass along the low p_2 to Consumer 1 and the low T_1 to Consumer 2.

When $1 < \lambda < 2$, at p_1, the slope of $SP(p)$ is positive by Equation 10B.10a so that $p_1 < p^*$. As Figure 10B.1 shows, $p_2 < p_1$; hence, $p_2 < p_1 < p^*$.

The most efficient method of price discrimination against Consumer 2—and one that does not interfere with self-selection by Consumer 1—is to set price at marginal cost and charge a high lump-sum fee. Hence $p_2 = m$ in the profit-maximizing solution,[3] and the optimal (T_2, p_2) combination therefore is a point like *a* in Figure 10B.1.

In summary, the optimal tariff depends on how large λ is. The per-unit price to Consumer 1 always exceeds that to Consumer 2, and the fixed fee to Consumer 1 is always less than that to Consumer 2. The per-unit price to Consumer 2 equals marginal cost. The presence of Consumer 1 constrains the T_2 and p_2 that Consumer 2 can be charged. The larger the relative demand of Consumer 2 (higher λ), the more profitable it is to forgo profits on Consumer 1 (that is, charge p_1 in excess of p^*) in order to charge a high lump-sum fee and low per-unit price to Consumer 2. Consumer 2 is better off when Consumer 1 is present in the market. From the solution to Equation 10B.5, we know that $U_2(T_2, p_2) > 0$, but, if Consumer 2 were the only customer, the optimal two-part tariff would extract all consumer surplus so that U_2 would be zero. Consumer diversity helps those with the greater willingness to buy the good. In contrast, Consumer 1's utility is completely unaffected by the presence of Consumer 2.

[3] We obtain this result by differentiating Equation 10B.7 with respect to p_2, setting the result equal to 0, and noting that $\partial S_2(p_2)/\partial p_2 = -q_2(p_2)$. This result of no marginal distortion for Consumer 2 is analogous to the results concerning optimal taxation in Mirrlees (1971).

CHAPTER 11

Strategic Behavior

Do unto others before they do unto you.

This chapter analyzes actions taken by firms to reduce competition by actual and potential rivals. These actions are loosely called *strategic behavior,* and they can be more complicated than simply setting prices or quantities. For example, the first firm in a market could build a gigantic plant so as to leave little room for potential rivals to enter.

This chapter first defines strategic behavior and then examines both noncooperative and cooperative strategic behavior. We explore the differences between cooperative and noncooperative strategic behavior and discuss the legal treatment of strategic behavior under U.S. antitrust laws.

The key questions examined in this chapter are

1. Under what conditions does a firm benefit from using noncooperative strategic behavior?
2. When do oligopolists benefit from using cooperative strategic behavior?
3. Should antitrust laws forbid all actions that appear to be noncooperative or cooperative strategic behavior?

Strategic Behavior Defined

Strategic behavior is a set of actions a firm takes to influence the market environment so as to increase its profits. The market environment comprises all factors that influence the market outcome (prices, quantities, profits, welfare), including the beliefs of customers and of rivals, the number of actual and potential rivals, the production technology of each firm, and the costs or speed with which a rival can enter the market.[1] By manipulating the market environment, a firm may be able to increase its profits. As in the theory of oligopoly, the equilibrium in models of strategic behavior crucially depends on what one rival believes another rival will do in a particular situation. This chapter describes how a firm can influence the conditions of rivalry and thereby affect the outcome of the rivalry.

We examine two types of strategic behavior: noncooperative and cooperative. Although the distinction between noncooperative and cooperative behavior is not sharp, for expositional purposes it is helpful to consider them separately. **Noncooperative strategic behavior** encompasses the actions of a firm that is trying to maximize its profits by improving its position relative to its rivals. Noncooperative strategic behavior generally improves the profits of one firm and lowers the profits of competing firms. **Cooperative strategic behavior** comprises those actions that make it easier for firms in a market to coordinate their actions and to limit their competitive responses.[2] Cooperative strategic behavior raises the profits of all firms in a market by reducing competition.

A firm's blowing up its rival's store is an example of noncooperative strategic behavior. On the other hand, a scenario in which two rivals who are distrustful of each other sit down in a room to work out a price-fixing agreement is an example of cooperative behavior. Their subsequent behavior (for example, attempts to cheat on the price-fixing agreement) may be noncooperative.

Antitrust laws, which attempt to limit the undesirable acquisition of market power, are used to attack certain types of strategic behavior. The first and perhaps most important U.S. antitrust law, the Sherman Act, was passed in 1890. Section I of the Sherman Act prohibits all contracts, combinations, and conspiracies in restraint of trade. Section I is used to attack explicit cooperative behavior, such as a price-fixing agreement. Section II of the Sherman Act prohibits attempts to monopolize. Section II has been used to attack noncooperative strategic behavior, such as pricing below cost to drive rivals out of business.

Noncooperative Strategic Behavior

> *All business sagacity reduces itself in the last analysis to a judicious use of sabotage.*
> *—Thorstein Veblen*

A firm engages in noncooperative strategic behavior to harm its rivals and thereby benefit itself. Firms use many techniques to prevent rivals from entering a market, to drive rivals out of a market, or to reduce the size of a rival. Some of these strategies are designed to allow a firm to scare off potential rivals by changing rivals' beliefs about how

[1] The term *market environment* is slightly more inclusive than the term *market structure* because the latter, as commonly used, does not include beliefs of market participants.

[2] The term *cooperative* does not necessarily imply that the firms have an *explicit* agreement to undertake the behavior.

aggressively the firm will behave in the future. Two conditions must be met for a noncooperative strategy to be successful:

1. *Advantage:* The firm must typically have an advantage over the rivals. For example, the firm may be able to act before its rivals. That is, a firm must be able to do unto its rivals before they can do unto it.
2. *Commitment:* The firm must demonstrate that it will follow its strategy regardless of the actions of its rival.

If two firms are identical, both firms are in an equal position to threaten each other. For a strategy to work, then, one firm must have an advantage that allows it to harm the other firm before that firm can retaliate. Asymmetry between firms allows one firm to make a commitment that makes its threatened behavior believable.

For a firm's strategic behavior to work, its rivals must believe that the firm will remain committed to its strategy for as long as necessary. For example, an incumbent firm may announce that it will do something drastic (such as produce large quantities of output, a tactic that drives price down) if another firm enters its market. Talk is cheap, however, so the rival does not believe the incumbent's claim unless it is rational for the incumbent to follow this strategy after entry occurs.[3] For the incumbent firm's claim to be a **credible threat**, its rivals must believe that its strategy is rational in the sense that it is in the firm's best interest to continue to employ it. By making a commitment that does not allow it to change its strategy even if it wants to later, a firm can make its threat credible.

This section begins by analyzing four well-known strategies: predatory pricing, limit pricing, investment to lower costs, and raising rivals' costs. These strategies may work where barriers to quick entry and exit prevent another identical firm from using the same strategies. Without these barriers there can be no asymmetry among firms, and these strategic behaviors do not work. Next, the text examines why incumbents may have a natural advantage over later entrants. The section concludes with a discussion of antitrust policy toward such behavior.

Predatory Pricing

A dead man can't bite. —*Plutarch*

A firm engages in **predatory pricing** by first lowering its price in order to drive rivals out of business and scare off potential entrants, and then raising its price when its rivals exit the market. In most definitions, the firm lowers price below some measure of cost (legal definitions are discussed below). That is, the firm incurs short-run losses to obtain long-run gains.

What does a firm have to do to drive its rivals out of business? It has to convince its rivals that it is willing to drive price below their costs and keep it there until they leave

[3] Talk may be cheap, but it can be effective in allowing one rival to communicate a complicated strategy to another rival and in making it possible for a firm to develop a reputation for telling the truth. See Farrell (1987).

the market. This strategy is likely to be successful only if the firm can survive low prices longer than its rivals can. In many cases, however, the firm has no ability to convince its rivals that it is willing to maintain low prices for as long as it takes to drive them out of business.

If the firm succeeds in driving out its current rivals and then raises its price, new rivals may enter the market, and the incumbent must again lower its price to drive out those firms. For the predation to be successful, potential entrants must believe that it does not pay to enter this business because of the incumbent's pricing behavior. Only then can the incumbent raise its price to the monopoly level with no fear of inducing entry.

If the predator succeeds in forcing its rivals into bankruptcy, it should try to gain control of their assets or see that they are permanently withdrawn from the market. Otherwise, when the incumbent raises its price, a rival could again use those assets or another firm could buy the assets and compete. Even if a rival's assets are purchased by a firm in another market, they could always be redeployed to compete against the predator.

This discussion of predatory pricing begins by examining a model of identical firms in which predation is unlikely to be successful. Next, we consider a model in which one firm has an advantage over its rivals so that predation may be a profitable policy. Then, we look at how courts identify predatory pricing, and conclude with a review of the empirical evidence.

Predation with Identical Firms. Does the model of predatory behavior make sense if firms are identical? During the period of predation, the predating firm loses much more money than an equally efficient rival. The predatory firm must meet all demands at the low price in order to maintain the low price, but the rival is free to reduce its output in order to minimize its losses. As a result, the predation is unlikely to succeed.

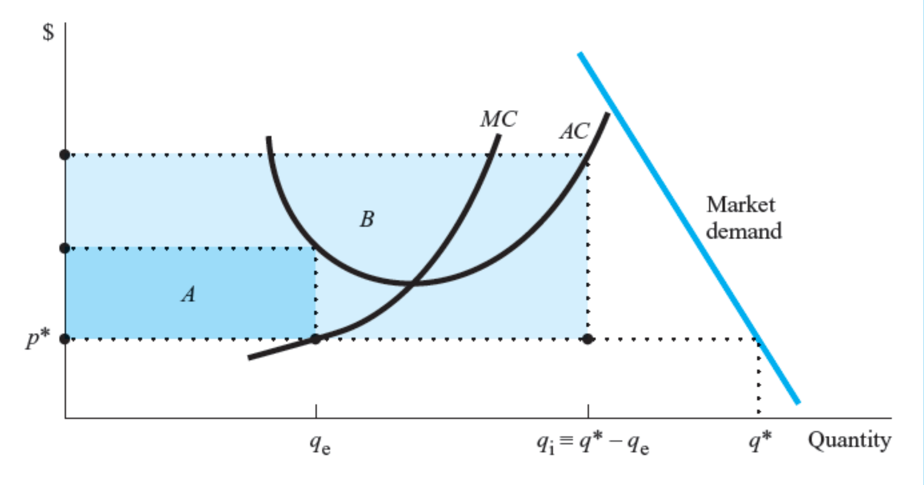
To illustrate this result, suppose that there are only two firms, an incumbent and a recent entrant, in the market with identical cost functions, as shown in Figure 11.1. The incumbent firm lowers the market price to p^* so as to inflict losses on its rival and drive it out of business. For the market price to be p^*, q^* units of output must be sold, as shown by the market demand curve in Figure 11.1.

If the rival does not exit the market, it produces q_e units, where p^* equals its marginal cost, and suffers a loss equal to area A in the figure. To keep the price at p^*, the incumbent must produce $q_i \equiv q^* - q_e$ units so that total market output is q^*. Thus, the incumbent produces at a higher marginal and average cost than its rival and suffers losses equal to area A plus area B. As a result, the incumbent's loss is greater than that of its rival by an amount equal to area B.

Consumers gain during the period of predation because they are able to purchase the product at price p^*, which is less than the duopoly price. If the predation is successful, consumers lose after the rival is driven out of business because the price rises to the monopoly level (which is greater than the duopoly price).

The major problem with this story of predatory pricing, when firms have identical cost functions, is that it is just as reasonable to suppose that the entrant can threaten the incumbent as to suppose the reverse. With no differences between the firms, why should any firm believe that another firm is willing to suffer losses greater than those of its rival for as long as necessary to drive the rival from the market?

FIGURE 11.1 Predation



For predation to work, the rival must believe that a firm will keep price low for as long as it takes to drive the rival out of business. Because the incumbent's proposed action is not viewed as rational by its rival, it is not viewed as a credible threat. See Example 11.1.

If an entrant did fear that its entry would precipitate a price war and drive prices below cost, it could avoid this problem in several ways. First, it could try to talk the in-

EXAMPLE 11.1

Supreme Court Says Alleged Predation Must Be Credible

In 1986, the Supreme Court reached an important decision regarding predatory pricing in *Matsushita Electric Industrial Co., Ltd. v. Zenith Radio Corporation et al.*, 106 S. Ct. 1348 (1986). A group of U.S. manufacturers claimed that certain Japanese firms had conspired for 20 years to sell consumer electronic products in the United States at prices below cost in an effort to drive the U.S. producers out of business.

The Supreme Court concluded that it was unlikely that any firm or group of firms would willingly inflict losses upon itself for 20 years in order to drive firms out of business eventually. Only a firm that made a very poor calculation of the discounted present value of the costs and benefits of the predatory strategy would predate for so long. The Supreme Court ruled that predation was not the reason for the low prices and that other explanations, such as legitimate competition, better described the reasons why the Japanese were able to price below American producers.

cumbent into merging, thereby enabling itself to charge a high price immediately and avoid the costly period of predation. U.S. antitrust laws, however, prohibit mergers to monopolize as well as predatory pricing (see Chapter 19).

A second approach is for the entrant to obtain contracts with buyers to set the price in advance of entry. A drop in the incumbent's price would not hurt it because its sales would be at the prespecified price. Buyers should be willing to sign fixed-price contracts at prices below the monopoly price that the incumbent initially charges.[4] Of course, it is not always possible to line up enough customers in advance on fixed-price contracts, especially when each customer is small. However, when there are large customers who realize that the entrant will prevent the incumbent from exercising market power, the entrant should have an easier time signing up customers in advance.

A third approach, as mentioned above, is for the rival to reduce its output during periods of predation to minimize the harm. In some markets, a rival can exit a market costlessly and redeploy its assets to another market during a period of predation. When the incumbent raises its price, the rival reenters the market. Exiting and entering can be repeated as long as necessary so that the predation can never inflict significant losses on the rival.

For example, suppose that the incumbent produces desks. The rival enters the market, and in response to its entry, the incumbent lowers the price of desks below cost. Suppose that the rival can quickly and profitably switch its factory to make tables instead of desks. As long as it is relatively inexpensive for the rival to switch between the manufacture of desks and the manufacture of tables, the incumbent cannot drive the rival out of business or credibly threaten to do so.

The rival can easily change industries if it does not have large *sunk costs* (costs that cannot be recovered once a business has been entered—see Chapter 2). Thus, if the rival has minimal sunk costs, an incumbent can have no hope in succeeding with predatory strategies because it has no way to impose costs on its rival. That is, in perfectly *contestable markets* (those in which instantaneous entry and exit at no cost disadvantage are possible), predation can never succeed.

Thus, the rival can avoid or mitigate the harms of predation in at least three ways. It can merge with the incumbent, sign long-term contracts in advance of the predation, or reduce its output during the period of predation.

Predation Where One Firm Has an Advantage. The reason that predation is unlikely to succeed where firms have identical costs is that the predating firm suffers greater losses than its intended victims. Thus, for successful predation to occur, the predating firm needs an inherent advantage over its rivals.

Not all differences between the predating firm and others, however, lead to successful predation. Many early studies of predatory pricing described the incumbent as a large firm and its rival as a small firm and argued that large firms can afford losses during predatory periods better than small firms. This assumption is questionable: Why

[4] Conversely, an incumbent facing the threat of entry may sign long-term contracts with buyers that limit entry of some lower-cost firms (Aghion and Bolton 1987).

wouldn't somebody lend to a small firm if it is not believable that the large firm will continue to incur losses forever?

Moreover, such a theory does not explain why other large firms fail to enter. For example, if small firms are at a disadvantage in competing with large firms, competition among large firms will ultimately dominate the economy. Therefore, predation should not necessarily lead to monopoly profits even when small firms are ineffective competitors.

More recent models of predatory behavior explain that differences in firms' beliefs about their rivals can result in successful predation.[5] For example, suppose that a firm can be either a high-cost firm or a low-cost firm and that only the firm knows its own costs with certainty. In response to entry, an incumbent firm may lower its price for one of two reasons. First, if the incumbent is a low-cost firm, the price decline might simply represent vigorous price competition that is profitable for the low-cost incumbent firm to pursue. Even if its new price is below the entrant's cost, it may be above the incumbent's cost. Second, if the incumbent is a high-cost firm, it may engage in predatory pricing.

The difference between this model and the previous models of predation is that it provides a possible explanation (that the firm has low costs) of why it is profit maximizing for an incumbent firm to drop its price in response to entry. The other firm, after observing the incumbent's pricing behavior, infers whether the incumbent firm is likely to have low or high costs. The lower its cost, the more likely the incumbent firm is to meet entry with very low prices.

As a result, an incumbent can acquire a reputation of being a low-cost firm by responding to entry with very low prices. Its pricing history is used by other potential entrants as an indicator, albeit not a perfect indicator, as to whether the incumbent firm has low or high costs. Because its pricing history is only a rough indicator, a high-cost firm might be able to price predate and convince potential entrants that it is really a low-cost firm. Of course, pricing histories can be used as an indicator of a firm's costs only if high-cost firms use low prices less frequently than do low-cost firms.

An entrant with no associated pricing history cannot influence the incumbent's beliefs about its costs, so there is a natural asymmetry between the firms. Because the entrant has no prior history whereas the incumbent has a history, the incumbent's beliefs about the entrant may differ from the entrant's beliefs about the incumbent. In this model, predatory pricing may be plausible. Pricing below cost for a high-cost firm turns out to be a rational strategy if it is able to create the illusion that it is a low-cost firm, and thereby deter entry.

Although these recent models show that it is possible to construct believable models of predatory pricing, it is still true that the practice is costly to an incumbent firm.[6] Moreover, the counterstrategy in which entrants contract with customers at a fixed price in advance may preclude successful predation.

[5] See, for example, Williamson (1977), Selten (1978), Ordover and Willig (1981), Easterbrook (1981), Kreps et al. (1982), Kreps and Wilson (1982a), and Milgrom and Roberts (1982b).

[6] See **www.aw-bc.com/carlton_perloff** "Spatial Predation" for a discussion of models where location is used to preempt rivals.

Finally, an entrant may have a reputation. For example, a firm's reputation in one market may carry over to any new market it enters. If so, there may be little asymmetry between the incumbent and the entrant, and hence little hope of successful predation.

Legal Standards of Predation. An extensive economic and legal literature suggests several standards for determining whether a firm is practicing predatory pricing. Many courts have adopted a rule proposed by Areeda and Turner (1975): A firm's pricing is predatory if its price is less than its short-run marginal cost. The logic behind this test is that no firm ever profitably chooses to operate where price is less than short-run marginal cost unless it is motivated by strategic concerns.[7] One possibility, if price is below short-run marginal cost, is that the firm is trying to drive rival firms out of business in order eventually to maximize profits. Pricing below short-run marginal costs would not make sense without some prospect of benefits in the future.

Areeda and Turner further suggest using average variable cost as a proxy for short-run marginal cost if data limitations prevent the determination of short-run marginal costs.[8] A strength of the Areeda-Turner rule is that it explicitly recognizes that pricing below average total cost is not, by itself, proof of predatory behavior. Indeed, price often is below average total cost in competitive industries such as agriculture due to short-run demand or supply fluctuations. (See Example 11.2 for a case in which the pricing-below-average-cost rule was used.)

Many economists and lawyers have responded to the article by Areeda and Turner. Some authors have suggested the use of long-run marginal cost, others have argued for the use of average cost, and still others have advocated observing price patterns over time or the amount produced over time to determine whether predation is actually occurring.[9]

Unfortunately, most of the suggested tests for predation can be difficult to implement, for two reasons. First, the data needed to determine short-run marginal production costs or even average variable production costs are often difficult to obtain. Second, other factors having nothing to do with price predation may explain violations of the tests.

It is common for a firm, upon entering a market, to attract consumer attention by running price promotions. During the start-up phase of a business, many firms give away their products as samples. Giving away a product may be a very effective promotional device to build business for the future, and it reflects rational, profit-maximizing behavior. This behavior appears to violate the Areeda and Turner rule and most other predation tests.

[7] In a one-period model, a profit-maximizing firm sets marginal revenue equal to marginal cost so that price is greater than or equal to marginal cost. If a firm is maximizing profits over time, however, it may operate at a price below short-run marginal cost, as is discussed below, even without a predatory strategy.

[8] In a multiproduct setting, the courts must use definitions that allow tests for predation involving only some of the many products that a firm produces. One standard, for example, could be that the price of one product could not be less than the average incremental cost of the product: the change in total cost from producing q units of a product divided by q, holding output of all other products at some prespecified level (Appendix 2A). This standard was used in *MCI Communication Corp. v. AT&T,* 708 F.2d 1081 [7th Circuit], cert. denied, 486 U.S. 891 [1983].

[9] Easterbrook (1981) and Posner (2001) discuss several of these alternative tests.

EXAMPLE 11.2

Evidence of Predatory Pricing in Tobacco

A firm may practice predatory pricing to force its rivals to sell their firms to it at a low price. By so doing, a firm can acquire its rivals cheaply and gain market power.

The Tobacco Trust allegedly engaged in predatory pricing against its rivals around the turn of the century. During the period 1881–1906, the Tobacco Trust acquired over 40 rivals and gained control of large shares of plug tobacco, smoking tobacco, snuff, and fine-cut tobacco sales. Frequently, the Tobacco Trust would identify a rival that it wished to buy and then introduce a competitive brand at a low price. The low profits would induce rivals to sell out to the Tobacco Trust at a low price.

For example, in 1901, the Tobacco Trust's American Beauty brand of cigarettes in North Carolina competed with a similar product of the Wells-Whitehead Tobacco Company of Winston, North Carolina. The American Beauty price was $1.50 per thousand, which was exactly equal to the required tax; it was therefore definitely below production costs. The Tobacco Trust claimed that the low price was an introductory offer. In 1903, the Tobacco Trust purchased its rival.

A detailed analysis of the value paid for the rivals purchased by the Tobacco Trust between 1881 and 1906 shows that predatory pricing had a large negative effect on the purchase price paid. Predation lowered the acquisition costs by about 25 percent.

As a result of violations of the antitrust laws, the Tobacco Trust was ordered dissolved and was broken into several separate firms (primarily American Tobacco, Liggett and Myers, and Lorillard) in 1911. By the 1920s, three firms dominated the cigarette industry, Reynolds (Camel brand), Liggett and Myers (Chesterfield brand), and American Tobacco (Lucky Strike brand).

In a famous antitrust suit (*American Tobacco Co. v. United States,* 328 U.S. 781 [1946]), these three firms were charged with explicit collusion to charge low prices to drive rivals out of business. During the Great Depression, the three major cigarette manufacturers increased their prices despite declining costs. New firms entered in response to this profit opportunity and sold, for roughly 5¢ less than the brands of the three major manufacturers, what were called *10-cent brands.* Between 1931 and 1932, the new brands increased their market share from below 1 percent to 23 percent.

In early 1933, the three major manufacturers dropped their wholesale prices by about 20 percent so that their retail prices were only slightly higher than those of the 10-cent brands, whose share fell to around 6 percent. The evidence suggests that these prices were not below the average total costs of the major cigarette companies. The 10-cent brands maintained a significant market share until the 1940s, when they disappeared. Even though price exceeded average cost and the 10-cent brands survived, the Court found that the companies had violated the antitrust laws by conspiring to lower prices with the intent to drive the 10-cent brands out of business.

Sources: Burns (1986), Tenant (1950, 43), Koller (1971), and *American Tobacco Co. v. United States,* 328 U.S. 781 (1946).

For most firms, a price of zero is lower than short-run marginal cost. A reasonable alternative view, however, is that the price of zero is a short-run promotional activity and is an investment designed to attract future customers. The price after the promotional period should be above the appropriate marginal cost measure where the price cut is treated as a promotional activity or cost. Just as investments in plant and equipment would not be expensed but would instead be amortized over time, so too should price promotions. Unfortunately, making such calculations can be difficult.

Similarly, a profit-maximizing firm may provide a product at a loss in the short run so as to signal the market that it will provide that product in the future. A firm may be concerned that potential customers may buy a rival's product and be unwilling to switch later, when it can produce cost-effectively. Pittman (1984) argues that IBM should have expected to make losses when it introduced its supercomputer but did not introduce the machine to engage in predatory pricing. Rather, IBM was signaling potential customers that it would provide supercomputers then and in the future.

Similarly, price can appear lower than short-run marginal cost when there is *learning by doing*: A firm's cost of production decreases as it produces more because it learns how to produce the product more efficiently. Because of this effect, a firm's costs are initially high but decline over time. By setting a very low price initially, the firm makes many sales and thereby accumulates experience that will enable it to lower its costs in the future. Even if the current price is lower than its current production costs, the prospect of reducing costs in the future by accumulating knowledge today justifies the lower price as an important investment for the firm. Again, the low price today should be viewed as an investment for the future. The short-run marginal cost of production that ignores future cost savings is not the relevant cost measure when a firm is involved in dynamic learning over time. Instead, one should look at the marginal production cost today plus (the present discounted value of) the change in production cost in the future that results from increased production today.

Most lawsuits alleging predatory pricing are brought by a firm against its rival. These rival firms may be complaining not about prices below cost but about price competition from a more efficient firm. If a firm is more efficient than another, one would expect the efficient firm to charge lower prices and take over the market. Indeed, the price could be below the inefficient firm's cost but equal to or above that of the efficient firm.

Therefore, predatory pricing suits could be a strategy by a less efficient firm to protect its market position. The evidence in a predatory pricing case of a lowering of price that inflicts losses on rivals is exactly what one expects when a more efficient firm competes in a market. If vigorous enforcement of predatory pricing laws prevents efficient firms from lowering their prices out of fear of a predatory pricing suit, it harms rather than helps consumers.

For this reason, Easterbrook (1981) suggests that the courts should not consider a predatory pricing suit until after a firm has been driven from business *and* the alleged predator has raised its price. Only then could one be sure that it was predation and not vigorous competition that drove the rival out of business.

Evidence on Predatory Pricing. Given all the theoretical difficulties with successful predatory pricing, it is not surprising that economists and lawyers have found few instances of successful price predation in which rivals are driven out of business and prices then rise. Although predation is frequently alleged in lawsuits, careful examination of these cases indicates that predation in the sense of pricing below cost usually did not occur.

For example, one of the most widely cited examples of price predation was the creation of Standard Oil. Supposedly, Rockefeller bought small, independent oil refineries after having lowered price to drive them out of business. McGee (1958), in his careful examination of this historical period, rejects that view and concludes that Rockefeller's rivals were bought out on rather favorable terms.

Koller (1971) reviews the available records in predatory pricing cases since 1890. Of the 26 cases for which adequate data existed, Koller finds evidence of below-cost pricing in 7 cases. Of these, only 4 represented successful predation, in that the rival vanished. Of these, 3 involved mergers.[10] A review of several predation cases illustrates that the evidence for predation in most cases is very weak and that defendants win over 90 percent of the time (Hurwitz et al. 1981). Isaac and Smith (1985) show that predation is rare in experimental settings.

The theory of predatory pricing relies on the incumbent's creation of a reputation for being a fierce competitor. The criticism of that theory is that it is unclear how such a reputation can be established and why rivals should believe it.[11] However, any theory that rests on a postulated set of beliefs cannot be logically *proven* wrong. Therefore, it is a mistake to think of price predation as inconceivable (see Example 11.2 and Weiman and Levin (1994), Genesove and Mullin (1997), and Morton (1997)).

Limit Pricing

Anybody can win unless there happens to be a second entry. —*George Ade*

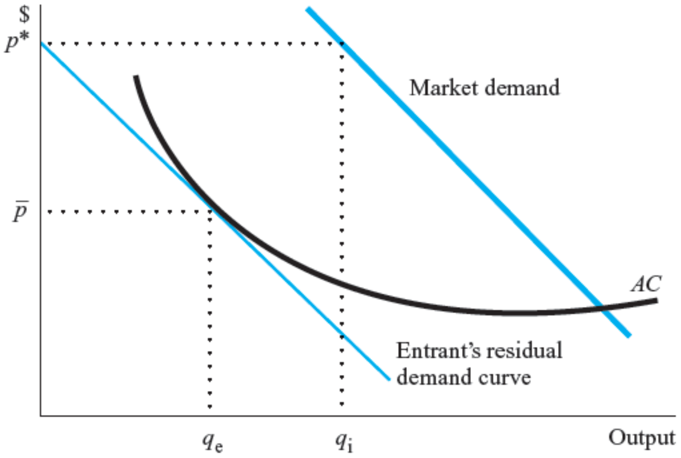
A firm is **limit pricing** if it sets its price and output so that there is not enough demand left for another firm to enter the market profitably. Early models of limit pricing were developed by Bain (1956), Modigliani (1958), and Sylos-Labini (1962). In the early limit-pricing models, the potential entrant believes that the incumbent firm will not change its output after the new firm enters. Therefore, a firm contemplating entry believes that total market output will equal its own output plus the current output of the incumbent. The extra output causes price to fall. In this model, the incumbent firm, given these beliefs by the potential entrant, chooses its output level and its associated price in such a way as to remove the incentive of a firm to enter.

Suppose that both the incumbent and a potential entrant have the same average cost, *AC*, curve (Figure 11.2). If the incumbent firm produces q_i units (and will continue to do so in the face of entry), then the demand curve facing an entrant equals the

[10] Mergers that lead to significant increases in market power are illegal under U.S. antitrust laws; thus, if successful predation requires a merger to exercise market power, there is no need for a law aimed at predatory pricing, because merger policy can be used to protect consumers.

[11] See Lott (1999), who argues that government enterprises, not private firms, are more likely to engage in predation.

FIGURE 11.2 Limit Pricing



market demand curve minus q_i. If the entrant believes that the incumbent will continue to produce q_i units of output, it believes its residual demand curve is the total demand curve minus q_i units.

If the potential entrant chooses not to enter, then the incumbent firm sells its q_i units for p^*, as Figure 11.2 shows. If instead the new firm enters the market and produces q_e units of output, then total market output equals $q_e + q_i$, and the market price is $\bar{p}$. Because of the incumbent's choice of q_i, $\bar{p}$. is just equal to the average cost for the potential entrant of producing q_e units, and the entrant is indifferent between entering and not (so presumably it does not enter).

If q_i is chosen so that the residual demand curve facing the potential entrant is just below (or equal to) its average cost curve, then the entrant cannot produce a quantity such that it earns a positive profit in this market. As Figure 11.2 shows, the incumbent can sell q_i at p^*, which is above its average cost of production, yet not induce entry. That is, the potential limit price $\bar{p}$ prevents entry. Indeed, the incumbent does not have to produce q_i to deter entry; it needs only to convince the potential entrant that it will produce q_i if entry occurs.

Limit Pricing with Identical Firms. The main problem with this model of strategic behavior is the same as in the model of predatory pricing: Why should an entrant with identical costs believe that the incumbent will carry out its threat to produce q_i units after entry occurs? It is not profit maximizing for the incumbent to continue to produce q_i in the event of entry. Thus, its threat to do so is not credible with identical firms.

Because both the incumbent and the potential entrant have identical costs, it is difficult to see how one firm could scare another based on assumed behavior after entry.

It is as plausible to believe that a potential entrant can scare an incumbent into exiting by threatening to enter and produce q_i as it is to assume that the incumbent can deter entry through limit pricing.

Further, as in the case of predatory pricing, a counterstrategy for the entrant would be to enter the market with existing fixed-price contracts already in hand. An entrant could induce customers to sign such contracts at a price slightly less than p^*, a price far above its minimum average cost.

Limit Pricing Where One Firm Has an Advantage. In order to make limit pricing believable and effective, an incumbent firm must pursue a strategy in which it is optimal for it to produce the q_i units at the limit price $\bar{p}$ after entry.[12] If the two firms have identical average cost curves, it is not believable that an incumbent would keep its output unchanged in the face of large-scale entry by another firm. The key to making limit pricing believable is for the incumbent firm to somehow manipulate the market environment when entry occurs so that the incumbent has the incentive to produce q_i units.

For example, suppose that in the first stage of a game between an incumbent and a potential entrant, the incumbent builds its plant. Only in the second stage can the potential entrant decide whether to build a plant so it can enter the market.

Further, suppose that the incumbent can construct its manufacturing facility so that it only can produce exactly q_i units.[13] Given such a plant, the potential entrant has no doubt that the incumbent will produce q_i units of output whether or not entry occurs. If the potential entrant knows that the incumbent has built such a plant, it will not enter. The incumbent has successfully practiced limit pricing: It has *committed* itself so that its threat to produce q_i units is believable.[14]

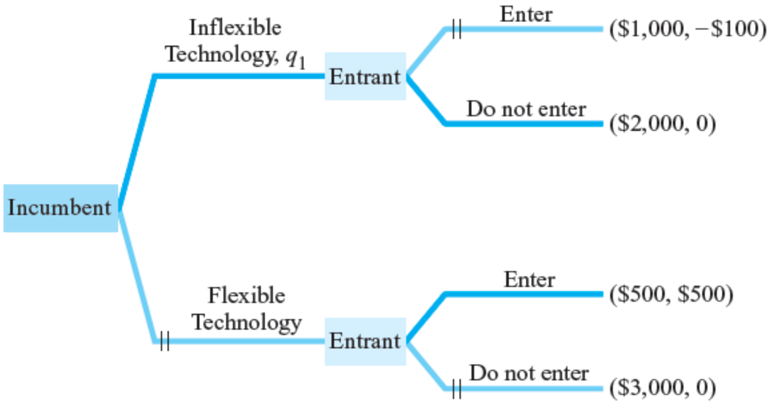
There is an inherent asymmetry in this model between the incumbent and the potential entrant. The incumbent chooses its investment first so that it can commit to produce q_i units of output whether or not entry occurs, whereas the entrant is not able to precommit to an output level before the incumbent acts. This fundamental asym-

[12] See, for example, Spence (1977a, 1979), Dixit (1979, 1980), Salop (1979b), Milgrom and Roberts (1982a), Fudenberg and Tirole (1983), Bulow, Geanakoplos, and Klemperer (1985b), Eaton and Ware (1987), Gilbert and Lieberman (1987), and Waldman (1987). Many of these papers stress the role of the incumbent's maintaining excess capacity. LeBlanc (1992) points out that a firm may choose between using limit pricing and predatory pricing. The stronger (relative to the entrant) incumbent is more likely to choose predatory pricing; and the weaker one, limit pricing. For intermediate cases, a combination of the two methods may be used. For an empirical analysis of entry deterrence, see Geroski (1991). For an analysis of limit pricing in an oligopoly setting, see Bagwell and Ramey (1991) and Martin (1995).

[13] More reasonably, it might build a large plant with very low marginal costs at q_i.

[14] Committing to a fixed capacity in the future becomes more difficult if capital depreciates over time because the incumbent may be unable to maintain the natural asymmetry that arises from moving first. With rapid depreciation, the incumbent's advantage erodes rapidly. Both the incumbent and the new entrant may be on equal footing in terms of their ability to precommit to replace capacity as it wears out.

FIGURE 11.3 Extensive-Form Representation of Limit Pricing Game



Note: The incumbent's profit is the first in the pair of payoffs.

metry is exploited by the incumbent to make its strategic behavior believable.[15] The incumbent manipulates the underlying environment (its ability to produce) in such a way as to give itself an advantage over the potential entrant.

The incumbent spends money in the first stage of this game to limit its productive options. That is, without such an investment, the incumbent could produce a wide range of output instead of only q_i units. At first glance, the incumbent is *purposefully shooting itself in the foot* by reducing its production options. Rather than only harming itself, however, the incumbent benefits from this restriction. The restriction makes believable the incumbent's threat to produce q_i units in the face of entry, so the potential entrant does not enter.

In general, a firm can benefit if it can precommit (limit its future options). By making commitments that render its threats credible, a firm raises its profit even though it restricts its future options.

Figure 11.3 illustrates this example of limit pricing using an *extensive-form representation* of the game (Chapter 6) that shows the sequence of all the possible actions and outcomes for both firms. Each line represents an action, and each box represents a decision point. The outcomes of actions are shown in parentheses, where the incumbent's profits are listed first. In the first stage of the game, the incumbent chooses between two production technologies: a flexible one that allows a wide range of output to be produced, and an inflexible one that can produce only q_i, where q_i is an output level that deters entry. In the second stage, the potential entrant decides whether to enter.

[15] Another way the incumbent might make a commitment is to sign a contract that penalizes it if it fails to produce q_i. However, because such contracts usually are not legally enforceable, they do not make the incumbent's threat credible.

To determine its optimal strategy in the first stage, the incumbent solves the game backward starting from the top right of the diagram. The potential entrant has to decide whether to enter. If the incumbent chose the inflexible technology and must produce q_i, the potential entrant earns \$0 if it does not enter and loses \$100 if it does enter. Thus, the potential entrant chooses not to enter. The two lines across the *Enter* line of action show that this strategy is ruled out.

Next the incumbent considers the bottom right corner of Figure 11.3. If the incumbent chose the flexible technology, the potential entrant earns a profit of \$500 if it enters and \$0 if it does not. Therefore, two lines block the action *Do not enter.* By this reasoning, the incumbent infers how the entrant would behave conditional on the incumbent's decision in the first stage of this game.

As a result, in the first stage, the incumbent decides whether to produce with the flexible or inflexible technology. If the incumbent chooses the inflexible technology, the entrant does not enter, so the incumbent earns a profit of \$2,000. If the incumbent chooses the flexible technology, the potential entrant enters, so that the incumbent earns a profit of only \$500. Thus, choosing the inflexible technology is more profitable.[16] Hence, the two lines across the *flexible technology* option block that line of action. The solution of choosing the inflexible technology is *subgame perfect* (see Chapter 6) because the threat to produce q_i is credible.[17]

Even if the incumbent chooses the flexible technology, it can still threaten a potential entrant that it will produce q_i if it enters. In so doing, the incumbent is trying to have it both ways: deterring entry through its threat and having a flexible technology if entry does not occur. Unfortunately for the incumbent, its threat is not credible without the commitment. The potential entrant knows that if it enters, the incumbent can make more money acting like a duopolist and reducing output below q_i than it can producing q_i. That is, the threatened post-entry strategy of producing q_i is not subgame perfect. Thus, the only way the incumbent can prevent entry is by committing to the inflexible strategy.

Dynamic Limit Pricing. If a firm sets prices (or quantities) over time so as to reduce or eliminate the incentives of rivals to enter a market, it is practicing **dynamic limit pricing**.[18]

Although a dominant firm may be able to set an extremely high price and maintain it in the short run, it may choose not to do so. A very high price attracts additional fringe firms, causing the market price to fall. Conversely, if a dominant firm

[16] If there were no threat of entry, the incumbent would prefer the flexible technology because its profit is higher (\$3,000) than with the inflexible technology (\$2,000). However, as the example shows, the inflexible technology is better for the incumbent when entry can occur.

[17] A Nash equilibrium in strategies is subgame perfect if the original strategies are Nash equilibria (best responses) in any subgame (a new game that starts in any period *t* and lasts to the end of the game). That is, no player wants to change strategies in a later period.

[18] See **www.aw-bc.com/carlton_perloff** "Dynamic Limit Pricing," Judd and Petersen (1986), Kamien and Schwartz (1971), De Bondt (1976), Gaskins (1971), Baron (1973), Stigler (1965), and Berck and Perloff (1988, 1990).

keeps its price very low to prevent entry, it has very low profits in both the short and the long run. Thus, a dominant firm that faces the threat of entry must trade off high profits in the short run against the entry of more competition and lower profits in the future.

It is often in the dominant firm's best interest to set a high price at first and then slowly to lower the price as entry occurs. Although the high price increases the rate of entry, profits today are worth more to the dominant firm than are profits in the future (given positive interest rates).

Because of this pricing behavior, it is common for dominant firms to lose market share over time (Example 11.3). When U.S. Steel was created in 1901, its share of the steel ingot market was thought to be 66 percent, but by 1982, its share had fallen to 19 percent.

EXAMPLE 11.3 *The Shrinking Share of Dominant Firms*

Generally, a dominant firm's share of an industry's sales shrinks over time. Consider 13 unregulated major industries in which firms compete on a national or international basis. Using the *Fortune 500* rankings to determine the leading firm in each industry, Pascale (1984) traced these firms' shares of industry sales over a 20-year period:

Industry Share Trends in 13 Key Industries

		Industry Share		
Leading Firm	Industry	1962	1982	Percent Change
Sears	Mass-market retailing	5	5	0
International Harvester	Farm tractors	24	18	−25
U.S. Steel	Finished steel	26	19	−27
Goodyear	OEM tires	29	27	−7
General Electric	Electrical appliances (refrigerators)	40	53	+33
RCA	Color TVs	49	20	−59
Boeing	Commercial wide-body jet aircraft	51	60	+18
General Motors	Passenger cars	52	46	−12
General Electric	Generators	59	61	+3
IBM	Mainframe computers	60	68	+13
Kodak	Photographic film	85	65	−24
Harley-Davidson	Motorcycles	100	36	−64
Xerox	Plain copiers	100	42	−58

Of these 13 leading firms, eight lost share, and one firm's share remained unchanged over the period. Three firms lost over half of their share of industry sales (RCA, Xerox, and Harley-Davidson), including both firms that made essentially all their industry sales in 1962.

EXAMPLE 11.4 *And Only a Smile Remained*

> *[The Cheshire Cat] vanished quite slowly, beginning with the end of the tail, and ending with the grin, which remained some time after the rest of it had gone.* —*Lewis Carroll*

Laszlo Jozsef Biro took out a patent on a ball-point pen in Paris in 1939. During World War II, he moved to Argentina, where his company, Eterpen S.A., started producing and selling the pens in 1943. Unlike a conventional fountain pen, it had a miniature socket that held a ball bearing, it used a special ink that dried almost instantly, and it held enough of this unconventional ink to work for months without refilling.

Also unlike a fountain pen, this pen could work at high altitudes without the risk of leakage. As a result, the U.S. Air Force was interested. It sent pens to various American manufacturers, saying it might want to buy ten thousand or so of them. The big three pen manufacturers—Parker, Sheaffer, and Eversharp—looked into the patent rights and discovered that Eberhard Faber, a pencil manufacturer, had obtained them but had run into difficulties producing the pens. Eversharp obtained the rights in 1945.

Eversharp redesigned the pen for mass production and instituted an advertising campaign to prepare the public for this new “miracle pen.” This advertising greatly benefitted Milton Reynolds, who ultimately beat Eversharp to the market.

Reynolds had seen the pen in South America. When he found that he was too late to buy the rights from Biro, he developed ways around Biro’s patent. What was unique about the Biro pen was its pressure-feed system that regulated the ink supply. Reynolds developed a different system that used gravity.

The Reynolds International Pen Company started production on October 6, 1945. A major New York department store, Gimbel’s, advertised extensively, claiming that the pen was guaranteed to write for two years without refilling, to write under water and at stratospheric altitudes, and to make a clear impression on six to eight carbons. These claims made Gimbel’s price of $12.50 (the maximum price allowed by the wartime Office of Price Administration) seem, if not a bargain, at least not the most staggering extravagance of all time.

With the initial cost of production around 80¢ per pen, healthy profits were realized when Gimbel’s sold 10,000 pens (worth about a third of the store’s average total daily sales volume) on the first day of sale, October 29, 1945. This success encouraged Reynolds to expand production; by early 1946, his 800 employees were producing 30,000 pens per day.

There are many examples of new industries in which a product is introduced at a high price that soon falls to a competitive level. When an industry is new, one or a few firms have large market shares and face relatively few competitors. Only over time do new entrants drive down the price. A particularly striking example from the early days of the ball-point pen industry is given in Example 11.4.

Production could not keep up with orders, and gift certificates were printed. By March 1946, Reynolds had banked $3 million. During one 10-day period, he deposited $1.5 million from orders of pens yet to be made. By February 1946, Reynolds had an after-tax profit of $1,558,607.81.

These enormous profits encouraged entry. Gimbel's rival department store, Macy's, sold the Biro pen for $19.98. It, too, did well, encouraging still more entry. Late in April, Eversharp finally entered the market with a $15 pen. The July 1946 *Fortune* magazine reported that Shaeffer was going to sell a pen at $15. Eversharp then announced plans to sell a retractable pen at $25.

Meanwhile, Reynolds introduced a new model with a retractable point protector that cost 60¢ per pen to produce but sold at the original price of $12.50. By late in the summer of 1946, his pens were being sold in 37 other countries (with prices in Hong Kong reaching $75). As profits remained high, still others entered.

The Ball Point Pen Company of Hollywood ignored a patent infringement suit and sold a $9.95 version. Another manufacturer, David Kahn, announced plans to sell a pen for less than $3. In October, Reynolds introduced a new pen that cost 30¢ to produce and sold for $3.85.

Approximately 100 manufacturers were producing pens by Christmas 1946, some selling for as little as $2.98. Reynolds again introduced a new model, priced at $1.69, but Gimbel's sold it for 88¢ in a price war with Macy's. At one point, Gimbel's changed prices five times during shopping hours. Reynolds then introduced a new, two-color model priced at 98¢ that was still highly profitable.

By mid-1948, some ball-point pens were selling for 39¢ and cost between 8¢ and 10¢ to produce. The price of some pens fell to 25¢ by 1951, and soon after, pens were available at 19¢. By this time, the large number of firms in the industry had driven the price down to the point where no unusual economic profits were being earned. Reynolds's market share went to zero and the firm stopped producing new pens in the United States.

This example shows that if a firm has no cost advantage or other advantage, it cannot maintain a large share of the market in the long run. Nonetheless, even a short-lived period of dominance can be highly lucrative. It is estimated that in a single month Reynolds earned profits as high as $500,000, or about 20 times his original investment of $26,000.

Sources: Lipsey and Steiner (1981) and Thomas Whiteside, "Where Are They Now?" *New Yorker,* February 17, 1951:39–58.

Investments to Lower Production Costs

In models of oligopoly behavior, the market outcome typically depends on the costs of the competing firms (Chapter 6). That is, the costs of each firm are part of the market environment that determines the outcome of the competition among the firms. In the following model, an incumbent firm manipulates the market environment to its advantage. We consider two examples where one of the firms has an advantage so that its strategy may be successful. In the first, one of the firms can engage in research and development (R&D) to lower its costs in a later period. In the second example, one firm lowers its cost through learning by doing.

Investing in R&D. Suppose that there are two time periods and two firms with identical initial cost functions. In Period 1, the incumbent firm is a monopoly and can invest in research and development (R&D) that will lower its costs in Period 2. In Period 2, the second firm may enter. The asymmetry in this model results from the assumption that only the incumbent firm, not the entrant, can invest in R&D to lower its costs. This asymmetry arises naturally when one firm is in a market before another firm.

Does the incumbent firm have an incentive to invest in R&D in order to lower its costs in Period 2? To illustrate the strategic choices of the incumbent firm, consider a specific example where the duopolists use Cournot strategies (Chapter 6) in Period 2.[19] In Period 1, the incumbent firm incurs $1 of fixed costs and has a constant marginal cost of $6. If the incumbent makes no investments in R&D in Period 1, then its fixed and marginal costs are the same in Period 2. The costs of the entrant in Period 2 are the same as the costs of the incumbent in Period 1. The linear market demand curve is $q = 12 - p$.

To decide whether to engage in R&D in Period 1, the incumbent needs to compare its profits in the equilibria with and without R&D. Table 11.1 shows the price and profits conditional on whether entry occurs in Period 2 and on whether the incumbent invests in R&D.

First, consider the equilibrium where the incumbent does not invest in R&D and the second firm enters in Period 2. In Period 1, the incumbent firm is a monopoly and equates marginal revenue based on the market demand curve to its marginal cost. It charges a price of $9 and produces 3 units, resulting in a profit of $8 in Period 1.[20] In Period 2, the incumbent and the entrant face the same cost conditions and play Cournot. In the equilibrium in Period 2, each firm produces 2 units at a price of $8 and earns a profit of $3.[21] Thus, the incumbent's total profit for the two periods is $11

[19] Although the quantitative results depend on the particular oligopoly behavior, the general insights from this model hold for all standard oligopoly models. The key feature of the example is that the incumbent's behavior in Period 1 influences the equilibrium in Period 2.

[20] The incumbent's profit in Period 1 is its total revenue minus its total cost: $\pi_i = q_i(12 - q_i) - (1 + 6q_i)$.The first-order condition for profit maximization is $12 - 2q_i - 6 = 0$, or $q_i = 3$. Consequently, p is $9 and π_i is $8.

[21] In Period 2, the incumbent maximizes its profit, $\pi_i = q_i[12 - q_i - q_e] - (1 + 6q_i)$. Consequently, the incumbent's best-response (reaction) function (see Chapter 6) is $q_i = 3 - q_e/2$. The entrant maximizes $\pi_e = q_e[12 - q_i - q_e] - (1 + 6q_e)$, so its best-response function is $q_e = 3 - q_i/2$. In the Cournot equilibrium (Chapter 6), which is determined by the intersection of these best-response functions, $q_i = q_e = 2$, $p = 8$, and $\pi_i = \pi_e = 3$.

TABLE 11.1 Strategic R&D Investment: Monopoly in Period 1, Cournot Competition in Period 2

	Period 1	Period 2	Total Profit in Periods 1 and 2
Entry			
No R&D investment	Profit of incumbent = $8	Profit of incumbent = $3	$11
		Profit of entrant = 3	3
	Price = 9	Price = 8	
R&D investment	Profit of incumbent = 8 − 7.01 = 0.99	Profit of incumbent = 10.11	11.10
		Profit of entrant = 0.77	0.77
	Price = 9	Price = 7.33	
No Entry			
No R&D investment	Profit of incumbent = 8	Profit of incumbent = 8	16
	Price = 9	Price = 9	
R&D investment	Profit of incumbent = 8 − 7.01 = 0.99	Profit of incumbent = 15	15.99
	Price = 9	Price = 8	

(assuming, for simplicity, that the discount rate is zero). These results are summarized at the top of Table 11.1.

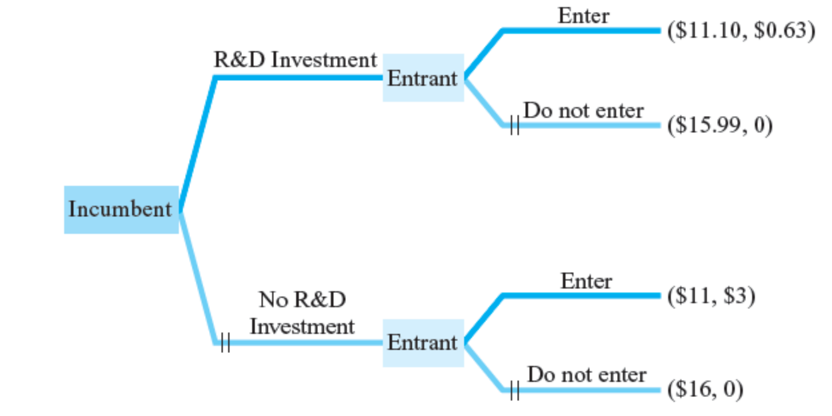
Now suppose that the incumbent firm invests in R&D in Period 1 and the rival enters in Period 2. For an investment of $7.01, the incumbent can lower its marginal cost next period by $2 (with certainty, for simplicity). If the incumbent makes the investment, it earns less money in Period 1. If it does not invest, it earns $8 in Period 1, whereas if the incumbent invests, it earns $8 − $7.01 = $0.99 in Period 1. The investment causes the incumbent's marginal cost to fall from $6 to $4 in Period 2. In the Cournot equilibrium in Period 2, the incumbent produces 3 1/3 units, the entrant produces 1 1/3 units, and the price equals $7.33.[22] The incumbent's profit in Period 2 is $10.11 (see Table 11.1).

Combined with its profit of $0.99 in Period 1, its total profit for the two periods is $11.10, which exceeds the $11 that it earns if it does not invest in R&D in Period 1. Thus, given entry in Period 2, the incumbent earns more by investing in R&D because its reduced earnings in Period 1 are more than offset by the increased earnings in Period 2. Moreover, consumers are better off when Firm 1 invests in R&D because the price in Period 2 is lower when R&D occurs.

Now, suppose that entry does not occur in Period 2. Would it still be profitable for the incumbent to invest in R&D? If the firm makes no investment in either period, it earns its monopoly profit of $8 in each of the two periods, for a total profit of $16. If the firm invests in R&D, its monopoly profit in Period 1 is $0.99 and its monopoly

[22] The incumbent chooses q_i to maximize its new Period 2 profit, $\pi_i = q_i(12 - q_i - q_e) - (1 + 4q_e)$. Its new best-response function is $q_i = 4 - q_e/2$. The entrant's best-response function remains $q_e = 3 - q_i/2$. The best-response functions intersect at $q_i = 3\,{}^{1}/_{3}$ and $q_e = 1\,{}^{1}/_{3}$.

FIGURE 11.4 Extensive-Form Representation of R&D Game



Note: The incumbent's profit is the first in the pair of payoffs.

profit in Period 2 is $15 (Table 11.1).[23] That is, it earns $7.01 less without the investment in Period 1 and earns $7 more in Period 2. Its total profit is $15.99, which is less than the $16 it would earn if it does not invest. Thus, it should not invest.

Figure 11.4 shows the extensive-form representation of this R&D game. The diagram illustrates that regardless of whether Firm 1 invests in R&D, its rival will enter the market (the double lines block the *Do not enter* action lines). As a result, it pays for the incumbent to invest in R&D: It earns $11.10 over the two periods instead of $11.

The incumbent makes an investment in Period 1 to alter the environment of Period 2 in its favor. This strategic behavior is attractive to the incumbent due to the asymmetry that allows it to act before the entrant. This example is similar to the Stackelberg model (Chapter 6) in that the incumbent benefits from being able to act before its rival and credibly committing itself to produce a relatively large output. In this example, the incumbent's ability to invest in R&D benefits both consumers and the incumbent. Without the threat of entry, the incumbent would not invest in R&D.

Learning by Doing. If the incumbent can reduce its costs in Period 2 through learning by doing in Period 1, it can gain an advantage over its rival that enters in Period 2. The incumbent has an incentive to sell more than it otherwise would in Period 1 to gain experience and lower its cost relative to that of its rival in Period 2. To increase its sales in Period 1, the incumbent must sell at a lower price than it otherwise would.

[23] If the incumbent invests in R&D, its monopoly profit in Period 2 is $q_i(12 - q_i) - (1 + 4q_i)$. Its first-order condition requires that it equate its marginal revenue to its marginal cost by setting q_i equal to 4. As a result, p is $8, and its profit is $15.

Thus, its profit in Period 1 is lower than it would be if it ignored the benefit of increased production on its costs in Period 2. Learning by doing, then, can be thought of as an investment that enables the firm to earn more in subsequent periods.

In a learning-by-doing model, the advantage of being able to go first depends on how much a firm can lower its cost relative to that of its rival and how long it takes to learn. If learning is either extremely rapid *or* extremely slow, the advantage of having a head start is not very great. When learning is very rapid, late entrants can quickly catch up with the incumbent. Conversely, when learning is very slow, the head start a firm gets does not matter very much. In the intermediate cases in which learning is neither very rapid nor very slow, the strategic importance of learning by doing is greatest for increasing profits (Spence 1981a). Indeed, if the learning-by-doing cost advantage is substantial enough, the second firm may choose not to enter the market.

Raising Rivals' Costs

A firm may benefit from strategic behavior that raises its rivals' costs.[24] In oligopoly models (Chapter 6), a firm's profit depends on its costs relative to those of its rivals. If a firm can costlessly raise its rivals' costs relative to its own, it can increase its profit at the expense of its rivals. In order to affect a rival's costs, usually a firm must have some market power or political power. This section examines a firm's strategies that raise its rivals' costs relative to its own and also those that raise everyone's costs. Then it discusses strategies that an entrant may use.

Raising the Rivals' Relative Costs. A firm clearly benefits if it can raise only its rivals' costs. Indeed, a firm may benefit from actions that raise its own costs if they raise its rivals' costs by more. The firm could use a direct method or one of several indirect methods.

Direct Methods: A firm may directly raise its rivals' costs if it can interfere with its rivals' production or selling methods. To take an extreme case, an unethical firm could blow up a rival's plant or sabotage a rival's machines. Both actions would raise its rival's costs, reduce competition, and raise the profit of the unethical firm practicing this strategic behavior (assuming that the firm is not caught). If the unethical firm must spend money to raise its rival's costs, then it must balance its increased expenditures for sabotage against its benefit from raising its rival's costs.

In 1993, British Airways (BA) admitted in court to playing dirty tricks on a smaller rival, Virgin Atlantic Airways (VAA) and is paying $2.5 million to settle a libel suit by the owner of VAA. BA's staff tapped into VAA's computers to obtain names and numbers of their passengers; phoned or met VAA passengers and falsely claimed that their flights were delayed or overbooked, and offered inducements to fly with BA; broke into homes and cars of VAA staff; hired a consultant to dig up dirt on VAA's owner and to plant negative news stories; and withdrew cooperation in maintenance and

[24] See Salop and Scheffman (1987), Krattenmaker and Salop (1986), Riordan and Salop (1995), and the papers in Salop (1981).

training.[25] No wonder Richard Branson, Chairman of VAA, once said that competing with BA was "like getting into a bleeding competition with a blood bank."[26]

The French government may have hidden microphones on Air France flights to Paris to gather information about American firms' marketing and technical plans.[27] The French Foreign Ministry portrays alleged spying as an essential way for France to keep abreast of international commerce and technology. The French may have won a billion-dollar contract to supply jet fighters to India by getting inside information on competing bids. The FBI also reported a French scheme to infiltrate foreign offices of IBM and Texas Instruments, perhaps to obtain information for the largely government-owned Compagnie des Machines Bull. Theft lowers a firm's costs relative to its rivals and is equivalent in its effects to raising a rival's relative costs.

Another example of a direct method is to make it difficult for a rival to gather information. For example, if an entrant conducts a marketing experiment to see whether its product is liked in certain locations, the incumbent can counteract the experiment by offering huge promotional discounts in those locations, making it more difficult for the entrant to judge consumer acceptance of its product relative to the incumbent's product (Fudenberg and Tirole 1986a).

Interference Through Government Regulation: A firm may raise its rivals' costs through government regulation. Many government regulations "grandfather" (exempt from regulation) existing firms and make it more onerous for new firms to operate in a market. For example, some environmental regulations impose more stringent requirements on new equipment than on old equipment and thus favor existing firms over entrants. By supporting government regulation so that a new rival cannot adopt their production techniques, incumbent firms can preserve and protect their market position and make it more costly for entrants to compete.

Tie-ins of Other Products: Sometimes an incumbent produces two products that must be used together, whereas the entrant produces only one of these products. Examples of products that complement each other are a camera and film or a computer and peripheral devices (printers, floppy drives, and so forth). Where products must be used together, the incumbent can disadvantage the entrant either through a contractual tie whereby the consumer must purchase both products together from the incumbent or through a product design decision that makes the entrant's product incompatible or difficult to use with the incumbent's other product. For example, a computer manufacturer could use a nonstandard plug to connect a printer. Even if a product's design reduces the amount consumers are willing to pay for it, the increased profits that come from hampering a rival may offset the loss (Farrell and Saloner 1986a,

[25] Paula Dwyer, "British Air: Not Cricket," *Business Week,* January 25, 1993:50–1; "Tactics and Dirty Tricks," *The Economist,* January 16, 1993:21–2.

[26] *London Times,* September 20, 1984.

[27] Larry Reibstein, Christopher Dickey, and Douglas Waller, "Parlez-Vous Espionage?" *Newsweek,* September 23, 1991:40.

Matutes and Regibeau 1988, Whinston 1990). See Example 11.5. Appendix 11A provides a detailed analysis of the strategic use of complements with applications to network industries (e.g., railroads, computers). Even where the products are independent, a tie-in can so reduce demand that rivals cannot efficiently produce (see Nalebuff forthcoming).

Raise Switching Costs: An incumbent can make it difficult for consumers of its product to switch to an entrant's product in the future (Schmalensee 1982; Klemperer 1987, 1990; Segal and Whinston 1996). That is, the incumbent may be able to raise the entrant's marketing costs to attract customers. For example, appropriate design may make it impossible to use computer programs written for one computer on another computer. Although the design may make the incumbent's product less desirable, it also serves to raise the switching costs to its consumers. As a result, a potential entrant faces a lower demand than it would otherwise, which reduces its incentive to enter.

Raising Wages or Other Input Prices: An incumbent firm that uses a different production technology than its rivals may be able to raise their costs disproportionately by raising the cost of an input to all firms in the market. If, for example, the rival uses more labor per unit output than does the incumbent firm, the incumbent's costs rise less from an increase in the wage rate than do the entrant's. Although total profits in the *market* must go down when wages rise, the market share of the less labor-intensive firm can increase by enough that its profits rise. This strategic behavior takes advantage of the natural asymmetry in production and assumes that the incumbent can influence market wages.

An incumbent may be able to increase wages by supporting union activities (Williamson 1968). For example, all the U.S. automobile manufacturers face a single union. Each time the union contract comes up for renewal, the union negotiates with (and strikes if necessary) one of these firms, and the others accept the outcome of these negotiations. A single firm, then, could negotiate an unusually high wage rate.

Similarly, an incumbent firm may be able to raise wages through direct market purchases. If the incumbent can purchase enough of the labor in a market to drive up the market wage, it has monopsony power. It can strategically use that market power to increase the costs of other firms more than its own if the other firms are more labor intensive.

To illustrate how raising a rival's costs can raise an incumbent firm's profits even if its own costs go up, consider the case of an incumbent that uses less labor per unit output than does a rival. Suppose that the incumbent has a constant marginal (and average) cost of m until its capacity, $\hat{q}_i$, is reached, whereupon its marginal cost is infinite, as Figure 11.5 shows. There are many rivals, a competitive fringe, all with the same constant marginal cost, m_1, as the figure illustrates.

In the absence of strategic behavior, the equilibrium price in the market is m_1, and it is optimal for the incumbent to produce at capacity, $\hat{q}_i$, where it earns profits equal to $(m_1 - m)\hat{q}_i$. Suppose now that the incumbent can raise the market wage rate. Because the incumbent's technology is different from everyone else's, the wage

EXAMPLE 11.5

Strategic Behavior and Rapid Technological Change: The Microsoft Case

In some markets, the product design of the primary product (such as a CD player) as well as that of complementary products (such as CDs) can change rapidly. In such settings, there is scope for strategic behavior in which a firm with market power in the primary product uses complementary products to increase its power in the primary market (see Appendix 11A). A recent government antitrust suit aimed at Microsoft attacks such behavior.

The government alleged in this case that Microsoft has market power in the market for personal computer operating systems, its primary product, because Microsoft's Windows operating system is the dominant system in use. The government charged that to preserve and enhance its power, Microsoft bundled its operating system together with its browser (software used to interact with the World Wide Web), a complementary product. The government then alleged that, as a consequence, the demand for Netscape's rival Internet browser declined substantially, and Netscape was forced to stop charging for its browser.

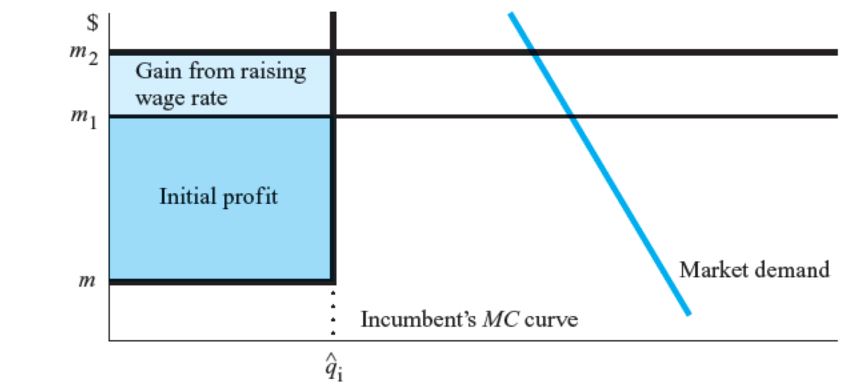
Although end users may benefit from obtaining free browsers, the government was concerned that potential competitors to Microsoft's Windows operating system might be driven out of business. Netscape's browser worked with all major operating systems, including Windows. Although Netscape's browser was designed initially to allow users to read Web pages, the government argued that Microsoft was concerned that Netscape's browser would evolve into an alternative "programming platform." According to the government, Microsoft executives worried that, in time, Netscape would add to its product "application programming interfaces" that software developers could use to write application programs. Such interfaces would pose a threat to Microsoft if they allowed application programs to run on all the operating systems that work with Netscape because that would erode the advantage that Windows had of having more application programs than other operating systems. The success of Netscape's browser would aid competition in operating systems, which might lead to a decline in Microsoft's dominance of operating systems. In its defense, Microsoft argued that the integration of its browser with Windows led to efficiencies that would be unattainable were the products sold separately and not integrated into a single program. The court ruled that Microsoft violated the antitrust laws.

Note: Carlton served as a consultant to Sun Microsystems, which sued Microsoft on related grounds.

Source: Carlton (2001), Carlton and Waldman (2002), and *U.S. v. Microsoft,* 253 F. 3d 34. See also Evans et al. (2000).

increase will have a different effect on m_1 than on the incumbent's marginal cost. Consider an extreme example in which the marginal cost of the incumbent firm does not change, and the marginal cost of rivals increases from m_1 to m_2. The equilibrium price rises from m_1 to m_2, the incumbent's optimal level of production is

FIGURE 11.5 Raising a Rival's Costs



unchanged at $\hat{q}_i$, and its profit increases to $(m_2 - m)\hat{q}_i$, as Figure 11.5 shows. By behaving strategically and raising the costs of the competitive fringe from m_1 to m_2, the incumbent is able to increase its profit. Moreover, even if the wage increase raises the incumbent's costs, it may still be profitable for the incumbent to raise wages as long as the cost increase is smaller than the gain in Figure 11.5.

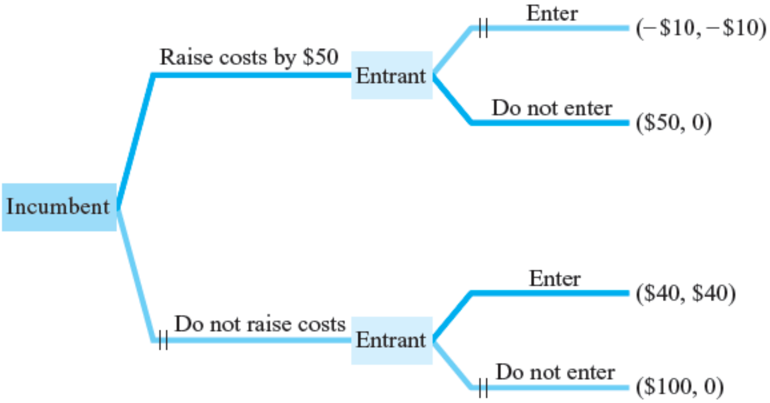
Another way to raise rivals' costs is to make distribution of the product more costly. If an incumbent can control most of the distributors (such as wholesalers or retailers), it can raise the cost of their rivals (Ordover, Saloner, and Salop 1990, Salop and Riordan 1995).

Raising All Firms' Costs. It may pay for the incumbent to raise the costs of all firms. A natural asymmetry often exists between an incumbent and potential entrants, in that the incumbent has already made expenditures (that is, sunk costs) that make it unlikely that it would exit the market.[28] Having made these expenditures before anyone else, the incumbent is committed earlier than anyone else to remaining in the market and derives a strategic advantage from this commitment. This strategic advantage creates incentives for the incumbent to spend more money to keep entrants out of a market than they are willing to spend to get into it (Salop 1979b, Gilbert 1989), as shown in the following example, illustrated in Figure 11.6.

Before entry the incumbent earns a monopoly profit, π_m = \$100. With entry the incumbent and entrant together earn duopoly profits, π_d = \$80, which are less than π_m because they cannot collude perfectly. If the incumbent and entrant equally share the duopoly profits, the entrant would pay $\pi_d/2$ = \$40 to enter, whereas the incumbent would pay $\pi_m - \pi_d/2$ = \$60 to keep the entrant out.

[28] Ghemawat and Nalebuff (1985) discuss the strategic issues that arise when a firm considers exiting a declining industry. Lieberman (1990) presents an empirical analysis.

FIGURE 11.6 Extensive-Form Representation of the Raising Costs Game



Note: The incumbent's profit is the first in the pair of payoffs.

This asymmetry is natural. Because π_m always exceeds π_d, it is always worth more to the monopoly to keep the entrant out than it is worth to the entrant to enter.

If the incumbent can raise the entrant's costs as well as its own by $50, the incumbent's profit is $50 if entry does not occur and its loss is $10 if entry occurs. With these increased costs, its rival loses $10 if it enters, so it does not enter. Because a profit of $50 exceeds the $40 that the incumbent would earn if it does not engage in strategic behavior, the incumbent has an incentive to raise costs for both its rival and itself by $50.

One way the incumbent could raise costs would be to support government legislation that raises both its own and its rival's costs by $50, for example, legislation regarding pollution controls. Alternatively, the $50 could be spent on advertising. Suppose that advertising does not change total consumption but does affect the relative market shares of firms within a market. A rival must match the advertising expenditures of the incumbent if anyone is to purchase the rival's product. Given the assumed asymmetry in this example, the incumbent can commit to spending on advertising first. As the example shows, the incumbent can gain by behaving strategically to raise both its own and its rival's costs.

Another implication of this natural asymmetry between an incumbent and a rival is that the incumbent is willing to bid more than an entrant for the right to a scarce resource that would enable entry (see Example 11.6). For example, suppose that only one distribution outlet is available to an entrant to market a product. Returning to the preceding numerical example, the incumbent firm is willing to bid $60 in order to purchase that distribution channel. An entrant would be willing to bid only $40. Moreover, the incumbent, which presumably already has a channel of distribution, would simply be purchasing the additional distribution outlet in order to foreclose its use by the entrant—it might not actually use the outlet. This strategy would deter

EXAMPLE 11.6 *Value of Preventing Entry*

In the 17th century, the Dutch were mad for tulips. After extensive experimentation, a Dutch shoemaker succeeded in producing the most sought-after specimen: a black tulip. He sold this marvel to a grower from Haarlem for 1,500 florins—a fortune. Immediately, the buyer threw it to the floor and stamped it to a pulp. That grower had already produced a black tulip of his own and wanted to prevent competition.

Why did no one outbid the Haarlem grower? The bulb was worth more to a monopolist than to others, because monopoly profit exceeds duopoly profit. If many such bulbs could be created, this policy of preventing entry would not be practical.

Source: Don Paarlberg, "Economic Pathology: Six Cases." *Choices,* 1994:17–21.

entry and guarantee that the incumbent's profit would be more than it would be if entry were allowed to occur. The scarcity of distribution channels is critical in this example. If distribution were easy to obtain, the incumbent could not profitably foreclose entrants by purchasing all distribution channels (Salop and Scheffman 1987).

Many antitrust suits allege that incumbent firms buy up industry supplies of scarce resources in an effort to prevent rivals from using them. For example, it was alleged that Alcoa (which had a monopoly in aluminum) signed contracts with power companies containing a provision that prevented the power companies from supplying power to any other firm for the purpose of making aluminum.[29]

Similarly, a firm may strategically obtain *sleeping patents*: patents that, once obtained, are "put to sleep" and not used (Gilbert 1981). By so doing, the firm prevents its rivals from obtaining and using these inventions.

Advantage of Entrants. Although an incumbent often has a natural advantage that it can exploit in its strategy, sometimes the entrant has an advantage. For example, a large firm with many locations or many products may not want to cut its price everywhere when confronted by new competition in only one location or in just one product. A large firm that has more to lose than a smaller firm if price falls may prefer to let an entrant develop a small foothold rather than engage in a price war. Thus, a firm that enters and remains at a small scale can compete without fear of retaliation.

If a large firm has stores at many locations that all charge the same price (perhaps because of economies of scale in national advertising), and another firm enters in just one of these locations, the large firm may abandon that location rather than lower its prices everywhere. Of course, if the large firm believed the small firm (or other entrants) would continue to expand, it might choose to fight vigorously.

[29] *United States v. Aluminum Co. of America,* 148 F.2d 416 (1945). Lopatka and Godek (1992) dispute this allegation.

If the large firm decides to fight, one alternative to reducing its uniform price is to introduce a new brand, sometimes called a **fighting brand**, whose price is low and whose availability is limited to those areas where a (small) rival is successful. In this way, the large firm can engage in competition without lowering price to all its customers. See Example 11.2.

Similarly, a firm that produces many substitute products views price competition in one product as costly because such competition also affects its revenues from other products. Conversely, a firm that produces complementary products does not find a price war in one product as costly if lost profits in one product are offset by increased profits in others. A firm has less to fear about competitive reaction to its aggressive pricing policy when its rival produces several substitute products, is relatively large, and believes that the entering firm only wants to occupy a small market niche (Bulow et al. 1985a, Fudenberg and Tirole 1984).[30]

Welfare Implications and the Role of the Courts

It is difficult to determine whether strategic behavior raises or lowers welfare. Moreover, it is difficult to distinguish competitive from strategic behavior.

Some strategic behavior lessens competition and harms consumers. For example, successful predatory pricing that leads to market power in the long run has no socially redeeming virtues.

Other types of strategic behavior, however, can produce socially desirable results. For example, even if R&D investments are a strategic action, consumers may ultimately benefit from lower prices. Even when strategic behavior leads to monopoly, consumers may benefit. Indeed, patents are designed to create monopolies because the incentive of monopoly profits encourages firms to develop new knowledge (see Chapter 16). These examples suggest that the welfare implications of strategic behavior need to be considered case by case. Strategic behavior may be socially undesirable in one set of circumstances and socially desirable in another.

In practice, it may be difficult to distinguish between strategic behavior and desirable competitive behavior. For example, passing along cost savings through lower prices, investing in R&D in order to lower costs, gathering marketing information, and arranging distribution channels for a product are all desirable features of competition that are difficult to distinguish from strategic behavior.

U.S. antitrust laws (Chapter 19) allow the government to intervene if it believes that firms are taking actions that lessen competition. The antitrust laws also give private plaintiffs who are the victims of such behavior the right to sue. The difficulty of distinguishing between beneficial competition and undesirable strategic behavior presents government enforcement agencies and the courts with a problem. Too little enforce-

[30] Bernheim and Whinston (1990) and Whinston (1990) provide examples of how multimarket contact and other devices can signal likely competitive responses. Products are called *strategic complements* if an aggressive action in one product induces an aggressive reaction (such as a firm's meeting its rival's price cut) and *strategic substitutes* when the reaction is dissimilar (a firm reduces output in response to a rival's expansion). See Bulow et al. (1985a).

ment leads to bad behavior and monopoly power, whereas too vigorous enforcement may deter firms from pursuing desirable forms of competition for fear that this competition will be misinterpreted. For example, if attempts to compete through lower prices trigger lawsuits for predatory pricing, then firms that lower prices for purely competitive reasons risk being sued for predatory pricing. Thus, the proper role of enforcement agencies and the court in dealing with strategic behavior is not easy to define.

When designing proper enforcement policies, one should consider the costs of making an error. The success of strategic behavior depends on the asymmetry between the firm practicing the strategic behavior and the firm that is its target. If a new firm can eventually model itself after an incumbent firm (for example, adopt its technology to remove any asymmetries in the use of labor), then strategic behavior is doomed to fail. In such cases, even if the courts fail to prevent its exercise, strategic behavior leads only to temporary market power that is eventually eroded as firms learn how to imitate the incumbent. (Of course, such market power causes more harm the longer it takes for new firms to enter.) In contrast, if the courts falsely condemn desirable behavior as strategic, this harm is not eliminated by future actions by firms.

Cooperative Strategic Behavior

Cooperative strategic behavior comprises those actions that rival firms take in their own self-interest that raise the oligopoly price closer to the monopoly level. The theory of cooperative strategic behavior relies on cartel theory (Chapters 5 and 6), which holds that oligopoly profits depend on the ability of each member of the cartel to assure the others that it is not trying to steal its rivals' customers.[31] The greater the mutual assurance that firms are not stealing each other's customers through lower prices, the easier it is for them to succeed in charging a price above the competitive level.

In Chapter 5, we examined several practices that can facilitate collusion (see also Salop 1986). Here we consider several additional practices and examine their treatment under the antitrust laws.

Practices That Facilitate Collusion

Oligopolies employ a variety of cooperative strategic actions to elevate price (for example, by facilitating collusion). Chapter 5 discusses the use of most-favored-nation and meeting-competition clauses in contracts, information sharing, dividing the market, and other methods. The following are some other important approaches.

Uniform Prices. If all of a firm's customers are charged exactly the same price, then it is costly for the firm to try to steal a rival's customers by offering them a slightly lower price. The reason is that the slightly lower price, by assumption, must also be offered to

[31] Explicit agreement is not necessary for an oligopoly to succeed in raising price above competitive levels (Chapters 5 and 6).

all of the firm's existing customers (this approach may be implemented through a most-favored-nation clause in contracts; Edlin 1997). This uniformity of price lowers the firm's gain from stealing away the rival's customers. Moreover, if all of the firm's customers pay identical prices, it is easier for a rival to learn when a firm has lowered price.

The question arises then as to what forces a firm to charge a single price to all consumers. One answer is government legislation. The Robinson-Patman Act requires that firms charge identical prices to customers who buy identical products.[32] Firms sometimes use this law as a justification for not granting selective discounts to particular customers. Therefore, the Robinson-Patman Act may facilitate collusion.

Penalty for Price Discounts. A more dramatic way of reducing a firm's incentive to steal another firm's customers by lowering price is for each firm to adopt a policy whereby any lower price is passed on not just to the firm's current customers (as occurs with a uniform price to all customers), but to all of its past customers over some time period. For example, if a firm signs a contract with buyers that entitles them to receive any price discount that occurs in the next year, then the firm has a great disincentive to lower price. Its rivals know that this firm is not likely to discount price because of the cost of applying the discount to past customers.

Advance Notice of Price Change. Cartels have difficulty maintaining a pricing arrangement when prices change (Chapter 5). At the time of a price change, firms distrust each other because each is likely to be selling at different prices.

Suppose that it is clear that the prices in an oligopoly that is not a cartel should rise. Which firm should increase the price? Some industries have a natural price leader, but many others do not. The first firm to raise price is at a serious disadvantage because it loses sales from its relatively high price. Of course, if rivals eventually match the higher price, all firms are better off with the higher prices. Nonetheless, if the firm that initiates the increase suffers a loss relative to its rivals who follow slowly, then no firm wants to be the price leader.

One way around this problem is to use advance notice of price increases, a tactic that allows other firms in the market to decide whether to go along with the price increase before it becomes effective. If rivals decide not to go along, the firm that announced the price increase can rescind it. In such a circumstance, firms need never find themselves selling at different prices in the market, and the disincentive to raise price is eliminated.[33]

Using the same logic, at times of decreased demand, the firm initiating the price decline gains relative to its rivals who take time to respond to the price cut. Thus, each firm

[32] The relevant portion of the Robinson-Patman Act applies if customers are firms that compete against each other and if the effect of the price discrimination is to substantially lessen competition.
[33] In industries in which price increases are announced in advance, firms compete over how much buying to allow before a price increase takes effect. Some firms have policies of allowing customers to buy an extra month's supply at the old price. More usually, the amount of buying at the old price is variable and differs over time and across firms.

has an incentive to cut its price first. Advance notice of price decreases mitigates this incentive by ensuring that no firm gains an advantage from taking the lead in cutting price.

Several industries give advance notice of price increases, and some have been the subject of lawsuits and investigations. For example, in the 1990s the Department of Justice investigated the major airlines' use of advance notices of fare changes and alleged that the airlines' communication of their pricing intentions led to elevated prices. The case was settled with the airlines agreeing to cease the practice, despite consumer groups' support of the practice. However, there is no evidence that cessation of the advance price announcements had any effect on fares (see Carlton, Gertner, and Rosenfield 1997, and Borenstein 2003).[34] Also see Example 11.7.

Information Exchanges. Information exchanges between firms can facilitate cartels or promote efficiency. One way a firm can convince rivals that it is not trying to steal customers through a price discount is to announce the identity of its new customers and the price and quantity terms offered. The firm makes this announcement so that when a customer shifts suppliers, a price war is not triggered by rivals who believe incorrectly that the shift was due to a lowered price. Another method of conveying information is disseminating publicly what the firm's strategy is so that rivals will not misinterpret the firm's actions and can coordinate with the firm's strategy. See Farrell (1987).

There may also be legitimate efficiency reasons for industry members to exchange information. When a centralized market does not exist, disseminating price information can improve market efficiency (see Example 11.8). Moreover, firms can monitor their own efficiency better if they can compare their costs to those of other firms.

Delivered Pricing. A delivered pricing system specifies the total delivered price (inclusive of freight) that a buyer must pay as a function of the buyer's distance from a specified location (a *basing point*), but not of the location of the seller. A delivered pricing system can be created by specifying the total delivered price as the sum of a going market price at the basing point plus freight from that point. For example, steel used to be sold with Pittsburgh as the basing point. If an Ohio steel mill shipped steel to Chicago, the price the buyer paid equaled the going price of steel in Pittsburgh plus freight from Pittsburgh to Chicago. The freight charges were calculated from standard published rate schedules.

At first glance, delivered pricing systems seem so bizarre that they inspire suspicion. Indeed, many economists believe that delivered pricing is an odd mechanism adopted only to facilitate collusion.[35] It facilitates collusion because it prevents competing firms from secretly granting discounts disguised as low freight charges. Forcing all firms to charge the same freight and same price makes it easy to detect deviations from a collusive price agreement.

[34] Carlton served as an expert for the airlines, and Borenstein served as an expert for the Department of Justice.

[35] See Thisse and Vives (1992) on how such a pricing scheme can arise in a static game only under somewhat contrived circumstances, but how it can be an effective punishment in a repeated game.

EXAMPLE 11.7 *The FTC versus Ethyl et al.*

In 1979 the FTC brought an antitrust suit [*E. I. du Pont de Nemours & Co. v. FTC,* 729 F.2d 128 (2d Cir., 1984)] against four producers (du Pont, Ethyl, Nalco, and PPG) of an additive to leaded gasoline. The FTC charged that certain business practices of these four firms had the effect of facilitating collusion in the industry. Some of the practices that the FTC attacked were the use of 30-day advance notice of price increases (not decreases) to buyers, most-favored-nation clauses, and public press announcements of all price changes. Although the evidence indicates that this industry did not behave perfectly competitively, the real issue was not the competitiveness of the industry but rather whether the practices decreased the competitiveness of the industry.

The primary economic theory of the FTC was that the industry practices eliminated uncertainty in the industry and improved the ability of rivals to match each other's prices. Empirical tests of the effect of some of the practices are possible. For example, public press announcements of price changes had ceased in the industry. Yet there appeared to be no difference in the speed with which rivals were able to match each other's prices before and after the cessation of the press announcements. Even though the advance notice provision applied only to price increases and not to decreases, rivals matched each other's price decreases as rapidly as they matched their price increases.

The trial record indicates that because many of the practices were adopted at a time when only one firm was in the industry, their adoption was not intended to facilitate collusion. Thus, these practices also might serve an efficiency function. For example, buyers might value advance notification of price increases so they could plan better. Moreover, some of the issues in the case, such as an attempt to ban public notice of price announcements in the *Wall Street Journal,* raised issues of free speech. The FTC ruled against the industry; however, the Second Circuit Court of Appeals overturned that verdict.

Note: Carlton appeared as an expert in this case on behalf of Nalco Industries.

The pricing system that many economists predict should emerge with competition is called **FOB pricing**: The buyer pays a *free-on-board* (FOB) price, where the seller loads the good onto the transport carrier at no cost to the buyer, plus the actual freight.[36] Under such a system, the freight charge varies with a buyer's location, and

[36] Because prices reflect costs in competition, economists expect purchasers to pay for FOB pricing and for actual freight under competition. In fact, firms in competitive industries often use delivered pricing because it is simple and saves on administrative costs. For example, firms may use uniform delivered pricing as long as freight does not vary much among customers. Typically, furniture stores include delivery in the price of an item, provided that customers live reasonably close to the store. Some firms have zone pricing, in which buyers who live in a firm's region pay lower freight charges than those in more distant regions. It appears that uniform delivered pricing is often followed as long as the variation in freight charges among customers is 10 percent or less (Carlton 1983c).

EXAMPLE 11.8 *Information Exchanges: The Hardwood Case*

In the *Hardwood* case [*American Column and Lumber Co. et al. v. United States,* 257 U.S. 377 (1921)], a group of lumber mills were accused of violating the Sherman Act. These producers ran the American Hardwood Manufacturers Association, which, under their *Open Competition Plan,* collected and disseminated price and production information. The number of mills in the industry was large—about 9,000 mills in 20 states. Participation in the Open Competition Plan was voluntary, and 465 mills participated (representing 30 percent of output).

Although monitoring output and prices can facilitate collusion, collusion can be difficult with such a large number of independent firms in the industry. Even though information-sharing arrangements are particularly suspect when the number of firms in the industry is small enough to make collusion likely, the Supreme Court ruled that the information exchange violated Section I of the Sherman Act, and the information dissemination ceased. Alexander (1988) contends that the information exchange had no anticompetitive impact on market output and instead was likely an attempt to disseminate valuable but costly information to a competitive industry.

sellers can cut price by undercharging for freight (which rivals cannot easily observe). In such a pricing system, firms at different locations generally quote different prices to a buyer, so that enforcing a collusive agreement about price is difficult.

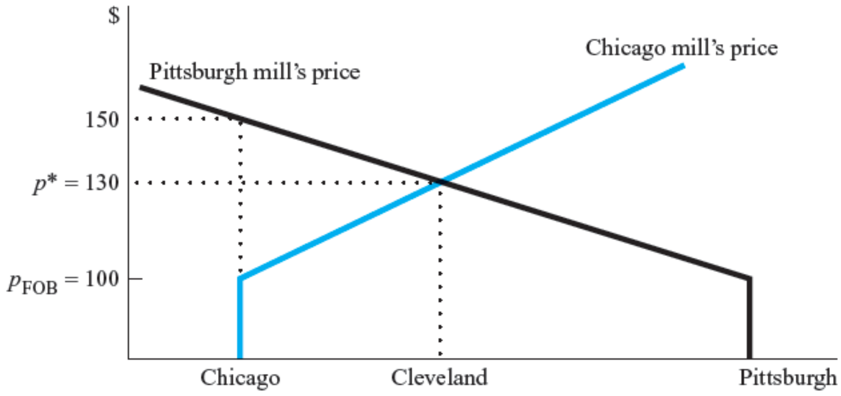
Underlying this story is the implicit assumption that colluding firms detect cheating by observing deviations from an agreed-upon price schedule. Another way firms detect cheating is by monitoring shifts in market share: If firms see that a rival is inexplicably gaining customers, they suspect cheating. When this latter method of detecting cheating is used, delivered pricing may be less effective than FOB pricing in facilitating collusion.

A great disadvantage to collusion through the use of delivered pricing is that it fails to allocate the market to sellers. For example, suppose that there are two sellers of steel—one in Chicago and the other in Pittsburgh. If Pittsburgh is the basing point for delivered pricing, it might be profitable for either a Chicago or a Pittsburgh steel mill to sell steel near Chicago. In such a case, no inference of cheating would follow if the Pittsburgh mill were seen selling in the Chicago area. If the firms instead adopted FOB plant pricing plus freight, the market would be clearly divided: Consumers closer to Chicago would buy from Chicago, and those closer to Pittsburgh would buy there.

Figure 11.7 illustrates that FOB pricing creates a clean market division precisely because firms charge most buyers different prices. As the figure shows, suppose that each firm agrees to charge the same FOB price at its plant and to charge actual transportation charges. The price lines represent the price that a buyer at any location must pay. They rise as one moves away from the location of each firm to show that transportation costs rise with distance. All customers to the west of Cleveland buy from Chicago, and all customers to the east of Cleveland buy from Pittsburgh.

If the Pittsburgh steel mill were seen selling in the Chicago area, one would infer that it was not adhering to the FOB pricing agreement. In contrast, under delivered

FIGURE 11.7 FOB Pricing Divides the Market



pricing, all firms charge the same price, and there is no neat market division. Moreover, under delivered pricing, the Chicago firm could be selling in Pittsburgh and the Pittsburgh firm could be selling in Chicago so that there is a lot of inefficient cross-hauling of the product. In general, the greater the distance between the firms and the more important the transportation charges, the better FOB pricing is as a means of market allocation and collusion compared to delivered pricing (Carlton 1983).

There is an important difference between delivered pricing and FOB pricing. In an equilibrium with FOB pricing, all firms must charge the *marginal* buyer the same price ($130 at Cleveland in Figure 11.7). At points other than Cleveland, however, the firms located at Pittsburgh and Chicago charge *different* prices under FOB pricing but not under delivered pricing. Despite this difference between delivered and FOB pricing, it is sometimes possible for the two pricing systems to look alike.

Suppose that many steel consumers are located in both Pittsburgh and Chicago, but that initially all of the many steel mills are in Pittsburgh. In a *competitive* equilibrium, if it costs $100 to produce a ton of steel, a buyer in Chicago pays $100 plus freight, say $50, from Pittsburgh, for a final price of $150. If a very small steel mill opens in Chicago, how will the equilibrium change? At any price below $150, the Chicago steel mill has more business than it can possibly handle because it is a very small mill. Therefore, there is no reason for the Chicago mill to charge below $150! In this competitive equilibrium, the Chicago mill charges what appears to be a delivered price based on Pittsburgh freight. Yet it is the competitive equilibrium. The small Chicago mill earns a rent because it has chosen such a desirable location.

Failure to understand that price-taking firms at some locations (like Chicago in the example) can charge $150 could lead one to mischaracterize a competitive FOB pricing arrangement as collusive delivered pricing. As more steel mills locate in the Chicago region, they will eventually be unable to sell all their steel in Chicago at $150, and they will start selling outside of Chicago (say, Cleveland). When that occurs, the

FOB price can no longer equal $150 in Chicago. If the FOB price is $150 in Chicago and the Chicago mill adds freight to Cleveland, the price to Cleveland exceeds the FOB price at Pittsburgh plus freight (of $30) from Pittsburgh, and no Chicago steel would be sold in Cleveland (see Figure 11.7). Therefore, the new competitive equilibrium must involve an FOB price below $150 in Chicago. Eventually, competitive entry of steel mills in the Chicago region will drive the Chicago price down to a marginal cost of $100. Buyers close to Pittsburgh buy from Pittsburgh, and those close to Chicago buy from Chicago.

The intensity of government scrutiny of spatial pricing schemes has varied over time. In *FTC v. Cement Institute* (333 U.S. 683 [1948]), the FTC alleged that a conspiracy among cement producers was facilitated by the use of delivered pricing with basing points. The FTC won the case, but subsequent political pressure from businessmen led to congressional hearings that seemed to stop the FTC from bringing many more cases.

More recently, pricing policies in the lumber industry were attacked in *Boise Cascade v. FTC* (63 F.2d 323 [9th Cir. 1980]). All plywood used to come from the Pacific Northwest. Beginning in the early 1960s, some plywood was shipped from the South. Initially, little southern plywood was shipped, but eventually significant amounts were shipped. The price of southern plywood always equaled the price of the wood plus freight, where the freight fee was based on shipping charges from the Pacific Northwest. Although it follows from the logic of the steel example (Figure 11.7) why the price of southern plywood was initially quoted in this way, it is harder to explain why these policies continued as the southern plywood industry grew.

The lumber firms practicing the policy claimed that it was just a convenient device to facilitate comparisons of price quotes from the South and the Pacific Northwest. They also claimed that the southern FOB price was different from the Pacific Northwest FOB price. The implication was that each FOB price was determined by the forces of supply and demand, and the resulting pricing equilibrium was the competitive FOB pricing equilibrium predicted by the reasoning in our steel example. For example, suppose that the true competitive price of southern plywood to a buyer in New York is $200, consisting of two parts: $100 of true freight plus $100 of true FOB price. If the freight from the Pacific Northwest is $150, then the quoted price for southern plywood is $50 FOB plus $150 freight, for a total price of $200. The court decided that the pricing scheme did not represent illegal collusive behavior. Gilligan (1992) shows that the subsequent replacement of the peculiar pricing scheme with FOB pricing caused the price to fall in the South but not in the Northwest.

In summary, the method of spatial pricing used can affect the ability to collude. Although using delivered pricing can facilitate collusion, it does not always do so. In some settings, delivered pricing can lead to greater competition than FOB pricing. Moveover, delivered pricing can be more efficient (reduced transaction costs) than FOB pricing, and therefore can appear in competitive industries.

Swaps and Exchanges. Firm C, located in Chicago, has a customer in Boston, and Firm B, located in Boston, has a customer in Chicago. To minimize shipping costs and service their customers, Firm C "swaps" or "exchanges" one unit of its output in Chicago for one unit of Firm B's output in Boston. Although this arrangement may appear odd, it is common in industries where the product (such as chemicals, gasoline, and paper) is relatively homogeneous and transport costs are high. A swap is not equivalent to two independent buy-sell transactions. Indeed, there are typically no prices involved in swaps, and the final customer transacts with only one of the firms.

Swaps have often been attacked as a facilitating device in antitrust cases. The theory is that swaps are a mechanism to divide the market, allow rivals to communicate, and prevent competition from occurring, yet still allow firms to service distant clients. Swaps also can make it difficult for a small entrant to service distant customers because the new entrant has few locations that it can use to engage in swaps.

Although this explanation is theoretically possible, there are other rationales for swaps. Swaps can be an enforcement mechanism to guarantee timely delivery in industries where supply availability is key. Firm C is reasonably assured that it can rely on Firm B in Boston as a supply source for its Boston customer because it knows that Firm B relies on it for deliveries in Chicago. If Firm B fails to deliver to Firm C's customer in Boston, then Firm C will not deliver in Chicago for Firm B. Frequently, firms closely monitor their swaps to make sure that Firm C and Firm B are "in balance" and don't "owe" any product to the other for long periods.

Cooperative Strategic Behavior and the Role of the Courts

Cooperative strategic behavior, requiring, as it does, that firms all choose similar actions, seems superficially easier to identify and condemn than other types of strategic behavior. After all, any agreement or practice that tends to reduce competition is likely to harm society. The problem is that many practices may be chosen not to restrict competition but for efficiency reasons (see Example 11.7).

For example, advance notice of price changes might benefit consumers even though it could also facilitate collusion. A policy that condemns practices reached through agreement to limit competition seems correct. A policy that condemns business practices, however chosen (for example, at the insistence of the buyer), that conceivably could affect collusion, is probably too broad and would leave firms in a quandary as to which of their policies would be subject to antitrust scrutiny. The result might be to deter firms from adopting efficient practices that their customers desire.

SUMMARY

Strategic behavior is an attempt by a firm to influence the market environment in which it competes. This environment includes the beliefs of rivals and customers, the technologies and costs of the firms, and the knowledge of customers. For noncooperative strategic behavior to be successful, the strategy must be believable to rivals. Asymmetry among firms is a key ingredient of successful strategic behavior.

Predatory pricing is a costly policy for a firm to follow. There have been only a few documented cases of successful predatory pricing in which price was below some measure of cost. Other noncooperative strategies, such as price cuts down to (but not below) cost, strategic R&D, and raising rivals' costs, might well be more profitable, and one should expect them to be used more often than predatory pricing.

Cooperative strategic behavior requires firms to act similarly and is a direct application of the theory of oligopoly studied in Chapters 5 and 6. Any practices that firms can use collectively to reduce uncertainty about each other can facilitate collusion.

The proper legal posture toward strategic behavior is complicated. Some strategic behavior helps consumers, for example, by encouraging investments. Other types of strategic behavior harm consumers. Even successful strategic behavior can maintain market power only until entry occurs. Strategic behavior is therefore likely to be harmful only in industries in which entry is difficult.

Distinguishing the good strategic behavior from the bad can be difficult for both economists and courts; hence, great care should be used in applying antitrust laws against apparent strategic behavior. Society faces a trade-off between too little enforcement, which leads to market power, and too much enforcement, which deters healthy competition.

PROBLEMS

1. Suppose that two rival manufacturers sell to the same retail stores. If one manufacturer imposes the condition that the retail stores carrying its product must charge the same retail price for its product as it charges for its rival's product, what happens to the prices of these manufacturers?
2. Describe how a joint venture for research and development can reduce competition.
3. If a firm has debt, it must pay interest to the debtholders. Suppose that there is a blot on a manager's record if the firm he or she operates goes bankrupt. Discuss whether the use of a high ratio of debt to equity among all firms in a market could be a practice that facilitates collusion. Consider the consequences if firms issue debt in different years and if interest rates vary from year to year.
4. Using a model of price predation, explain why driving a rival into bankruptcy does not, by itself, enable the predator to charge monopoly prices. [*Hint:* What happens to the assets of the bankrupt firm?]
5. Suppose that the Japanese firms in Example 11.1 were indeed predating for 20 years in the hope that in the 21st year and thereafter they could charge a monopoly price. Suppose that the annual loss is $1 million for each of the first 20 years, and let π_m be the annual flow of monopoly profits thereafter. If the interest rate is 10 percent, calculate how high π_m would have to be in order for the predation strategy to be profitable. [*Hint:* The discounted present value of the 20 years of annual loss is

$$\frac{1}{r}\left[1 - \left(\frac{1}{1 + r}\right)^{20}\right],$$

and the discounted present value of an annual profit of π_m beginning in year 21 is

$$\frac{\pi_m}{r}\left(\frac{1}{1 + r}\right)^{20},$$

where r is the interest rate.]

Answers to odd-numbered problems are given at the back of the book.

SUGGESTED READINGS

Gilbert (1989), Ordover and Saloner (1989), Tirole (1988), and Wilson (1992) provide excellent overviews of modern theories of strategic behavior.

Farrell and Klemperer (2003) survey the literature on switching costs and network effects.

APPENDIX 11A

The Strategic Use of Tie-in Sales and Product Compatibility to Create or Maintain Market Power with Applications to Networks

In this appendix, we study the strategic incentive to use tie-ins and product compatibility decisions to create or maintain market power. (In Chapter 10, we studied the use of tie-in sales as a method of price discrimination.) We apply these strategic concepts to network industries.

Tie-in Sales

A tie of product *B* to the sale of product *A* can, under certain circumstances, affect the market power of a monopoly producer of *A* in the market for either *A* or *B.* By tying the sale of product *B* to *A,* the monopoly producer of *A* can reduce the size of the market available to the rival producers of *B.* If *B* is not produced in a constant returns environment with competition, then the tie can affect the market structure of product *B* (Whinston 1990) and benefit the monopoly producer. With constant returns to scale in the production of *B,* such a tie would not benefit the monopoly producers under certain conditions (see Chapter 10).

For example, suppose that local residents share their island with a hotel.[1] The residents frequent two local tennis clubs, each of which also serves some of the hotel's guests. If the hotel builds a health club and ties the guests to that club (for example, by allowing free use of the facilities), such an action may deprive the local clubs of the necessary size to support themselves, causing the hotel to become the monopoly owner of a tennis club on the island. This result depends critically on the presence of scale effects in the provision of tennis clubs. This "foreclosure of competition" for product *B* has been the traditional antitrust concern with tie-in sales. Nalebuff (forthcoming) shows that a monopoly of good *A* may sell its product tied to another independent good *B* so as to prevent rivals from achieving sufficient scale to enter the market for *B.*

Tie-in sales can also have an effect on competition in the market for *A* and can enable a monopoly supplier of *A* to maintain or even extend its market power into new products (Carlton and Waldman 2002). For example, suppose that consumers are willing to consume products *A* and *B* only if they can consume them together, and

[1] We thank R. Gertner for this example.

that one firm is a monopoly supplier of *A* and uses product *B* as a complementary good. By tying *B* to *A,* the monopoly producer of *A* monopolizes *B.* So far, the story is the same as our earlier one. However, now imagine that another firm wishes to enter the market and compete in producing *A.* That firm, even if it is an especially efficient entrant, can be deterred from entering because it will have no supply of product *B* at least initially.

The original monopoly remains the monopoly producer of *A* by using tying to control (or raise the cost of) key complementary products that entrants to produce *A* require. Indeed, the firm that controls *B* can prevent entry into any new market *A** that requires *B,* and in this way can swing its original monopoly of *A* into one for *A**. This scenario is most likely in a market where technology changes rapidly so that the market size for *B* does not remain large for extended periods, where sunk costs (such as R&D costs) are large relative to the scale of the market for *B,* and where entry into *B* takes a long time (without these conditions, firms will enter the market for *B*).

One possible example of such strategic behavior comes from the computer industry. Suppose that one firm is the only producer of an operating system for personal computers. All software is therefore designed to work with this (and only this) operating system. A new device, a hand-held computer, is invented that could use many different operating systems. However, because the available software works only with the monopoly's operating system, the monopoly has an advantage in selling the operating system for the new device and could thereby become the monopoly supplier of the operating system for the new device.

Product Compatibility

Many systems consist of complementary products *A* and *B* that work together: A stereo and its headphones, a computer and its printer, and a camera and its film. Should a manufacturer of computers produce printers that are compatible with its competitors' computers or not?

To demonstrate the role that compatibility plays, we consider two situations. Initially, two firms each manufacture products *A* and *B.* The unit costs for the first firm are $1 for *A* and $2 for *B,* while those for the second firm are $2 for *A* and $1 for *B.* With Bertrand competition on each component, the price of *A* is $2, and the price of *B* is $2, so that the total system price is $4.

In contrast, suppose that each firm produces a version of product B that will not work with a rival's product *A.* Now consumers are only willing to buy *A* and *B* together from one firm. The Bertrand system price for both products is $3.

This numerical example illustrates a general, if somewhat counterintuitive, principle: Product incompatibility may lead to more vigorous competition. When there is product compatibility, a cut in the price of one component stimulates demand for that component but does not stimulate demand to the same extent for the complementary component produced by the same firm (because some consumers use the complementary component produced by the other firm). In contrast, if firms produce products that are incompatible with rivals' products, a firm's price cut in one component automatically increases demand for its complementary component. Hence, the gain to price cutting is greater with product incompatibility, and the consequence is more vigorous competition and lower prices. Accordingly, firms may choose compatibility in order to avoid competition.[2]

If the various components produced by rival firms are not identical but differ in their characteristics, then the ability to mix and match will lead to a greater variety of products, and this greater variety of products stimulates demand and benefits some consumers. Moreover, with compatibility, the more efficient firm for each component will wind up producing that component. Matutes and Regibeau (1988, 1992) show that full compatibility maximizes social welfare, ignoring the costs to achieving compatibility.

Networks

In the last decade, interest in what has been labeled "network industries" has increased greatly. Loosely speaking, these are industries in which activities in one part of the network affect other parts. An example is a phone network, whose consumer value depends on the number of consumers hooked up to the network. The literature often stresses the failure of competition to work in network industries but unfortunately has not always been precise in tracing the failure of competition to the network feature. In this section, we first define network effects and then discuss strategic use of tie-in sales and product design in networks.

A physical network consists of pathways connecting nodes. Good examples are a railroad network (tracks connect stations), a telephone network (wires connect phones), and an electricity grid (wires connect generators and users). In the operation of such a network, there can be interactions between the various parts of the network. For example, the cost of shipping electricity on one path depends on the electricity loads on the other paths. The cost of shipping by rail from *A* to *B* depends on the amount of other traffic on the track between *A* and *B* and whether that traffic can be easily shifted to another track.

In such networks, Koopmans and Beckman (1957) showed that the use of prices alone does not necessarily lead to the optimal use of the network among decentralized firms, each owning different parts of the network. If the network is a single firm, that firm will internalize these network interactions and have an incentive to operate its network efficiently. Hence, there is an incentive for a single firm to operate a network. The problem that arises here is identical to Coase's insight that a firm is created when it can produce and allocate goods more efficiently than can a market (Chapter 12). A firm controlling a network can achieve scope economies by virtue of its superior allocation ability compared to a decentralized price system attempting to coordinate independent parts of one network (Carlton and Klamer 1983). The recent widespread consolidation into large national networks in the airline, railroad, and telecommunication industries vividly illustrates these forces.

[2] The idea that product incompatibility can increase the vigor of price competition has a direct implication for the incentive to use tie-in sales. A tie-in sale effectively creates product incompatibility, thereby intensifying competition, with the result that entry can be deterred (see Whinston 1990).

Even though there are efficiency gains from a national network, offsetting problems occur due to monopoly. Whether a network industry is a natural monopoly depends on how costs change as the network gets larger. Just as there can be many competing multiproduct firms, so too can there be many competing national networks.[3]

It may be impossible for any one firm to establish a property right in some network. No one firm can take over part of Chicago's roads and allocate cars over its streets. Moreover, there are circumstances in which networks aren't physical and exclusion from the network cannot occur. For example, imagine three people: *A, B,* and *C. A* and *B* wish to speak together, as do *B* and *C.* The trio form a "network." If *A, B,* and *C* all learn English, that is more efficient than if *A* and *B* learn English while *B* and *C* learn French. Yet no one person or firm owns the "right" to set language standards. (Of course, countries often do try to influence language choice.) When standards are adopted voluntarily and no one can exert property rights over their use, the familiar economic problem of an inefficiency due to an unpriced resource can result.

Networks Where the Type of Interaction Depends on Size

The recent literature focuses on two effects in networks where the type of interaction depends on size. In the "direct" network effect, the benefit to a network user depends directly on how many other users are hooked up to the network, as in a telephone network. In the "indirect" network effect, the benefit to a user arises indirectly because the number of users of the network affects the price and availability of complementary products. For example, an increase in the number of computer users of Windows (an operating system) leads to the development of additional software that is compatible with Windows. Let us examine each of these two effects in some detail. See Dranove and Gandal (2003) and Saloner and Shepard (1995) for empirical measurement of network effects.

Direct Network Effect. The value that a phone user places on the phone network rises with the number of people that user can call. Moreover, the decision of additional users to join such a network benefits existing users, who can now communicate with the new subscribers.

It is sometimes suggested that, because a new user provides a benefit to old users, there is an externality and a market failure. That need not be true. The theory of clubs (Buchanan 1965) was developed to deal with cases in which the size of a firm influences the quality of its product. In such a case, a competitive firm has the appropriate incentive to choose the efficient size[4] of its customer base—provided that the size is finite. Although there is a benefit to size, there may also be costs; hence, the optimal

[3] A complication arises when it is desirable for networks to interact. Then, the separate networks can no longer be regarded as independent competitors. Moreover, strategic denial of interconnection could occur in an attempt to inflict costs on a rival. See Laffont, Rey, and Tirole (1998a, 1998b).

[4] A variant of this point arises in Chapter 17, where we discuss a firm's incentive to create the optimal composition of heterogeneous consumers when the heterogeneity of consumers influences the firm's costs (Carlton 1991).

network size can be finite. The firm will generally charge both a variable user fee and a membership fee.

If the optimal size of the network is infinite, then the market may have a natural monopoly. Even here, competition can occur among a few networks and be a stable equilibrium outcome as long as the scale economies are not too great. Indeed, the new technologies in telephony have allowed competition among several phone networks to become a reality for many communications products.

Indirect Network Effect. The indirect network effect arises when the benefit of size leads to increased variety or lower pricing of complementary products. Such an effect occurs typically because of scale economies in production of the complementary product so that lower costs and more competitors go along with bigger markets. As more people use a particular operating system, more software is developed for that operating system. This type of effect is the same as occurs in nonnetwork industries: As more people play tennis, there could be increased variety and lower prices for tennis balls.[5]

Networks and Strategy

We now examine how tie-in sales and product design can be used strategically in network industries. The analysis can become quite complicated for several reasons.[6]

First, since price competition in our example is reduced and profits are elevated under product compatibility, an incentive is created for new firms to enter the industry or for firms to engage in R&D to produce better-quality products. In other words, the suppression of price competition due to product compatibility leads to an offsetting benefit caused by the increased incentive for nonprice competition. Even when theory can guide us about the presence of these trade-offs, the empirical magnitude of the trade-offs is uncertain and is an area for future research. In assessing these trade-offs, we need to keep in mind the key point that the social rate of return to innovation typically exceeds the private rate (Chapter 16).

Second, in network industries, competing networks can "tip" so that if one network overtakes the other, the other becomes insignificant. For example, the VHS (video recorder) recording format overtook and put out of business the rival Beta recording

[5] Some people contend that this indirect network effect necessarily creates a market failure because the lowered price of the complementary product is not considered by the primary users of a network. This inference is not correct for at least two reasons. First, given the usual assumptions of how markets work, one can show that this indirect effect leads to precisely the correct market incentives. Second, even if the conditions for efficiency fail, the empirical relevance of such resulting inefficiencies is a matter of debate. See Liebowitz and Margolis (1994). The alleged failure of competition when indirect effects are present is related to claims that standard setting can be inefficient.

[6] There is an enormous literature of networks and strategic behavior. See Farrell and Saloner (1985, 1986a), Economides (1988a), Katz and Shapiro (1985b, 1994), and the special issue of the *International Journal of Industrial Organization* (1996) on networks, especially the Economides (1996) article. For a discussion of strategic issues related to the Microsoft case, see Carlton (2001), Carlton and Waldman (2002), Evans et al. (2000), and Whinston (2002), and the references they cite.

format. In such settings, strategic behavior to prevent a rival from prospering has enormous competitive benefits. A firm can become and remain dominant by using the standard setting process to impede rivals or by creating unnecessary product incompatibilities, such as making it difficult for rivals to use complementary products.

Third, the expectation of future network size influences the desirability of buying a product today that will last into the future. A key marketing claim of a firm that produces a durable good is that new complementary products for its network will be available and that its network will be larger in the future. Advance announcements of future new products can harm a rival if those announcements are believed. Computer firms frequently announce that a new software product will be available next year in order to dissuade consumers from buying a rival's product today. Sometimes these announcements turn out to be false: The alleged product is nonexistent "vaporware."

A firm can try to convince consumers of the availability of future products and of its future size by signing contracts in advance with providers of complementary products, or by licensing in advance its intellectual property to others. If the expectation of size is important to consumers, then in such a network industry, there can be multiple equilibria with differing numbers of competitors.

Fourth, the expectations of future network circumstances can be affected by how fast the network is growing and how costly it is for consumers to switch networks. If consumers are locked into their network, the network might find it profitable to concentrate on introducing new products that appeal only to new consumers (unless price discrimination can be used). This could lead to excessive product introduction, especially in rapidly growing networks. Conversely, in a stable environment, a network might fail to introduce a new, desirable technology that benefits its locked-in customers, unless it can charge for the innovation, resulting in sluggish technological change. One way to avoid this problem is with pricing that depends on whether one is a new or an old customer: New purchasers of the latest version of a software program are often charged more than existing customers who are upgrading.

CHAPTER 12

Vertical Integration and Vertical Restrictions

> *Outside the firm, price movements direct production, which is co-ordinated through a series of exchange transactions on the market. Within a firm, these market transactions are eliminated and in place of the complicated market structure with exchange transactions is substituted the entrepreneur-coordinator, who directs production. It is clear that these are alternative methods of coordinating production.*
>
> *—Ronald Coase (1937)*

A firm that participates in more than one successive stage of the production or distribution of goods or services is **vertically integrated**. Nonvertically integrated firms buy the inputs or services they need for their production or distribution processes from other firms. A nonintegrated firm may write long-term, binding contracts with the firms with which it deals, in which it specifies not only price, but also other terms or forms of behavior. Contractual restraints on nonprice terms are called **vertical restrictions** (or restraints). For example, manufacturers commonly restrict their distributors by limiting their sales territories, setting inventory requirements, and, where legal, setting the minimum retail price they can charge.

Some firms choose to vertically integrate and perform all production and distribution activities themselves. Most firms partially vertically integrate. For example, they may produce goods, but rely on others to market them. A restaurant that bakes its own pies instead of buying them ready-made is partially integrated.

Some firms are not vertically integrated but buy from a small number of suppliers or sell through a small number of distributors. These firms often write complex contracts that restrict the actions of those with whom they deal. These vertical restrictions can approximate the outcome from vertically merging. Other firms buy in the open market from any number of anonymous firms. For example, they may buy wheat from a wheat broker without knowing who grew it or using any formal long-term contracts. Such firms place no restrictions on their suppliers.

An example of an integrated firm is Perdue, a prominent chicken supplier.[1] In the 1950s, Frank Perdue began to mix his own feed rather than buy what he felt was an inferior commercial mix. In 1961, he bought a soybean plant to make feed. In 1968, he bought the first Perdue processing plant so that his firm could kill, dress, and deliver chickens to market rather than rely on meat packers. In 1969, Perdue started appearing in his own television ads.

A firm's decision to vertically integrate, write complex contracts with vertical restrictions, or rely on markets is a basic strategic decision. It affects the subsequent pricing and promotional behavior of that firm and other related firms. Chapter 2 notes that a firm may choose to vertically integrate because it is cost effective to do so. This chapter expands on that analysis and examines vertical restrictions. Vertical restrictions between manufacturers and distributors are of particular interest and have been the subject of lengthy antitrust litigation. This chapter explores the procompetitive as well as the anticompetitive reasons for such restrictions.

The analysis begins by examining why some firms vertically integrate, whereas others do not. That analysis provides a story of the life-cycle of firms, in which they integrate at certain times and not at others. We then examine how some firms use vertical restrictions to achieve many of the advantages of vertical integration. Finally, we present some empirical evidence on franchising, an increasingly important vertical relationship, and the motives for vertical integration and vertical restrictions.

We analyze four key issues:

1. Why do firms vertically integrate? Why not rely on the market (other firms) to supply inputs and distribute products?
2. What should public policy be toward vertical integration? We know that horizontal mergers sometimes have anticompetitive effects; is the same true for vertical mergers?
3. Why do some manufacturers establish vertical restraints that give their distributors some of their monopoly power?
4. What should public policy be toward vertical restraints? Do these restrictions necessarily hurt retailers and consumers?

The Reasons for and Against Vertical Integration

If you want something done right, do it yourself.
He is a slave of the greatest slave, who serves nothing but himself.

Most of the reasons that firms choose to vertically integrate have to do with reducing costs or eliminating a market externality. Firms choose the least costly approach: Only

[1] Glenn Plaskin, "How Perdue Found Success," *San Francisco Chronicle*, January 27, 1993:B4.

EXAMPLE 12.1 *Outsourcing*

Whether a firm performs a task itself or relies on the market depends on the relative costs. A firm may find that it can save money by having outsiders provide services that the firm originally performed. This divestiture of activities is called *outsourcing*.

Many industries use outside firms for specific activities, such as payments. A 2003 Payroll Manager's Report survey found that more than half the firms surveyed outsourced payroll activities. Also in 2003, Accenture, a consulting firm, reported that two-thirds of U.S. retail and commercial banks with assets of at least $3 billion outsourced one or more business functions.

Outsourcing is particularly common in high-tech industries. Ingram Micro Inc., the world's largest wholesale distributor of computers, builds and distributes personal computers for rival firms that sell one-third of U.S. computers, including Acer, Apple, Hewlett-Packard, and IBM. U.S. high-tech firms are expected to outsource 1 in 10 jobs to low-cost emerging markets by the end of 2004. Worldwide, one-fifth of major firms are outsourcing their programming projects—often to India. The motive is cost savings. A programmer in Ireland may cost the firm up to 10 times as much as a comparably skilled programmer in India. Other countries that are effectively bidding for software work include Canada, China, Mexico, the Philippines, Russia, and Singapore.

Governments also rely on outside firms. Accenture reported in 2003 that 90 percent of government executives in 23 governments in Asia, Europe, North America, and South America outsource various functions. New Jersey purchases its welfare processing from an Arizona firm with a call center in Bombay, India. An advisory panel urged the Japanese government to outsource part of the management of a test module for the International Space Station to cut costs.

Even colleges and universities use outside firms for services such as janitorial, accounting, and teaching. Indeed, we once investigated the outsourcing of pithy sayings for this book, but decided against it when we considered the possible pithfalls.

Sources: Saul Hansell, "Is This the Factory of the Future?" *New York Times*, July 26, 1998, Section 3:1, 12, 13; Jon Surmacz, "Offshore Outsourcing Still Popular Despite Political Tensions," *CIO Metrics*, July 17, 2002; "Two-Thirds of U.S. Banks Outsource One or More Functions," *Business Wire*, February 24, 2003; "Vast Majority of Government Executives Report Outsourcing 'Important' or 'Critical' Activities, Accenture Report Finds," *Financial News*, May 15, 2003; "Exclusive ONR Survey," *2003 IOMA Payroll Manager's Report*, June 2003; "Cheap Labor at America's Expense," *Insight on the News*, June 9, 2003:32; "Japan to Outsource Management of Space Module Kibo," *BBC Monitoring International Reports*, June 25, 2003; "One Out of 10 Jobs at US Tech Firms to Go Offshore by 2004," *Agence France Presse*, July 29, 2003.

if a firm can perform most of the necessary production steps less expensively than if it relied on other firms does it vertically integrate. In general, a firm needs a good reason to vertically integrate because integration can involve substantial costs. In some cases, firms can avoid integration by having outside firms perform some functions for them (Example 12.1) or by using detailed contracts (Example 12.2).

EXAMPLE 12.2 *Preventing Holdups*

As Central and Eastern European countries convert from communism to capitalism, they are confronting serious transition problems. Communist Central and Eastern European countries had relatively few large, highly vertically integrated firms. The transition to capitalism has caused disruptions of traditional exchange systems, and the restructuring of firms both upstream and downstream has led to serious contracting problems.

One major problem is that contract terms are hard to enforce in these countries at this time. When contract terms cannot be enforced (or are not fully specified), holdup problems occur, typically leading to underinvestment in relation-specific capital.

Gow and Swinnen (1998) carefully examined sugar processors in Slovakia. A typical holdup problem is delayed payments by food processors to farmers. By delaying payment, processors effectively obtained interest-free loans and reduced their indebtedness due to high rates of inflation. According to surveys in 1994 and 1995, the average delay in payments for delivered products was 94 days—77 days for commercial farms and more than 100 days for state farms.

Faced with these adverse conditions, some farmers left the market, while others reduced their investments in land, equipment, and seed. Consequently, the amounts of beet sugar available to processors fell.

Given the large number of farmers involved, vertical integration was not feasible, even with foreign direct investment. Instead, long-term contracts and the establishment of trust through long-term relationships are used by Juhocukor a.s. in Slovakia to deal with holdup problems and to encourage investment in relationship-specific assets. Juhocukor a.s. is a subsidiary of Eastern Sugar BV, a firm that also owns operations in other Central and Eastern European countries, including the Czech Republic and Hungary. In 1993, Eastern Sugar BV purchased a 51 percent stake in

There are at least three possible costs of vertical integration. First, the cost of supplying its own factors of production or distributing its own product may be higher for a firm that vertically integrates than for one that depends on competitive markets, which serve these needs efficiently. Second, as a firm gets larger, the difficulty and cost of managing it increase. The advantage of dealing with a competitive market is that someone else supervises production. Third, the firm may face substantial legal fees to arrange to merge with another firm. For example, lawyers may be used to defend the merger before the U.S. Federal Trade Commission or the U.S. Department of Justice.

Because of these costs, firms vertically integrate only if the benefits outweigh the costs. Six major advantages to integrating are[2]

[2] See Perry (1989) for an excellent survey that discusses these and other explanations.

Juhocukor a.s. from Slovakia's privatization program. Later, it increased its holding to 76 percent and started a four-year development program to inject capital.

Before the takeover, Juhocukor a.s. had a reputation for late payments to farmers. To ensure sufficient high-quality sugar beet deliveries and to stimulate farm investment, the new management made three changes. First, it paid contracts on time at a price above that of the rest of the market and paid a premium for high sugar content. Second, it started a development program that included increasing processing and production productivity; made available high-quality seeds, fertilizer, and harvesters to farmers; and extended financing to growers. Third, it initiated a two-year information and media campaign directed at farmers and the agricultural community, explaining what the firm's long-term contract offered and how the contract would be beneficial for growers.

In short, the new management worked to establish credibility in a number of ways. By providing prompt payments and above-market prices, it showed that it was not going to engage in further holdup behavior. The new firm also signaled that it wanted long-term contracts and relationships with farmers based on mutual trust. By making large investments in its own plant, it signaled it was in the business for the long haul. Similarly, the firm could only benefit by helping supply farmers with financing and providing information, better-quality seeds, fertilizer, and harvesters if it continued to buy from the farmers in the long run. Finally, the firm increased quality by offering premiums and penalties for quality variations in sugar content of the beets.

The program worked. From 1992 to 1997, the amount of land the firm had under contract rose by 91 percent, yields per hectare increased by 140 percent, sugar content rose 119 percent, and total sugar production climbed 234 percent. These increases were not in response to market conditions as market prices were stable during this period, and the increases were large compared to those of other firms, which have since attempted to imitate these programs, with limited success.

1. *Lower transaction costs:* A firm may lower its transaction costs by vertically integrating. For example, the transaction costs of buying from or selling to other companies are avoided.
2. *Assure supply:* A firm may vertically integrate to assure itself a steady supply of a key input. To do so, the firm may *vertically integrate backwards,* buying or building the capacity to produce that input. Delivery problems may thus be reduced, because it is often easier to exchange information within a firm than between firms.
3. *Correct market failure:* A firm may vertically integrate to correct market failures due to externalities by internalizing those externalities. For example, by owning or controlling all its restaurants, McDonald's can ensure a uniform quality, which results in a positive reputation (externality). Wherever their travels bring

them, consumers know that they can expect a certain minimum quality at any of this chain's restaurants.

4. *Avoid government rules:* A firm may be able to avoid government restrictions, regulations, and taxes by vertically integrating. Examples of government interventions include price controls, regulations that restrict profit rates (see Chapter 20), and taxes on revenues or profits.
5. *Gain market power:* A firm may vertically integrate to better exploit or to create market power. For example, a sole supplier of a vital input may *vertically integrate forward,* buying the manufacturing firms, so as to monopolize the final product market and thereby increase its monopoly profits. Similarly, a firm may try to buy its sole supplier to increase combined profits. By vertically integrating, a firm may create or increase its monopoly profits by being able to price discriminate, eliminate competition, or foreclose entry.
6. *Eliminate market power:* A victim of another firm's market power may vertically integrate to eliminate that power. For example, around the turn of the century, dairy farmers contended that they faced a single processor that bought their milk at a low, monopsonistic price. To raise the price of milk, dairy farmers vertically integrated forward to form their own processors.

Integration to Lower Transaction Costs

A key reason why a firm performs productive activities itself rather than relying on other firms has to do with *transaction costs,* such as the expenses associated with writing and enforcing contracts (Williamson 1975, 1985; Alchian and Demsetz 1972; Klein, Crawford, and Alchian 1978). When such costs are high, a firm may engage in **opportunistic behavior**: taking advantage of another when allowed by circumstances. Each side may try to interpret the terms of a contract to its advantage, especially when terms are vague or even missing.

If contracts are simple (for example, a transaction involving one bushel of a specific variety of corn in Chicago on a particular date), opportunistic behavior is unlikely. The more unpredictable the future and the more complicated the contract, however, the harder it is to specify contractual terms. People have *bounded rationality:* a limited ability to enumerate and understand all future possibilities. In complicated contracts, it is often too difficult to specify all possible contingencies, and a signed contract may contain provisions that turn out to be undesirable to one of the parties.

Opportunities for exploitation are greater when one firm is dependent on another. For example, to respond to a sudden increase in demand, an automobile manufacturer needs more supplies. If there is only one supplier of a critical part, that supplier can raise its prices, and the auto manufacturer has nowhere to turn in the short run. Even when such complications and dependencies can be foreseen, it may be difficult to structure a contract that completely removes the incentives for either firm to behave opportunistically toward the other. For example, the Intel Corporation designs and sells many embedded control-function semiconductor chips, which are customized to do one job quickly and well. But buyers who start using these chips in their products have only one source because Intel does not allow other companies to produce the new chips. As one observer noted, "If they can get customers to make the transition, they now have captives."[3]

A firm chooses to perform activities itself rather than to rely on the market when transaction costs are likely to be high. Vertical integration transforms the monitoring problem from monitoring between firms to monitoring employees within the firm. Within a firm, a boss can coordinate the decisions of different divisions and can monitor workers in ways that are not possible when firms are completely independent. On the other hand, an employee on a fixed salary may work less hard than an owner of a subcontracting firm.

The desirability of integrating increases as the transaction costs of using the marketplace rise. There are four types of transactions in which transaction costs are likely to be substantial enough to make vertical integration desirable. They involve *specialized assets, uncertainty* that makes monitoring difficult, *information,* or *extensive coordination.*

Specialized Assets. A **specialized asset** is tailor-made for one or a few specific buyers. To illustrate why the use of specialized assets provides a reason to integrate, consider a supplier that has custom-designed its facility to suit a particular buyer's needs. That supplier will be at the mercy of the buyer should any disputes arise subsequent to the construction of the supplier's plant. In this case, we expect to see vertical integration because of asset specificity, which takes three main forms involving specific physical capital, specific human capital, and site-specific capital (Williamson 1985, 95–96).

Specific physical capital includes buildings and machines that can be used for only one or a few buyers. As an example, suppose that specific dies (molds used to make parts) are needed on a machine press to produce a particular part for one buyer. If the supplier that owns the machine press also owns the dies, there is a chance for opportunistic behavior: The supplier can raise the price, and the buyer may find it prohibitively expensive to switch suppliers in the short run. If the buyer owns the dies and has other firms bid to provide the machine-press services, no opportunistic problems arise. In this case, complete vertical integration is not necessary. Only partial or **quasi-vertical integration** (or quasi-integration), where the firm owns the specific physical asset (the dies) and not the entire supplying firm, is required to avoid opportunistic behavior. If the machine press itself is unique, however, this method cannot be used, and vertical integration may be necessary.

Ownership by the buyer diminishes the incentive for opportunistic behavior on both sides. For example, automobile manufacturers that rely on outside suppliers for

[3] Michael Slater, editor of *Microprocessor Report,* quoted in Don Clark, "Intel Corp. Planning New Chip Campaign," *San Francisco Chronicle,* April 2, 1988:B1, B20.