8th edition

URBAN ECONOMICS

Arthur O’Sullivan

Urban Economics

EIGHTH EDITION

Urban Economics

EIGHTH EDITION

Arthur O'Sullivan
Department of Economics
Lewis & Clark College

McGraw-Hill Irwin

The McGraw-Hill Companies

McGraw-Hill Irwin

URBAN ECONOMICS, EIGHTH EDITION

Published by McGraw-Hill/Irwin, a business unit of The McGraw-Hill Companies, Inc., 1221 Avenue of the Americas, New York, NY, 10020.

This book is printed on acid-free paper.

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

ISBN 978-0-07-351147-4
MHID 0-07-351147-1

Publisher: *Douglas Reiner*
Sponsoring Editor: *Daryl Bruflodt*
Director of Marketing & Sales: *Jennifer J. Lewis*
Project Manager: *Melissa M. Leick*
Buyer: *Susan K. Culbertson*
Design Coordinator: *Margarite Reynolds*
Media Project Manager: *Balaji Sundararaman*
Cover Designer: *Mary-Presley Adams*
Cover Image Credit: *Imagezoo/Getty Images*
Typeface: *10/12 Times Roman*
Compositor: *MPS Limited, a Macmillan Company*
Printer: *Quad/Graphics*

Library of Congress Cataloging-in-Publication Data

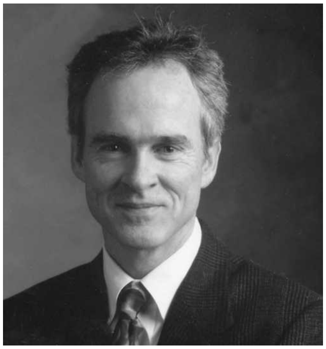
O'Sullivan, Arthur.
Urban economics / Arthur O'Sullivan.—8th ed.
p. cm.
Includes index.
ISBN 978-0-07-351147-4
1. Urban economics. I. Title.
HT321.O88 2012
330.9173'2—dc23

2011022545

www.mhhe.com

To Professor Ed Whitelaw, the most talented teacher I've ever known.
It has been almost 30 years since I've been in Ed's class, but whenever
I start thinking about how to teach some new material,
my first thought is "How would Ed present this material?"

About the Author



ARTHUR O'SULLIVAN is a professor of economics at Lewis and Clark College in Portland, Oregon. After receiving his B.S. degree in economics at the University of Oregon, he spent two years in the Peace Corps, working with city planners in the Philippines. He received his Ph.D. degree in economics from Princeton University in 1981 and taught at the University of California, Davis and Oregon State University, winning teaching awards at both schools. He is the Robert B. Pamplin Junior Professor of Economics at Lewis and Clark College in Portland, Oregon, where he teaches microeconomics and urban economics. He is the coauthor of the introductory textbook, *Economics: Principles and Tools*, currently in its seventh edition.

Professor O'Sullivan's research explores economic issues concerning urban land use, environmental protection, and public policy. His articles appear in many economics journals, including *Journal of Urban Economics, Regional Science and Urban Economics, Journal of Environmental Economics and Management, National Tax Journal, Journal of Public Economics*, and *Journal of Law and Economics*.

Preface

This book is on urban economics, the discipline that lies at the intersection of geography and economics. Urban economics explores the location decisions of utility-maximizing households and profit-maximizing firms, and it shows how these decisions cause the formation of cities of different size and shape. Part I of the book explains why cities exist and what causes them to grow or shrink. Part II examines the market forces that shape cities and the role of government in determining land-use patterns. Part III looks at the urban transportation system, exploring the pricing and design of public transit systems and the externalities associated with automobile use (congestion, environmental damage, collisions). Part IV explores the economics of urban education and crime, two factors that play key roles in household location decisions. Part V explains the unique features of the housing market and examines the effects of government housing policies. The final part of the book explains the rationale for our fragmented system of local government and explores the responses of local governments to intergovernmental grants and the responses of taxpayers to local taxes.

The text is designed for use in undergraduate courses in urban economics and urban affairs. It could also be used for graduate courses in urban planning, public policy, and public administration. All of the economic concepts used in the book are covered in the typical intermediate microeconomics course, so students who have completed such a course will be able to move through the book at a rapid pace. For students whose exposure to microeconomics is limited to an introductory course— or who could benefit from a review of the concepts covered in an intermediate microeconomics course—I have provided an appendix ("Tools of Microeconomics") that covers the key concepts.

CHANGES FOR THE EIGHTH EDITION

The eighth edition improves on the previous edition in two ways. First, I've rewritten Chapter 11 (Urban Transit) to incorporate the most recent developments in economic theory, empirical results, and practical experience with transit systems. Included in the revised chapter is a thorough analysis of the rationale for transit subsidies and a discussion of the size of the socially efficient subsidy. In addition, the chapter has a full accounting of the relative costs of light rail versus buses.

The second improvement is a new chapter on education (Chapter 12). This chapter uses the education production function as a framework to explore the economics of K–12 education. The chapter identifies the key inputs to the production process—teachers, the home environment, and classroom peers. One of the insights from the production function is that teacher productivity varies significantly across teachers. For example, if we replace an average teacher with an above-average teacher for one year, the benefit is roughly $210,000. At the other end of the productivity scale, if we were to replace the bottom 8 percent of teachers with average teachers, aggregate earnings in the national economy would increase by roughly $112 trillion. The education chapter also looks at spending inequalities across schools and evaluates the effects of intergovernmental grants on spending and achievement inequalities.

WEB SITE

The Web site for the book (www.mhhe.com/osullivan8e) has the following resources.

- Color versions of the maps in the book
- Maps for other cities
- For each chapter
  - PowerPoint presentations, which include all the figures and tables from the text
  - Lecture notes
- A chapter, "The Core-Periphery Model of Regional Development," that presents some key ideas from economic geography
- A list of corrections. The author has a typo-bounty program that pays $5 to the first person to identify a particular error.

The instructors' version of the Web site also has model answers to the exercises in the book.

Acknowledgments

I am indebted to many people who read the book and suggested ways to improve the coverage and the exposition. In particular I would like to thank those instructors who participated in surveys and reviews that were indispensable in the development of the Eighth Edition of *Urban Economics*. The appearance of their names does not necessarily constitute their endorsement of the text or its methodology.

Oliver D. Cooke
Richard Stockton College of New Jersey

Jonathan Diskin
Earlham College

Kristie A. Feder
Bard College

Gary Frimann
Gavilan College

Anthony Joseph Greco
University of Louisiana-Lafayette

Peter E. Howe
Cazenovia College

Haydar Kurban
Howard University

Thomas J. Muench
State University of New York Stony Brook

Steven R. Nivin
St. Mary's University

Joseph Michael Pogodzinski
San Jose State University

Jeffrey Pompe
Francis Marion University

Margaret A Ray
University of Mary Washington

Jesse J. Richardson, Jr
Virginia Tech

Paul Rivera
Wheaton College

Frederica Shockley
California State University, Chico

John D. Wong
Wichita State University

In addition, dozens of instructors provided feedback and suggestions for earlier editions of the book.

Richard Arnott
Boston College

Randall Bartlett
Smith College

Charles Becker
Department of Economics, Duke University

Charles Berry
University of Cincinnati

Bradley Braun
University of Central Florida

Jerry Carlino
University of Pennsylvania

Paul Carrillo
George Washington University

Suparna Chakraborty
Dept. of Economics and Finance,
Zicklin School of Business,
Baruch College, CUNY

Brian J. Cushing
West Virginia University

Maria N. DaCosta
University of Wisconsin–Eau Claire

Joseph Daniel
University of Delaware

Minh Quang Dao
Eastern Illinois University

Gilles Duranton
University of Toronto

Steven Durlauf
University of Wisconsin

Ingrid Gould Ellen
Wagner School, New York University

Erwin F. Erhardt, III
University of Cincinnati

David Figlio
University of Oregon

Edward J. Ford
University of South Florida

Tom Fullerton
University of Texas at El Paso

Andrew Gold
Trinity College

Alan Day Haight
SUNY–Cortland

Bríd Gleeson Hanna
Rochester Institute of Technology

Julia L. Hansen
Western Washington University

Daryl Hellman
Northeastern University

Barry Hersh
Steven L. Newman Real Estate Institute,
Baruch College, City University of New York

Diane Hite
Auburn University

Bruce K. Johnson
Centre College

Christopher K. Johnson
University of North Florida

Stanley Keil
Ball State University

Sukoo Kim
Washington University

MaryJane Lenon, Ph.D.
Providence College, Providence, RI

James P. LeSage
Texas State University–San Marcos

Kenneth Lipner
Florida International University

Roxanne Ezzet-Lofstrom
University of Texas at Dallas

Vijay Mathur
Cleveland State University

Dr. Warren McHone
University of Central Florida

Kevin J. Murphy
Oakland University

James K. O'Toole
California State University, Chico

Bruce Pietrykowski
University of Michigan–Dearborn

Florenz Plassmann
Binghamton University, State University of New York

Michael J. Potepan
San Francisco State University

David A. Quart
NYU–Wagner Graduate School of Public Service

Steven Raphael
University of California, San Diego

Donald Renner
Minnesota State University

Jesse J. Richardson, Jr.
Virginia Tech

Craig Rogers
Canisius College

Jonathan Rork
Vassar College

Stuart S. Rosenthal
Syracuse University

Jeffrey Rous
University of North Texas

William A. Schaffer
Georgia Institute of Technology

Steve Soderlind
Saint Olaf College

Dean Stansel
Florida Gulf Coast University

Mary Stevenson
University of Massachusetts, Boston

Will Strange
University of Toronto

Edward F. Stuart
Northeastern Illinois University

Timothy Sullivan
Towson University

Jacques-Francois Thisse
Universite Catholique de Louvain-la-Neuve

Wendine Thompson-Dawson
University of Utah

Mark R. Wolfe
University of California, Berkeley

Anthony Yezer
George Washington University

King Yik
Idaho State University

John Yinger
Syracuse University

Arthur O'Sullivan

Brief Contents

Contents

CHAPTER 1

Introduction and Axioms of Urban Economics

Cities have always been the fireplaces of civilization, whence light and heat radiated out into the dark.

—Theodore Parker

I'd rather wake up in the middle of nowhere than in any city on earth.

—Steve McQueen

This book explores the economics of cities and urban problems. The quotes from Parker and McQueen reflect our mixed feelings about cities. On the positive side, cities facilitate innovation, production, and trade, so they increase our standard of living. On the negative side, cities are noisy, dirty, and crowded. As we'll see in the first part of the book, firms and people locate in cities because the obvious costs of being in a city are more than offset by subtle benefits of producing in close proximity to other firms and people. As we'll see later in the book, policies that combat urban problems such as congestion, pollution, and crime are likely to increase the vitality of cities, causing them to grow.

WHAT IS URBAN ECONOMICS?

The discipline of urban economics is defined by the intersection of geography and economics. Economics explores the choices people make when resources are limited. Households make choices to maximize their utility, while firms maximize their profit. Geographers study how things are arranged across space, answering the question, Where does human activity occur? Urban economics puts economics and geography together, exploring the geographical or location choices of utility-maximizing households and profit-maximizing firms. Urban economics also identifies inefficiencies in location choices and examines alternative public policies to promote efficient choices.

Urban economics can be divided into six related areas that correspond to the six parts of this book.

1. **Market forces in the development of cities.** The interurban location decisions of firms and households generate cities of different size and economic structure. We explore the issues of why cities exist and why there are big cities and small ones.
2. **Land use within cities.** The intraurban location decisions of firms and households generate urban land-use patterns. In modern cities, employment is spread throughout the metropolitan area, in sharp contrast to the highly centralized cities of just 100 years ago. We explore the economic forces behind the change from centralized to decentralized cities. We also use a model of neighborhood choice to explore the issue of segregation with respect to race, income, and educational level.
3. **Urban transportation.** We explore some possible solutions to the urban congestion problem and look at the role of mass transit in the urban transportation system. One issue is whether a bus system is more efficient than a light-rail system or a heavy-rail system like BART (San Francisco) or Metro (Washington).
4. **Crime and public policy.** We look at the problem of urban crime and show the links between crime and two other urban problems, poverty and low educational achievement.
5. **Housing and public policy.** Housing choices are linked to location choices because housing is immobile. We'll discuss why housing is different from other products and how housing policies work.
6. **Local government expenditures and taxes.** Under our fragmented system of local government, most large metropolitan areas have dozens of local governments, including municipalities, school districts, and special districts. In making location choices, households consider the mix of taxes and local public goods.

WHAT IS A CITY?

An urban economist defines an urban area as a geographical area that contains a large number of people in a relatively small area. In other words, an urban area has a population density that is high relative to the density of the surrounding area. This definition accommodates urban areas of vastly different sizes, from a small town to a large metropolitan area. The definition is based on population density because an essential feature of an urban economy is frequent contact between different economic activities, which is feasible only if firms and households are concentrated in a relatively small area.

The U.S. Census Bureau has developed a variety of geographical definitions relevant to urban economics. Since much of the empirical work in urban economics is based on census data, a clear understanding of these definitions is important.

The appendix to this chapter provides the details of the census definitions. The key census definitions, some of which are new for the 2000 Census, are as follows.

1. **Urban area:** A densely settled geographical area with a minimum population of 2,500 people and a minimum density of 500 people per square mile. In 2000, there were 3,756 urban areas in the United States.
2. **Urban population:** People living in urban areas. In 2000, the urban population was 79 percent of the total population.
3. **Metropolitan area:** A core area with a substantial population nucleus, together with adjacent communities that are integrated, in an economic sense, with the core area. To qualify as a metropolitan area, the minimum population is 50,000 people. In 2000, there were 361 metropolitan statistical areas in the United States.
4. **Micropolitan area:** A smaller version of a metropolitan area with a concentration of 10,000 to 50,000 people. In 2000, there were 559 micropolitan statistical areas in the United States.
5. **Principal city:** The largest municipality in each metropolitan or micropolitan statistical area. A municipality is defined as an area over which a municipal corporation exercises political authority and provides local government services such as sewage service, crime protection, and fire protection.

This book uses three terms to refer to spatial concentrations of economic activity: *urban area, metropolitan area,* and *city*. These three terms, which will be used interchangeably, refer to the economic city (an area with a relatively high population density that contains a set of closely related activities), not the political city. When referring to a political city, we will use the term *central city* or *municipality*.

WHY DO CITIES EXIST?

This is the fundamental question of urban economics. People need land to produce food and other resources, and living in dense cities separates us from the land where food is produced. As Bartlett (1998) points out, no other creatures in the animal world form anything like cities. Herbivores such as wildebeests and bison form larger herds but constantly migrate to fresh land to ensure a steady supply of food. Coral is concentrated in stationary reefs, but ocean currents provide a steady supply of food to the stationary coral. Perhaps the closest thing to a city in the natural world is a bee hive or an anthill. Eusocial insects such as bees and ants form colonies with thousands of inhabitants, with highly specialized castes—soldier ants, drones, breeders, nurses, and cleanup crews. In contrast with human cities, these insect agglomerations are closed to non-natives and not based on voluntary exchange.

Cities exist because human technology has created systems of production and exchange that seem to defy the natural order. Three conditions must be satisfied for a city to develop.

1. **Agricultural surplus.** People outside cities must produce enough food to feed themselves and city dwellers.

2. **Urban production.** City dwellers must produce something—goods or services—to exchange for food grown by rural workers.
3. **Transportation for exchange.** There must be an efficient transportation system to facilitate the exchange of food and urban products.

Figure 1–1 shows the share of people living in cities in the United States from 1800 to 2010. Over this period, the urban share increased from 6 percent to 82 percent, a remarkable transformation that also occurred in other parts of the world. As we'll see in the next three chapters of the book, the transformation of a rural society into an urban one occurred because technological advances increased the agricultural surplus (condition 1), increased the productivity of urban workers (condition 2), and increased the efficiency of transportation and exchange (condition 3).

Figure 1–2 shows urbanization rates for different regions around the world, with projections for the year 2030. In 1950, urbanization rates were relatively low in Africa and Asia, and highest in Oceania and North America. Between now and the year 2030, urbanization rates are expected to increase everywhere, with the largest increases in Africa and Asia. For the world as a whole, the urbanization rate was 30 percent in 1950 and is expected to double by the year 2030.

Table 1–1 (page 6) shows the population figures for the nation's 30 largest metropolitan areas. The New York area tops the list, followed by Los Angeles, Chicago, Dallas, and Philadelphia. The third column shows the percentage growth of each

FIGURE 1–1 Percent of U.S. Population in Urban Areas, 1800–2010

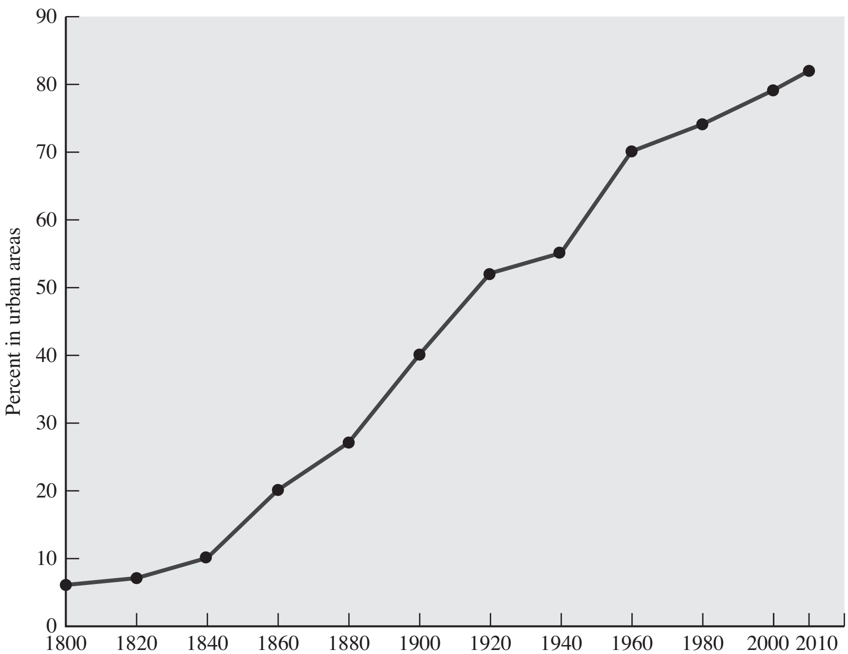
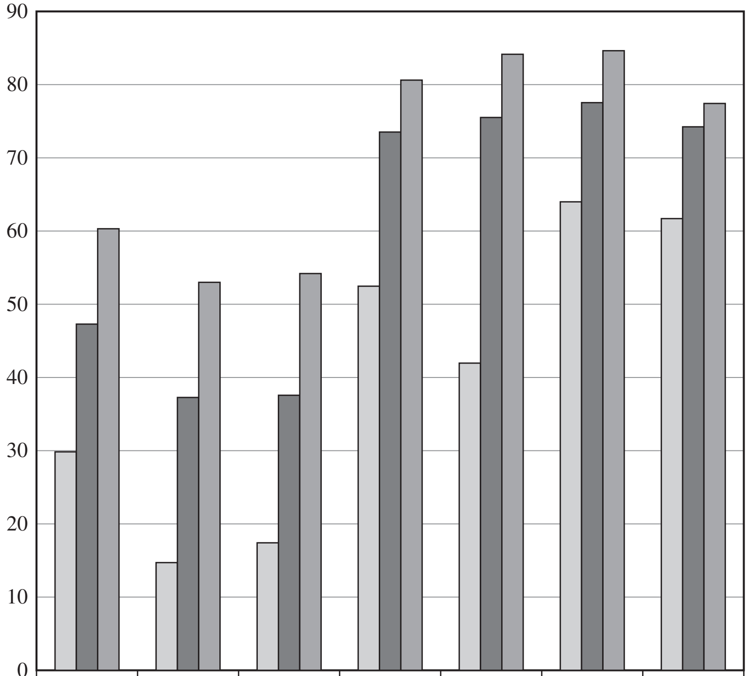


FIGURE 1–2 Urbanization Rates, by World Region, 1950–2030



	World	Africa	Asia	Europe	Latin America and the Caribbean	North America	Oceania
1950	29.8	14.7	17.4	52.4	41.9	63.9	61.6
2000	47.2	37.2	37.5	73.4	75.4	77.4	74.1
2030	60.2	52.9	54.1	80.5	84	84.5	77.3

Percent in urban areas

Source: United Nations: World Urban Prospects, 2001 Revision.

metropolitan area over the period 2000 to 2005. The most rapidly growing metropolitan areas were in the South, the Mountain States, and the West. In three metropolitan areas—Detroit, Pittsburgh, and Cleveland—population decreased over this period, continuing a two-decade trend of decreasing population. These metropolitan areas experienced large losses in manufacturing employment.

Table 1–2 (page 7) shows the population figures for the world's largest metropolitan areas outside the United States. The table shows actual populations in 1975 and 2005, and projected populations for the year 2015. Eight metropolitan areas, all of which are in the developing world, are expected to grow by at least 20 percent over the 10-year period. In contrast, three cities in the developed world (Tokyo,

TABLE 1–1 Largest Metropolitan Areas in the United States, 2009

Metropolitan Area	Population in 2009	Percentage Change 2000–2009	Rank
New York-Northern New Jersey-Long Island, NY-NJ-PA	19,069,796	4.1	1
Los Angeles-Long Beach-Santa Ana, CA	12,874,797	4.1	2
Chicago-Naperville-Joliet, IL-IN-WI	9,580,567	5.3	3
Dallas-Fort Worth-Arlington, TX	6,447,615	24.9	4
Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	5,968,252	4.9	5
Houston-Sugar Land-Baytown, TX	5,867,489	24.4	6
Miami-Fort Lauderdale-Pompano Beach, FL	5,547,051	10.8	7
Washington-Arlington-Alexandria, DC-VA-MD-WV	5,476,241	14.2	8
Atlanta-Sandy Springs-Marietta, GA	5,475,213	28.9	9
Boston-Cambridge-Quincy, MA-NH	4,588,680	4.5	10
Detroit-Warren-Livonia, MI	4,403,437	–1.1	11
Phoenix-Mesa-Scottsdale, AZ	4,364,094	34.2	12
San Francisco-Oakland-Fremont, CA	4,317,853	4.7	13
Riverside-San Bernardino-Ontario, CA	4,143,113	27.3	14
Seattle-Tacoma-Bellevue, WA	3,407,848	12.0	15
Minneapolis-St. Paul-Bloomington, MN-WI	3,269,814	10.1	16
San Diego-Carlsbad-San Marcos, CA	3,053,793	8.5	17
St. Louis, MO-IL	2,828,990	4.8	18
Tampa-St. Petersburg-Clearwater, FL	2,747,272	14.7	19
Baltimore-Towson, MD	2,690,886	5.4	20
Denver-Aurora-Broomfield, CO	2,552,195	17.1	21
Pittsburgh, PA	2,354,957	–3.1	22
Portland-Vancouver-Beaverton, OR-WA	2,241,841	16.3	23
Cincinnati-Middletown, OH-KY-IN	2,171,896	8.1	24
Sacramento–Arden-Arcade–Roseville, CA	2,127,355	18.4	25
Cleveland-Elyria-Mentor, OH	2,091,286	–2.6	26
Orlando-Kissimmee, FL	2,082,421	26.6	27
San Antonio, TX	2,072,128	21.1	28
Kansas City, MO-KS	2,067,585	12.6	29
Las Vegas-Paradise, NV	1,902,834	38.3	30

Source: U.S. Census Bureau, "Table 1–Annual Estimates of the Population of Metropolitan and Micropolitan Statistical Areas: April 1, 2000 to July 1, 2009 (CBSA-EST2009-01)," March 2010.

Osaka, and Paris) are expected to grow slowly. In the United States, New York is expected to grow 6 percent over the period, and Los Angeles is expected to grow 7 percent.

Figure 1–3 (page 8) shows the time trend of large urban agglomerations in the world, defined as metropolitan areas with at least 1 million people. The figure distinguishes between cities in the developed and less developed regions. In 1970, the two types of regions had roughly the same number of large cities. By 1996, however, the number of large cities in the less developed regions nearly doubled, and by 2015 there will be roughly four times as many large cities in less developed regions.

TABLE 1–2 Populations and Projected Populations of Large World Cities

Metropolitan Area	Nation	Population 1975 (million)	Population 2005 (million)	Population 2015 (million)	Percent Change 2005–2015
Tokyo	Japan	26.6	35.2	35.5	1
Ciudad de México (Mexico City)	Mexico	10.7	19.4	21.6	11
Sao Paulo	Brazil	9.6	18.3	20.5	12
Mumbai (Bombay)	India	7.1	18.2	21.9	20
Delhi	India	4.4	15.0	18.6	24
Shanghai	China	7.3	14.5	17.2	19
Kolkata (Calcutta)	India	7.9	14.3	17.0	19
Jakarta	Indonesia	4.8	13.2	16.8	27
Buenos Aires	Argentina	8.7	12.6	13.4	7
Dhaka	Bangladesh	2.2	12.4	16.8	35
Karachi	Pakistan	4.0	11.6	15.2	31
Rio de Janeiro	Brazil	7.6	11.5	12.8	11
Osaka-Kobe	Japan	9.8	11.3	11.3	0
Al-Qahirah (Cairo)	Egypt	6.4	11.1	13.1	18
Lagos	Nigeria	1.9	10.9	16.1	48
Beijing	China	6.0	10.7	12.9	20
Manila	Philippines	5.0	10.7	12.9	21
Moskva (Moscow)	Russian Federation	7.6	10.7	11.0	3
Paris	France	8.6	9.8	9.9	0
Istanbul	Turkey	3.6	9.7	11.2	15

Source: United Nations. Urban Agglomerations 2005.

THE FIVE AXIOMS OF URBAN ECONOMICS

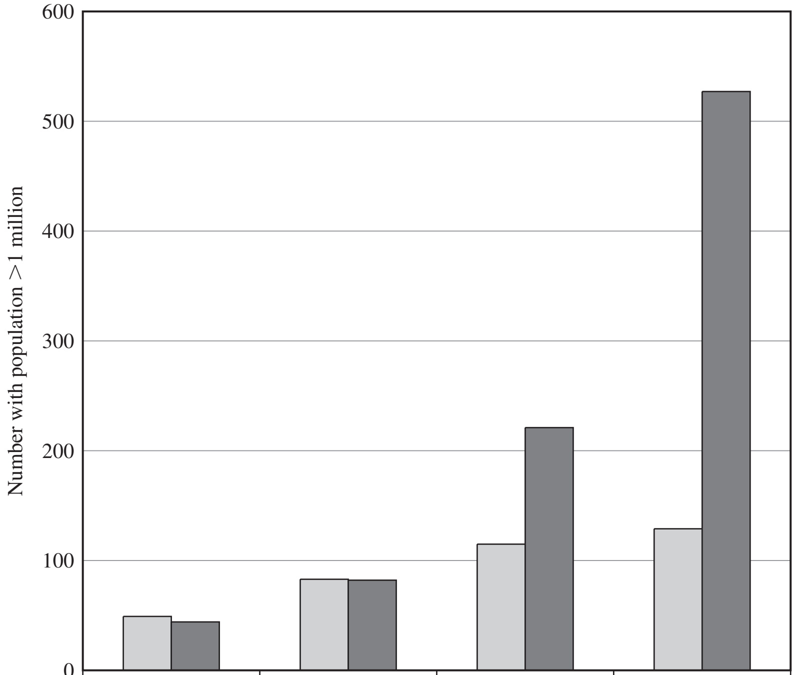
Urban economics explores the location choices of households and firms, and so it is natural to assume that people and firms are mobile. Of course, people don't instantly change their workplaces and residences when circumstances change; therefore, a model of perfect mobility tells us more about long-term changes than short-term ones. The average household changes its residence every seven years, meaning that about 14 percent of the population moves every year. Although most models of urban economics assume perfect mobility, there are exceptions, and we will highlight the analysis that assumes less than perfect mobility.

In this part of the chapter, we introduce five axioms of urban economics. An axiom is a self-evident truth, something that most people readily understand and accept. For our purposes, "most people" are people who have taken at least one course in economics. The five axioms lie at the heart of urban economics and together provide a foundation for the economic models of location choices. As you go through the book, these five axioms will appear repeatedly.

1. Prices Adjust to Achieve Locational Equilibrium

A locational equilibrium occurs when no one has an incentive to move. Suppose that you and Bud are competing for two rental houses, one along a beautiful beach

FIGURE 1–3 The Number of Large Agglomerations in the World, 1950–2015



	1950	1970	1996	2015
More developed regions	49	83	115	129
Less developed regions	44	82	221	527

Source: United Nations World Population Prospects (New York: United Nations, 2001).

and one along a noisy highway. If the two houses have the same price (the same monthly rent), you would prefer the beach house, and so would Bud. Flipping a coin and giving the beach house to the winner wouldn't generate a locational equilibrium because the unlucky person in the highway house would have an incentive to move to the more desirable house.

Locational equilibrium requires a higher price for the beach house. To eliminate the incentive to move, the price of the beach house must be high enough to fully compensate for the better environment. The question is, How much money are you willing to sacrifice to live on the beach? If your answer is $300 and Bud agrees, then the equilibrium price of the beach house will be $300 higher than the price of the highway house. In general, prices adjust to generate the same utility level in different environments, getting people to live in both desirable and undesirable locations.

The same sort of economic forces operate in the labor market. Workers compete for jobs in desirable locations, causing lower wages in more desirable locations. Suppose you are competing with Ricki for two jobs, one in Dullsville and one in Coolsville, a city with a more stimulating social environment. If a $500 gap in the monthly wage fully compensates for the difference in the social environment, the equilibrium wage will be $500 lower in Coolsville. The two workers will be indifferent between the two cities because a move to Coolsville means a $500 wage cut. In the labor market, wages adjust to get people to work in both desirable and undesirable environments.

The price of land also adjusts to ensure locational equilibrium among firms. Office firms compete for the most accessible land in a city, and land at the center is the most accessible and thus the most expensive. In equilibrium, office firms on less accessible land far from the center pay lower prices for land, and can be just as profitable as firms on the most accessible land.

2. Self-Reinforcing Effects Generate Extreme Outcomes

A self-reinforcing effect is a change in something that leads to additional changes in the same direction. Consider a city where the sellers of new automobiles are initially spread evenly throughout the city. If one seller relocates next to another seller on Auto Road, what happens next? Auto consumers compare brands before buying, and the pair of sellers on Auto Road will facilitate comparison shopping and thus attract buyers. The increased consumer traffic on Auto Road will make it an attractive site for other auto sellers, so they will move too. The ultimate result is an "auto row," a cluster of firms that compete against one another, yet locate nearby.

Self-reinforcing changes also happen in the location decisions of people. Suppose artists and creative types are initially spread out evenly across a dozen cities in a region. If by chance one city experiences an influx of artists, its creative environment will improve as artists (1) are exposed to more ideas and fabrication techniques and (2) can share studios, print shops, tool suppliers, and other facilities. The cluster of artists will attract other artists from the region, causing a concentration of artistic production in one city. In recent decades, cities that have attracted artists and creative folks have experienced relatively rapid growth (Florida, 2002).

3. Externalities Cause Inefficiency

In most transactions, the costs and benefits of the exchange are confined to the individual buyer and seller. The consumer pays a price equal to the full cost of producing the good, so no one else bears a cost from the transaction. Similarly, the consumer is the only person to benefit from the product. In contrast, an externality occurs when some of the costs or benefits of a transaction are experienced by someone other than the buyer or seller, that is, someone *external* to the transaction.

An external cost occurs when a consumer pays a price that is less than the full cost of producing a product. The price of a product always includes the costs of the labor, capital, and raw materials used to produce the product, but it usually does not include the environmental costs of producing the product. For example, if burning gasoline in automobiles generates air pollution, part of the cost of driving is borne by people who breathe dirty air. Similarly, when you enter a crowded highway, you slow down everyone else, meaning that other drivers bear a cost.

An external benefit occurs when a product purchased by one person generates a benefit for someone else. For example, painting my peeling house improves the appearance of my neighborhood, increasing the value of my neighbor's house as well as mine. Education generates external benefits because it improves communication and thinking skills, making a person a better team worker. In other words, some of the benefits of education are experienced by a person's fellow workers, who become more productive and thus earn higher wages.

When there are external costs or benefits, we do not expect the market equilibrium to be socially efficient. In the case of external cost, people pay less than the full social cost of an action like driving, so they drive too much. In the case of external benefit, people get less than the full social benefit from an action like education, so they stop short of the socially efficient level of education. As we'll see later in the book, cities have all sorts of external costs and benefits. In many cases there is a simple solution: Internalize the externality with a tax or a subsidy, and let individuals, who then bear the full social cost and benefits of their actions, decide what to do.

4. Production Is Subject to Economies of Scale

Economies of scale occur when the average cost of production decreases as output increases. For most products, if we start with a relatively small production operation and double all inputs, the average cost of production decreases. In the jargon of economics, when the long-run average cost curve is negatively sloped, we say that there are scale economies in production. Scale economies occur for two reasons:

- **Indivisible inputs.** Some capital inputs are "lumpy" and cannot be scaled down for small operations. As a result, a small operation has the same indivisible inputs as a large operation. For example, to manufacture frisbees you need a mold, whether you produce one frisbee per day or a thousand. Similarly, to produce microprocessors you need a clean room and other expensive equipment, whether you produce one processor per day or a thousand. As output increases, the average cost decreases because the cost of the indivisible input is spread over more output.
- **Factor specialization.** In a small one-person production operation, a worker performs a wide variety of production tasks. In a larger operation with more workers, each worker specializes in a few tasks, leading to higher productivity because of continuity (less time is spent switching from one task to another) and proficiency (from experience and learning). The notion of factor specialization is captured in the old expression, "A jack of all trades is master of none." Adding to this expression, we can say that a specialized worker is a master of one task.

As we'll see later in the book, scale economies play a vital role in urban economies. In fact, as we'll see in Chapter 2, if there are no scale economies, there will be no cities. It is costly to transport products from a production site to consumers, so centralized production in cities will be sensible only if there is some advantage that more than offsets transport costs.

The extent of scale economies in production varies across products. Microprocessors are produced in $5 billion fabrication facilities with a highly specialized workforce performing hundreds of complex tasks, resulting in large scale economies in production. In contrast, pizza is produced with a $5,000 pizza oven with just a few production tasks, so scale economies are exhausted sooner. In general, the extent of scale economies is determined by the lumpiness of indivisible inputs and the opportunities for factor specialization.

5. Competition Generates Zero Economic Profit

When there are no restrictions on the entry of firms into a market, we expect firms to enter the market until economic profit is zero. Recall that economic profit equals the excess of total revenue over total economic cost, where economic cost includes the opportunity costs of all inputs. Two key components of economic costs are the opportunity cost of the entrepreneur's time and the opportunity cost of funds invested in the firm. For example, suppose an entrepreneur could earn $60,000 in another job and invests $100,000 in the firm, taking the money out of a mutual fund that earns 8 percent. The economic cost of the firm includes $60,000 in time cost and $8,000 in investment cost. Once we account for all the opportunity costs, the fact that economic profit is zero means that a firm is making enough money to stay in business, but not enough for other firms to enter the market. Earning zero economic profit means earning "normal" accounting profit.

In urban economics, competition has a spatial dimension. Each firm enters the market at some location, and the profit of each firm is affected by the locations of other firms. Spatial competition looks a lot like monopolistic competition, a market structure in which firms sell slightly differentiated products in an environment of unrestricted entry. Although this sounds like an oxymoron such as "tight slacks" and "jumbo shrimp," the words are revealing. Each firm has a monopoly for its differentiated product, but unrestricted entry leads to keen competition for consumers who can easily switch from one differentiated product to another. With spatial competition, each firm has a local monopoly in the area immediately surrounding its establishment, but unrestricted entry leads to keen competition. Firms will continue to enter the market until economic profit drops to zero.

WHAT'S NEXT?

This introductory chapter sets the stage for the economic analysis of cities in the rest of the book. Here are some of the big questions we'll address in coming chapters:

- Why do cities exist?
- Are cities too big or too small?
- What causes urban economic growth?
- Why is employment in modern cities so widely dispersed?
- Why is there so much segregation with respect to race and income?

- Why do economists advocate a tax of about 7 cents per mile for all driving and about 27 cents per mile for driving on congested roads?
- Why do so few people take mass transit?
- What are the key inputs in the education production function?
- Why is crime higher in cities?
- Why does the typical metropolitan area have dozens of municipalities?

In answering these and other questions, we will use the five axioms of urban economics. In addition, we will use a number of economic models to explore the spatial aspects of decision making. It's worth noting that much of the analysis in the book reflects advances in urban economics in the last 10 to 15 years, in both theoretical modeling and empirical analysis.

REFERENCES AND ADDITIONAL READING

1. Audretsch, David, and Maryann Feldman. ''Knowledge Spillovers and the Geography of Innovation,'' Chapter 61 in *Handbook of Regional and Urban Economics 4: Cities and Geography,* eds. Vernon Henderson and Jacques-Francois Thisse. Amsterdam: Elsevier, 2004.
2. Bartlett, Randall. *The Crisis of American Cities.* Armonk, NY: Sharp, 1998.
3. U.S. Government. "Standards for Defining Metropolitan and Micropolitan Statistical Areas." *Federal Register* 65, no. 249 (December 17, 2000).
4. Florida, Richard. *The Rise of the Creative Class*. New York: Basic Books, 2002.

Appendix: Census Definitions

The U.S. Census Bureau has developed a variety of geographical definitions relevant to urban economics. Since much of the empirical work in urban economics is based on census data, a clear understanding of these definitions is important. This appendix provides the details of the census definitions.

URBAN POPULATION

The first three definitions deal with the urban population and are based on the census block, the smallest geographical unit in census data. A *census block* is defined as an area bounded on all sides by visible features (streets, streams, or tracks) or invisible features (property lines or political boundaries). The typical census block has between a few dozen and a few hundred residents. A *block group* is a group of contiguous census blocks. There are two types of urban areas:

1. **Urbanized area.** An *urbanized area* is a densely settled core of census block groups and surrounding census blocks that meet minimum population density

requirements. In most cases, the density requirement is 1,000 people per square mile for the core block groups and 500 people per square mile for the surrounding blocks. Together, the densely settled blocks must encompass a population of at least 50,000 people. In 2000, there were 464 urbanized areas in the United States.

2. **Urban clusters.** An *urban cluster* is a scaled-down version of an urbanized area. The total population of the census blocks that make up an urban cluster is between 2,500 and 50,000 people. In 2000, there were 3,112 urban clusters in the United States.
3. **Urban population.** The Census Bureau defines the nation's *urban population* as all people living in urbanized areas and urban clusters. Based on this definition, 79 percent of the population lived in urban areas in 2000.

METROPOLITAN AND MICROPOLITAN STATISTICAL AREAS

The census bureau has a long history of changing its definitions of metropolitan areas. The general idea is that a metropolitan area includes a core area with a substantial population nucleus, together with adjacent communities that are integrated, in an economic sense, with the core area. Over the years, the labels for metropolitan areas have changed from standard metropolitan area (SMA) in 1949, to standard metropolitan statistical area (SMSA) in 1959, to metropolitan statistical area (MSA) in 1983, to metropolitan area (MA) in 1990, which referred collectively to metropolitan statistical areas (MSAs), consolidated metropolitan statistical areas (CMSAs—the largest metropolitan areas), and primary metropolitan statistical areas (PMSAs—parts of CMSAs).

The new label for areas considered metropolitan, implemented in 2000, is *core based statistical area* (CBSA). Each CBSA contains at least one urban area (either an urbanized area or an urban cluster) with at least 10,000 people and is designated as either a metropolitan area or a micropolitan area.

1. **Metropolitan area.** A *metropolitan statistical area* includes at least one urbanized area with at least 50,000 people.
2. **Micropolitan area.** A *micropolitan statistical area* includes at least one urban cluster of between 10,000 and 50,000 people.

In 2000, there were 361 metropolitan statistical areas and 559 micropolitan statistical areas in the United States.

The building blocks for metropolitan and micropolitan areas are counties. For a particular CBSA, central counties are ones in which at least 5,000 people or 50 percent of the population resides within urban areas with at least 10,000 people. Additional outlying counties are included in the CBSA if they meet minimum thresholds of commuting rates to or from the central counties. Specifically, at least 25 percent of workers in an outlying county must work in one of the central counties, or at least 25 percent of the jobs in an outlying county must be filled by residents of one of the central counties.

Together CBSAs contain 93 percent of the nation's population, with 83 percent in metropolitan areas and 10 percent in the smaller micropolitan areas. The percentage of the population in CBSAs (93 percent) exceeds the percentage in urban areas (79 percent) because CBSAs encompass entire counties, including areas outside urban areas (defined by the smallest geographical unit, the census block).

PRINCIPAL CITY

The largest municipality in each metropolitan or micropolitan statistical area is designated a *principal city*. Additional cities qualify as "principal" if they meet minimum requirements for population size (at least 250,000 people) and employment (at least 100,000 workers). The title of each metropolitan or micropolitan statistical area consists of the names of up to three of its principal cities and the name of each state into which the metropolitan or micropolitan statistical area extends. For example, the name for Minneapolis metropolitan area is Minneapolis-St. Paul-Bloomington, MN-WI, indicating that it includes parts of two states with two other municipalities large enough to merit listing. For most metropolitan areas, the label includes only one principal city. About a dozen large metropolitan areas are divided into smaller groupings of counties called metropolitan divisions.

PART ONE

Market Forces in the Development of Cities

In a market economy, individuals exchange their labor for wage income, which they use to buy consumer goods and services. How do these market transactions affect cities? As we'll see in Chapter 2, cities exist because of the benefits of centralized production and exchange. We'll look at the rationale for the development of cities based on trade, production, and processing raw materials. Chapter 3 explores agglomeration economies, the economic forces that cause firms to cluster in cities to share the suppliers of intermediate inputs, share a labor pool, get better skills matches between workers and firms, and share knowledge. Chapter 4 explores the economic forces behind the development of cities of different size and scope. We'll look at how worker utility varies with city size and see why the equilibrium city size often exceeds the optimum size. Chapter 5 explores the sources of urban economic growth (increases in per-capita income) and urban employment growth. It also addresses the question of who benefits from employment growth and describes some of the techniques used by economists to predict future employment growth.

CHAPTER 2

Why Do Cities Exist?

Nobody ever saw a dog make a fair and deliberate exchange of one bone for another with another dog.

—Adam Smith

Cities exist because individuals are not self-sufficient. If each of us could produce everything we consumed and didn't want much company, there would be no reason to live in dirty, noisy, crowded cities. We aren't self-sufficient, but instead specialize in a labor task—writing software, playing the accordion, performing brain surgery—and use our earnings to buy the things we don't produce ourselves. We do this because labor specialization and large-scale production allow us to produce and consume more stuff. As we'll see in this chapter, production happens in cities, so that's where most of us live and work. By living and working in cities, we achieve a higher standard of living but put up with more congestion, noise, and pollution.

To explain why cities exist, we'll start with a model that implies that they don't. In the model of backyard production, every consumer is a producer, and all production occurs in backyards (or apartment roofs). In other words there is no need for concentrated production or population. As we drop the assumptions of the backyard-production model, the new models imply that cities will develop. In other words, the short list of assumptions in the model identifies the key factors behind the development of cities.

A REGION WITHOUT CITIES—BACKYARD PRODUCTION

Consider a region that produces and consumes two products, bread and shirts. People use the raw materials from land (wool and wheat) to produce the two consumer products. The following assumptions eliminate the possibility of cities.

- **Equal productivity.** All land is equally productive in producing wheat and wool, and all workers are equally productive in producing shirts and bread.
- **Constant returns to scale in exchange.** The unit cost of exchange (the cost of executing one transaction, including transportation cost) is constant, regardless of how much is exchanged.

- **Constant returns to scale in production.** The quantity of shirts produced per hour is constant, regardless of how many shirts a worker produces. The same is true for bread production.

Together these assumptions eliminate the possibility of exchange and guarantee that each household will be self-sufficient. If a person were to specialize in bread and then trade some bread for shirts, she would incur a transaction cost equal to the product that could be produced in the time required to execute the trade. Under the assumption of equal productivity, there is no benefit from specialization because everyone is equally productive. Under the assumption of constant returns to scale, there is no benefit from producing shirts in factories because an individual is just as efficient as a shirt factory. In sum, exchange has costs without any benefits, so every household will be self-sufficient, producing everything it consumes.

The absence of exchange guarantees a uniform distribution of population. If population were concentrated at some location, competition for land would bid up its price. People in the city would pay a higher price for land without any compensating benefit, so they would have an incentive to leave the city. In the locational equilibrium, the price of land would be the same at all locations, and population density would be uniform. Recall the first axiom of urban economics:

Prices adjust to ensure locational equilibrium

In this case, all sites are equally attractive, so locational equilibrium requires the same price of land at all locations.

A TRADING CITY

Now that we have a short list of assumptions under which cities don't develop, let's drop the assumptions, one by one, and see what happens. We'll start by dropping the assumption of equal productivity for all workers. Suppose households in the North are more productive in producing both bread and shirts. This could result from differences in soil conditions, climate, or worker skills. Table 2–1 shows the output per hour for the two regions. While each worker in the South can produce one shirt or one loaf per hour, workers in the North are twice as productive in producing bread and six times as productive in producing shirts.

TABLE 2–1 Comparative Advantage

	North		South	
	Bread	Shirts	Bread	Shirts
Output per hour	2	6	1	1
Opportunity cost	3 shirts	1/3 loaf	1 shirt	1 loaf

Comparative Advantage and Trade

A region has a comparative advantage in producing a particular product if it has a lower opportunity cost. For every shirt produced, the North sacrifices 1/3 loaf of bread, so that's the opportunity cost of a shirt. In the South, the opportunity cost of a shirt is one loaf. The North has a lower opportunity cost for shirts, so it has a comparative advantage in producing shirts. It is sensible for the North to specialize in shirts (and not produce any bread) because, although the North is twice as productive as the South in producing bread, the North is *six* times as productive in producing shirts.

Comparative advantage may lead to specialization and trade. Suppose the two regions are initially self-sufficient, with each household producing all the bread and shirts it consumes. Table 2–2 shows what happens if a North household switches one hour from bread to shirt production, and a South household goes the other direction, switching two hours from shirt to bread production. The first row shows the changes in production: −2 loaves and +6 shirts for North; +2 loaves and −2 shirts for South. As shown in the second and third rows, if the households exchange two loaves and four shirts, each has a gain from trade of two shirts. After specialization and exchange, each household has just as much bread as before and two additional shirts.

What about transaction costs? The transaction cost is the opportunity cost of the time required to exchange products and is equal to the amount of output that could be produced during that time. For example, a North household can produce six shirts per hour, so the opportunity cost for a 10-minute (1/6 hour) transaction is one shirt. In this case, the net gain from trade is the gross gain of 2 shirts minus the transaction cost of 1 shirt, or a net gain of one shirt. As long as the transaction time is less than 1/3 hour (two shirts), trade is beneficial for a North household. The South household, with lower productivity and thus a lower opportunity cost, has a lower transaction cost. For example, if the opportunity cost is 1/6 hour, the transaction cost is 1/6 loaf of bread, and the net gain is 2 shirts minus 1/6 shirt, or 11/6 loaves. For a South household, the threshold transaction time is 2 hours (2 loaves of bread).

Scale Economies in Exchange

The presence of specialization and trade will not necessarily cause a city to develop. The second assumption of the backyard-production model is that there are constant returns to scale in exchange. Under this assumption, an individual household is just as efficient in executing trades as a trading firm, so there is no reason to pay a firm

TABLE 2–2 Specialization and Gains from Trade

	North		South	
	Bread	Shirts	Bread	Shirts
Change in production from specialization	−2	+6	+2	−2
Exchange 4 shirts for 2 loaves	+2	−4	−2	+4
Gain from trade	0	+2	0	+2

to execute an exchange. Therefore, each North household will link up with a South household to exchange shirts and bread directly, without intermediaries.

Trading firms will emerge if there are economies of scale associated with exchange and trade. Recall the fourth axiom of urban economics:

Production is subject to economies of scale

A trading firm could use indivisible inputs such as a large truck to transport output between North and South. Similarly, workers who specialize in transportation tasks will be more efficient in transporting goods than workers who spend most of their time producing bread or shirts. In general, because trading firms have lower transaction costs, individual households will pay trading firms to handle exchanges.

The emergence of trading firms will cause the development of a trading city. To fully exploit scale economies, trading firms will locate at places that can efficiently collect and distribute large volumes of output. The concentration of trade workers will bid up the price of land near crossroads, river junctions, and ports. The increase in the price of land will cause people to economize on land by occupying smaller residential lots. The result is a place with a relatively high population density—a city.

TRADING CITIES IN URBAN HISTORY

Our simple model of the trading city suggests that trading cities develop when comparative advantage is combined with scale economies in transport and exchange. This observation provides some important insights into the history of cities before the Industrial Revolution of the 1800s. Most of the workers in these trading cities didn't produce goods, but instead collected and distributed goods produced elsewhere, such as agricultural products from the hinterlands and handcrafted goods from various locations. Trade was a risky business, and firms in the trading city provided insurance, credit, investment opportunities, banking, and legal services.

Trading Cities in World History

Trading cities have a long history. In the third millennium B.C., Phoenicians used fast sailing ships to serve as traders for the entire Mediterranean basin, trading dye, raw materials, foodstuffs, textiles, and jewelry. They established trading cities along the Mediterranean coast in present-day Lebanon. Around 500 B.C., Athens was a thriving site for regional trade, exchanging household crafts and olive products for food and raw materials from the countryside. During the 11th and 12th centuries, Italian city-states forged agreements with the Byzantine and Islamic rulers for trade with North Africa and the East. The Europeans traded wood, iron, grain, wine, and wool cloth for medicines, dyes, linen, cotton, leather, and precious metals. This trade was the major force behind the growth of Venice, Genoa, and Pisa.

Some cities were built on coercive transfer payments rather than voluntary trade. The Athenian empire developed in the aftermath of the successful war against Persia in the fifth century B.C. After the Greek city-states repelled the Persian invasion, they formed the Delian League for joint defense and later to carry the war into

Asia Minor. By the end of the successful campaign, Athens controlled the league and transformed the voluntary contributions of member city-states into payments of tribute to Athens. The system of homage and tribute led to the Peloponnesian War between the Athenian Empire and Sparta (431 to 404 B.C.). The war ended when Athens renounced control over its empire and demolished its defensive walls.

By the third century A.D., Rome had a population exceeding 1 million. The Romans established colonial cities throughout Europe and focused on collecting the agricultural surplus while they neglected urban production activity (Hohenberg and Lees, 1985). Instead of exchanging urban goods for agricultural products, Rome used conquest and tribute to feed its population. In the fourth and fifth centuries, attacks from Germanic tribes disrupted the Roman collection system. It appears that there was little interest outside of Rome in restoring the "trade" routes, so the losses from successive attacks were cumulative. If Rome had relied to a greater extent on voluntary exchange, the colonies would have had a greater stake in maintaining the exchange network and the Western empire might have recovered from the Germanic raids.

What are the lessons from the rise and fall of Athens and Rome? Early in its history, Athens thrived under a system of voluntary trade with other areas, exchanging urban goods for food from the countryside. The Athenians eventually switched to a system of conquest and tribute, resulting in war and the decline of the city. Mumford (1961) suggests that the city of Rome should have been called "Parasitopolis" to indicate its dependence on the labors of outsiders. The decline of Rome was caused in part by the disruption of its collection system by the Germanic raids. Perhaps the lesson is that cities based on coercive transfer payments are not sustainable.

Trading Cities in American History

The history of urban America illustrates the role of transport costs and comparative advantage in trading cities (Bartlett, 1998). In the 1700s, most cities served largely as trading posts for ocean trade. On the eastern seaboard, cities collected agricultural products from their hinterlands to the west and shipped them overseas. The volume of trade was limited by the dirt roads serving the interior: Travel was always slow and, in times of rain and melting snow, slippery. The Pennsylvania Turnpike, built with stone and gravel in 1792, increased travel speeds to a steady two miles per hour, increasing the market area and trading volume of the city of Philadelphia.

Farther to the north, New York State took more drastic steps, completing the 360-mile Erie Canal in 1825. The canal linked New York City, with its natural harbor, to vast agricultural areas to the north and west, and it cut freight costs from about 20 cents per ton mile to 1.5 cents. An additional canal connecting Lake Champlain to the Hudson River extended the market area of New York City to northern New England. The vast transportation network increased the volume of trade through New York City, increasing its size. By 1850, the city had a population of half a million, about 20 times its size at the end of the American Revolution. Other cities, including competitors to the south (Baltimore and Philadelphia), responded by building canals to connect hinterlands and ports, and by 1845 there were over 3,300 miles of artificial waterways in the United States.

Comparative advantage also plays a role in urban history. Eli Whitney's cotton gin (1794) provided a means of removing the sticky seeds of green-seed cotton, which could be grown throughout the south. The total output of cotton increased by a factor of 50 over a 15-year period, with most of the output coming from inland areas far from the east coast ports. American cotton was transported along rivers to New Orleans for shipment to textile firms in New England and Europe. The increase in cotton trade caused the rapid growth of New Orleans at the mouth of the Mississippi, and the development of upriver commercial cities such as Mobile, Alabama, and Natchez, Mississippi.

Later innovations in transportation reduced transport costs and contributed to the development of trading cities. Before the introduction of the steamboat in 1807, traffic was strictly downstream: After cargo was unloaded at the terminal point, wooden boats were broken up for lumber. The steamboat allowed two-way traffic and cut river freight costs, increasing the volume of trade and the size of river cities. Later, the steam engine was used to power locomotives, and railroad freight replaced river shipping as the principal means of transporting goods. Between 1850 and 1890, the ratio of railroad freight to river freight went from 0.10 to 2.0, and the volume of railroad freight increased by a factor of 240. The shift from river to railroad caused the decline of commercial cities along rivers and the rise of cities along the vast railroad network.

A FACTORY CITY

The third assumption of the backyard-production model is constant returns to scale in production. We'll maintain this assumption for bread production, but apply the fourth axiom of urban economics to shirt production:

Production is subject to economies of scale

A shirt factory will use indivisible inputs (machines) and allow workers to specialize in narrowly defined tasks, leading to a higher output per worker and lower average cost. Suppose a household can produce either a loaf of bread or one shirt per hour. A worker in a shirt factory is six times as productive as a home worker, so the factory worker produces six shirts per hour.

Determining Wages and Prices

We assume that workers are perfectly mobile, so the utility level of a city worker must be the same as the utility level for a rural worker. Recall the first axiom of urban economics:

Prices adjust to ensure locational equilibrium

A factory must pay its workers enough to make them indifferent between working in the factory city and in the rural area. A rural worker earns one loaf of bread

TABLE 2–3 Cost of Factory Shirt

Labor cost per hour	3⁄2 loaves
Cost of indivisible inputs per hour	1⁄2 loaf
Total cost per hour	2 loaves
Cost per shirt with 6 shirts produced per hour	1⁄3 loaf

per hour, so city workers must earn one loaf per hour plus an amount high enough to offset the higher cost of living in the factory city, such as higher land prices. For example, if the cost of urban living is 50 percent higher, locational indifference requires an hourly wage of 3⁄2 loaves of bread. A city worker will pay 1⁄2 loaf for land, leaving one loaf per hour of factory work, the same that she could earn producing bread in a rural area.

What's the price of factory shirts? The price must be high enough to cover the costs of labor and the indivisible inputs used to produce shirts. In Table 2–3 the labor cost per hour is the wage (3⁄2 loaves) and the hourly cost of indivisible inputs is 1⁄2 loaf. Adding these together, the hourly cost of producing shirts is two loaves of bread. To translate this into a cost per shirt, recall that a factory worker produces six shirts per hour, so the cost per shirt is one-sixth of the cost per hour, or 2⁄6 = 1⁄3 loaf. Therefore, for zero economic profit, the price per shirt must be 1⁄3 loaf of bread.

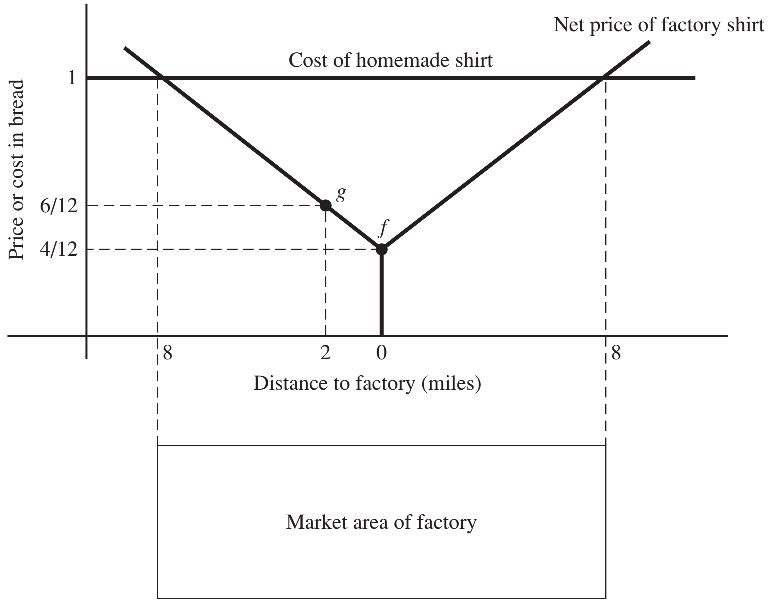
Suppose there is a single shirt factory in the region. The factory competes with homemade shirts, and will sell shirts to any household for which the net price of factory shirts is less than the cost of a homemade shirt. The cost of a homemade shirt is the one loaf of bread that is sacrificed to produce a shirt. The net price of a factory shirt equals the price charged by the factory (1⁄3 loaf) plus the opportunity cost of travel to and from the factory to buy the shirt.

The Market Area of a Factory City

Figure 2–1 (page 24) shows the net price of factory shirts and the market area of the shirt factory. As shown by point *f*, the net price for a consumer located just across the road from the factory (distance = 0) is the factory price, equal to 1⁄3 = 4⁄12 loaf of bread. Other consumers bear a travel cost when they buy factory shirts, so the net price is higher. Suppose the travel time is 1⁄12 hour per round-trip mile: It takes 1⁄12 hour to complete a round-trip of one mile in each direction. In an hour, a rural household can produce one loaf of bread, so in 1⁄12 of an hour of travel, it sacrifices 1⁄12 loaf. For example, at point *g* (two miles from the factory), the net price of a factory shirt is 6⁄12 loaves, equal to 4⁄12 paid at the factory plus 2⁄12 in travel cost (forgone bread production at home).

The market area of the factory is the area over which it underprices the home production of shirts. In Figure 2–1, the horizontal line shows the opportunity cost of homemade shirts, which is one loaf of bread. The net price of factory shirts is 4⁄12 at the factory and increases by 1⁄12 per mile, reaching one loaf at a distance of eight miles (4⁄12 + 8⁄12). In other words, the factory underprices home

FIGURE 2–1 Market Area of Factory



The net price of a factory shirt is the factory price (1⁄3 = 4⁄12 loaf of bread) plus transport cost (1⁄12 loaf per round-trip mile). The market area of the factory is the area over which the net price of a factory shirt is less than the cost of a homemade shirt (one loaf).

production up to eight miles away, so households within eight miles of the factory buy shirts rather than producing them at home. Beyond this point, households are self-sufficient, producing their own bread and shirts.

A factory city will develop around the shirt factory. Workers will economize on travel costs by living close to the factory, and competition for land will bid up its price. The higher price of land will cause workers to economize on land, leading to a higher population density. The result is a place of relatively high population density, a factory city. Note that we have already incorporated the higher land price into the factory wage and the factory price: Workers receive an hourly wage of 3⁄2 loaves to cover the opportunity cost of their time (1 loaf) and land rent (1⁄2 loaf).

THE INDUSTRIAL REVOLUTION AND FACTORY CITIES

Our simple model of the factory city suggests that a factory city develops because scale economies make factory shirts cheaper than homemade shirts. The Industrial Revolution of the 19th century produced innovations in manufacturing and transportation that shifted production from the home and the small shop to large factories in industrial cities. In contrast to the earlier trading cities, workers in factory cities produced products rather than simply distributing products produced elsewhere.

Innovations in Manufacturing

One of the key innovations of the Industrial Revolution was Eli Whitney's system of interchangeable parts for manufacturing, developed around 1800. Under the traditional craftsman approach, the component parts of a particular product were made individually—and imprecisely. Skilled craftsmen were necessary to produce the parts and then fit them all together. Under Whitney's system, the producer made a large batch of each part, using precise machine tools to generate identical parts. The identical parts were interchangeable, so unskilled workers could be quickly trained to assemble the parts. The replacement of handcraft production with standardized production generated large scale economies, causing the development of factories and factory cities.

Whitney applied this system to the production of muskets for the army. To prove to President-elect Jefferson and other government officials that his system would work with unskilled labor, he unloaded a random collection of parts onto the floor and had the officials assemble the muskets. He got the contract to manufacture 10,000 muskets and built a factory in New Haven, Connecticut, close to a stream that he used to power the factory. His system, which became known as the American System of Manufacturing, became the standard system for mass production.

The new system of manufacturing caused the development of factory cities. New machines, made of iron instead of wood, were developed to fabricate products in large factories. Manual production by skilled artisans was replaced by mechanized production using interchangeable parts, specialized labor, and steam-powered machines. Mass production decreased the relative cost of factory goods, causing the centralization of production and employment in large industrial cities.

As an illustration of the role of scale economies in the development of cities, consider the sewing machine, which was developed in the middle of the 19th century. At the beginning of the century, about four-fifths of the clothing worn in the United States was hand-sewn in the home for members of the household, and the rest was hand-sewn by tailors. The sewing machine (patented in 1846) allowed factories to underprice home producers, and by 1890 nine-tenths of U.S. clothing was being made in factories. New cities developed around the clothing factories.

A similar story line applies to shoes. Before 1700, most shoes were produced in the home or the local village. The cost of transportation was so high that local production was efficient. Over time, transportation costs decreased, and the putting-out system was implemented in the 1700s: Shoe producers distributed raw materials to cottage workers, collected their rough output, and finished the shoes in a central shop. As new shoemaking machines were developed, the number of operations performed in the central shops increased. The McKay sewing machine (for which a patent was granted to Lyman Blake in 1858) mechanized the process of sewing the soles to the uppers. The scale economies in shoe production increased to the point that shops became genuine factories, and cities developed around the shoe factories.

Innovations in Transportation

Innovations in intercity transportation contributed to industrialization and urbanization. As we saw earlier in the chapter, the dirt roads of the 1700s were replaced by turnpikes, and the construction of canals allowed a more dense network of inland water transport. The development of the steamship allowed two-way travel on major rivers, and the railroad system increased the speed and reach of the transportation system. All of these innovations decreased the relative price of factory goods, contributing to the growth of factory cities.

Innovations in Agriculture

One of the three conditions for the development of cities is an agricultural surplus to feed city dwellers. The Industrial Revolution generated a number of innovations that increased agricultural productivity. Farmers substituted machinery for muscle power and simple tools, increasing the output per farmer. The increased agricultural productivity freed people to work in urban factories and commercial firms. Between 1800 and 1900, the share of the population living in cities increased from 6 percent to 35 percent, reflecting the decrease in the number of agricultural workers required to feed city dwellers.

Consider first the sowing side of agriculture. At the start of the 19th century, plows were fragile, awkward, and often made of wood. These inefficient plows were replaced in the 1830s by the cast-iron plow, which was produced in factories in Pittsburgh and Worcester. In the 1840s, John Deere introduced the steel plow, which was lighter, stronger, and easier to handle. Later innovations allowed the farmer to adjust the depth and angle of the plow blade, increasing productivity further.

Consider next the reaping side of agriculture. In 1831, McCormick combined several earlier innovations into a horse-drawn harvesting machine that increased the productivity of the most labor-intensive part of agriculture. Using a horse-drawn reaper, two people could harvest the same amount of grain as eight people using traditional harvesting methods.

Other innovations contributed to higher agricultural productivity. The development of agricultural science led to innovations in planting, growing, harvesting, and processing. Innovations in transportation cut transport costs and allowed each farmer to serve a wider market area. Because of rising productivity, the share of employment in agriculture decreased over the 19th and 20th centuries, from over 90 percent to less than 3 percent.

Energy Technology and Location Decisions

During the Industrial Revolution, the location pattern of factory cities reflected changes in energy technology. The first factories used waterwheels turned by waterfalls and fast-moving streams to translate moving water into mechanical motion. The power was transmitted by systems of belts and gears. Textile manufacturers built factories along backcountry streams in New England and used waterwheels to run their machines. Some examples of waterwheel cities are Lowell, Lawrence, Holyoke, and Lewiston.

The refinement of the steam engine in the second half of the 19th century made energy a transportable input. A key innovation was John McNaught's development of a compounding engine (using steam twice, at descending pressures, to drive pistons) in 1845. The steam engine could be operated anywhere, with the only constraint being the availability of coal to fuel the engine. Some energy-intensive manufacturers located near the coal mines in Pennsylvania. Others located along navigable waterways and shipped coal from the mines to their factories. In New England, textile firms shifted from backcountry waterfall sites to locations along navigable waterways. Production shifted to the Fall River–New Bedford area along the south coast of New England. The later development of the railroad gave coal users another transport option, causing the development of factories along the vast network of rail lines. In general, the steam engine widened the location options for factories.

The development of electricity changed the location patterns of factories. Electricity generators were refined in the 1860s, and the electric motor was developed in 1888. Factories replaced belt-and-gear systems driven by a central steam engine with small electric motors for individual machines. The first factory to use electric power was adjacent to a hydroelectric generating facility at Niagara Falls. Rapid improvements in the electricity transmission soon allowed factories to be hundreds of miles from hydroelectric and coal-powered generating plants. Between 1900 and 1920, the share of factory horsepower from electric motors increased from 2 percent to 33 percent.

The development of electricity made factories more footloose. A firm could tap water power without locating close to the stream and use coal without shipping the bulky fuel to the factory. In general, the development of electricity decreased the importance of energy considerations in location decisions, causing firms to base their location choices on the accessibility to other inputs and to consumers.

A SYSTEM OF FACTORY CITIES

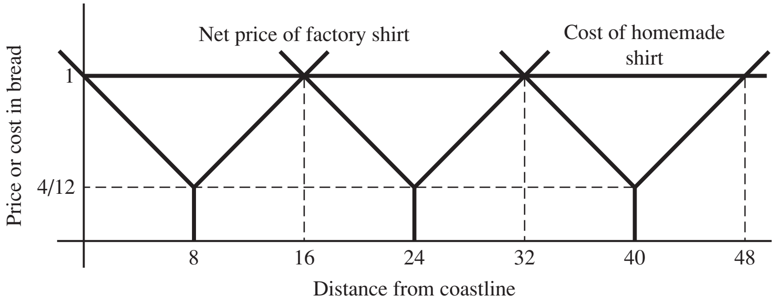
We can widen our horizon by looking at the entire region and consider the possibility of additional factory cities. Firms can enter the shirt industry by building shirt factories at different locations, and each firm will have a local monopoly in the area surrounding its factory. Recall the fifth axiom of urban economics:

Competition generates zero economic profit

If there are no restrictions on entry, firms will continue to enter the market until economic profit is zero.

Figure 2–2 (page 28) shows the equilibrium in the region. The horizontal axis measures distance from a coastline. The rectangular region is 48 miles wide, and in equilibrium has three shirt factories, each with a market area 16 miles wide. The

FIGURE 2–2 System of Factory Cities



Each factory's market area is 16 miles wide, so a system of factory cities develops with a distance of 16 miles between cities. In this equilibrium, workers specialize, with shirt workers in cities and bread producers in rural areas between the cities.

market areas of the factories span the region: Every location in the region lies within the market area of some factory. There is complete labor specialization: Workers in factory cities produce shirts (and receive bread as wages), and workers in rural areas produce bread (and pay bread to get factory shirts).

This is an equilibrium because each firm makes zero economic profit and workers are indifferent between rural and city life:

- **Zero economic profit.** The factory price of 4/12 loaf equals the average cost of producing shirts, including the cost of urban workers and the cost of indivisible inputs.
- **Locational indifference for workers.** The wage for factory workers is high enough to cover (1) the opportunity cost of working in factories rather than producing bread in the rural area and (2) the higher cost of urban living (land rent).

What about rural residents? For a rural resident just outside the factory cities, the net price of a factory shirt is 4/12 loaf of bread, compared to a homemade cost of one loaf. At the other extreme, a rural household eight miles from the factory pays a net price of one loaf per shirt (4/12 + 8/12 in travel cost). Recall the first axiom of urban economics:

Prices adjust to ensure locational equilibrium

In this case, the price of land in rural areas will adjust to make people indifferent between locations that differ in their accessibility to the shirt factory. The shorter the distance to the factory, the lower the net price of factory shirts, and the more a household is willing to pay for land. In other words, the price of land adjusts to fully compensate for differences in accessibility.

Landowners benefit from the scale economies in production that generate the regional system of factory cities. In the rural areas, the price of land is higher at locations close to the factory city. In the factory city itself, competition among workers for locations near the factory bids up the price of land.

RESOURCES-ORIENTED FIRMS AND PROCESSING CITIES

Up to this point, we have ignored the cost of transporting the raw materials required to produce urban goods (shirts). We have implicitly assumed that factory workers harvest wool from wild sheep who wander by the factory at just the right time to be sheared for shirts. In the language of urban economics, we have assumed that the raw materials required for production are ubiquitous—available at all locations at the same price. This is an extreme case of a market-oriented industry, defined as an industry in which the cost of transporting output is large relative to the cost of transporting inputs. The Appendix to this chapter explores the location decisions of market-oriented firms.

Consider the opposite extreme. Suppose it is costly to transport material inputs, but output can be transported at zero cost. This is the extreme case of a materials-oriented industry, defined as an industry for which the cost of transporting material inputs is large relative to the cost of transporting output. The Appendix to this chapter explores the location decisions of materials-oriented firms. For example, the sugar content of sugar beets is roughly 15 percent, so it takes seven tons of beets to produce one ton of sugar. Beet-sugar firms locate their plants close to the beet fields to economize on transport costs (Holmes and Stevens, 2004). Beet-sugar producers will cluster in the regions of the country where weather and soil conditions are favorable for the production of sugar beets.

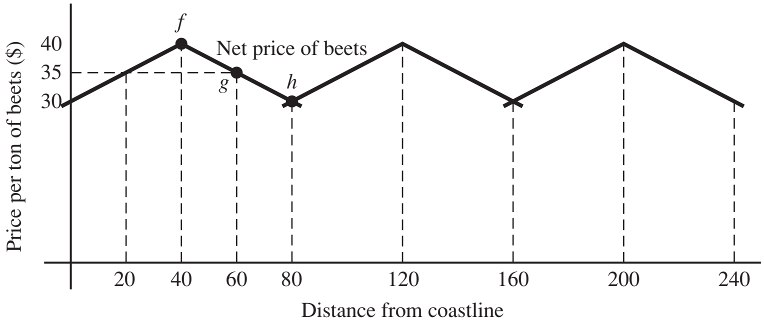
Scale Economies and Market Areas

The process of transforming sugar beets into sugar is subject to scale economies. Processors use indivisible inputs and engage in factor substitution, so the average processing costs decrease as the quantity increases. The typical sugar-beet processing plant employs 186 workers, about four times the average number of employees per plant in manufacturing.

The market area of a processing plant is determined by the net price farmers receive. The net price equals the price paid by the processor minus the cost of transporting the beets from the farm to the processing plant. In Figure 2–3, the horizontal axis measures the distance from a coastline. Consider a processing plant located 40 miles from the coastline. If the price paid by the processor is \$40, the net price is \$40 for a farmer across the road from the processor (point *f*), and drops to \$35 for a farmer 20 miles away (point *g*). Farmers naturally sell to the processing plant that generates the highest net price, so this processing plant has a market from the coastline to 80 miles inland.

Figure 2–3 shows a regional equilibrium with three processing plants, each with a market area of 80 miles. Each firm is the single buyer of sugar beets within

FIGURE 2–3 System of Processing Cities



The net price of beets (received by farmers) decreases as the distance to the processing plant increases. The market area of the typical sugar-beet processing plant is 80 miles wide, so a system of processing cities develops with a distance of 80 miles between them.

its 80-mile market area, so it has a local monopsony (as opposed to monopoly for a single seller). Recall the fifth axiom of urban economics:

Competition generates zero economic profit

If there are no restrictions on entry, firms will continue to enter the market until economic profit is zero.

System of Processing Cities

The location of sugar-beet processing plants leads to the development of a system of processing cities. The people who work in the processing plants live nearby to economize on commuting, resulting in a place with a relatively high population density. As Holmes and Stevens (2004) show, beet-sugar processing plants locate in the regions where beet production occurs, and they carve out input market areas within each beet-growing region.

Note the similarities of the beet-sugar industry and the shirt industry. In the shirt industry, with relatively high cost of transporting output, each firm gets a local monopoly, with all consumers patronizing the nearest factory. In the beet-sugar industry, with relatively high cost of transporting input, each firm gets a local monopsony, with all farmers selling their output to the nearest beet-sugar plant.

Other Examples of Resources-Oriented Industries

The same logic applies to other resources-oriented industries (Kim, 1999). The production of leather requires hides and tannin (from tree bark) for the tanning process. The tannin content of bark is only 10 percent, so it takes a lot of bark to produce a ton of leather. In 1900, U.S. leather producers located close to forests to economize on the transport costs of tannin. In the 20th century, improvements in the extraction process and the development of synthetic tannin reduced the orientation toward forest sites, and firms moved closer to other input sources.

The location decisions of steel producers reflected changes in the input requirements of coal and iron ore. Early in the history of the industry, a ton of steel required five tons of coal and two tons of ore, and steel production was concentrated near coal deposits. Technological innovations reduced the coal content, and steel producers were pulled toward locations that provided access to ore deposits, including sites on the Great Lakes that offered water access to ore from the Masabi Range in Minnesota. Each ton of steel required 175 tons of water, and the Great Lakes sites also provided a plentiful supply of water.

The location decisions of these and other resources-oriented industries caused the development of processing cities. Leather cities developed around tanneries and steel cities developed around steel mills. Lumber producers locate near forests, causing the development of lumber cities centered on sawmills. Ore processors locate near mineral deposits, causing the development of mining cities.

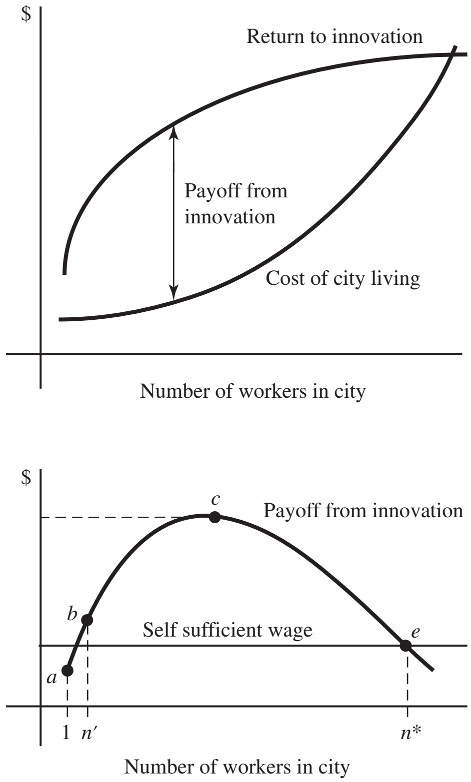
INNOVATION CITIES

Cities are centers of innovation because they facilitate knowledge spillovers, the exchange of knowledge and ideas among innovators. The bulk of patents for new products and production processes are issued to people in cities. As we'll see in the next chapter, the incidence of patents (the number of patents per capita) increases with city size and employment density. A key factor in innovation is the education level of the metropolitan workforce: an increase in the share of the population with college degrees increases patent intensity.

To illustrate the role of knowledge spillovers in urban development, consider a region that does not experience scale economies in production or exchange. The absence of scale economies means that there will be no trading cities or factory cities. Unless a worker finds an alternative source of income, he or she will be self-sufficient. The alternative activity is innovation—generating ideas that can be sold to people outside the region. The innovations could be in consumer goods, or in production—new intermediate goods used in production or improved techniques for producing goods.

The key assumption of the model is that innovation is facilitated by collaboration. The larger the number of people who share knowledge and ideas—either formally or informally—the higher the payoff to innovation. In the upper panel of Figure 2–4, the return to innovation increases with the number of workers in the cluster, but at a decreasing rate. The cost of living in a city increases with the number of workers, a result of increased competition for land and higher prices. In the lower panel of Figure 2–4, the payoff to innovation (the return minus the cost of city living) is shaped like a hill, with a peak at point c.

FIGURE 2–4 Innovation City



The payoff from innovation equals the return to innovation minus the cost of living in a cluster. The payoff for solo innovation (point a) is less than the self-sufficient wage, but the innovation payoff exceeds the wage for larger workforces. The equilibrium number of workers in the innovation city is n^*.

Workers in the region choose either self-sufficiency or working with other workers in an innovation cluster (a city). In the lower panel of Figure 2–4, the horizontal line shows the wage from self-sufficiency. In this case, the wage exceeds the payoff from solo innovation (an innovation cluster of one worker). Suppose that initially all workers are self-sufficient. This is an equilibrium allocation because solo innovation is less lucrative than self-sufficiency, meaning that no single worker has an incentive to switch to innovation. As a result, the distribution of population is uniform—there are no cities.

The equilibrium without cities is unstable. Suppose that a small group of workers form an innovation cluster, with n' workers in the cluster. The innovation payoff per worker is shown by point *b*. The innovation payoff now exceeds the self-sufficient wage, so each worker in the cluster will be better off than a self-sufficient worker. And if an additional worker joins the cluster, the payoff will increase as the city moves up the payoff hill. Workers will continue to join the innovation cluster as long as the payoff from innovation exceeds the self-sufficient wage. The stable equilibrium is shown by point *e*: the payoff from innovation equals the wage, so workers no longer have an incentive to switch from self-sufficiency to innovation. There are n^* workers in the innovation cluster, and we have an innovation city. The innovations produced in the city are sold outside the region, generating income that city workers use to purchase consumer goods.

Under what conditions will an innovation city develop? A key factor is the education level of the workforce. The collaborative nature of the innovation process means that innovators are most productive when they interact with educated and creative people. For a region with a highly educated workforce, the "payoff hill" will be relatively tall and wide, so the population of the innovation city will be relatively large. In contrast, a region with a poorly educated workforce will have a payoff hill that is short and narrow, leading to either a small innovation city or no innovation at all.

In the last few decades, changes in telecommunication have increased the payoff from innovation. In the modern world of instant communication by word and video, innovation has a worldwide market: a good idea can be sold to people throughout the world. The payoff from a great idea is very large, and the best way to develop a great (and lucrative) idea is to collaborate with educated and creative people in a city.

SUMMARY

Cities exist because of the benefits of centralized exchange (trading cities) and centralized production (factory and processing cities). We have focused on the market forces that generate cities. For a discussion of other possible reasons for cities, such as religion and defense and their role in the development of the first cities, see Mumford (1961) and O'Sullivan (2005). Here are the main points of this chapter:

1. A trading city develops when comparative advantage is combined with scale economies in exchange.
2. A factory city develops when there are scale economies in production.
3. The Industrial Revolution caused massive urbanization because of its innovations in agriculture, transportation, and production.
4. Changes in energy technology altered the location decisions of firms, with water power generating factories along streams, steam power generating factories along rivers and railroads, and electricity making firms more footloose.
5. Spatial competition among firms generates a market area for each firm and a system of cities.
6. An innovation city develops if there is a relatively large payoff from collaborative innovation.

APPLYING THE CONCEPTS

For exercises that have blanks (_____), fill each blank with a single word or number. For exercises with ellipses (. . .), complete the statement with as many words as necessary. For exercises with words in square brackets ([increase, decrease]), circle one of the words.

1. **Innovation and Trading Cities: Numbers**
   Consider the example shown in Tables 2–1 and 2–2. Suppose a single number changes: The shirt output per hour in the North is four shirts instead of six shirts. To specialize, a North household switches one hour from bread to shirts, and a South household switches two hours from shirts to bread. The exchange rate is three shirts for two loaves. There are scale economies in exchange, and a transaction takes 30 minutes to execute.
   a. The gains from trade are _____ for North and _____ for South.
   b. A trading city [will/won't] develop because. . . .
   c. Suppose an innovation in transportation decreases the transaction time. A trading city will develop if the transaction time is less than _____ minutes because. . . .
2. **Matter Transmitter and Trading City**
   Consider a region with two standardized products (bread and shirts) that are transported by horse-drawn wagons. There is a single trading city. Consider the effects of a new matter transmitter, an indivisible input that is economical for trading firms in the city, but not economical for an individual household. A transmitter instantly transports goods from a production site (a farm or workshop) to the trading firm, and the marginal cost of transport is zero.
   a. The labor used in transporting products from production sites to trading firms will [increase, decrease, not change] because. . . .
   b. The volume of trade in the region will [increase, decrease, not change] because. . . .
   c. The labor used in processing transactions (banking, accounting, insuring) will [increase, decrease, not change] because. . . .
   d. The trading city will grow if _____ exceeds _____.
   e. Suppose the transmitter technology changes, and it becomes economical for an individual household. The trading city will [grow, shrink, disappear] because. . . .
3. **Drilling for Cities**
   Consider a country with two regions that are separated by a mountain range. Initially each region is self-sufficient in shirts and bread, and there are no cities. Suppose that a tunnel is drilled through the mountain, decreasing travel costs between the two regions. The tunnel will cause the development of a trading city if three following conditions are met. . . .
4. **Spring-Loaded Sneakers**
   Consider the example shown in Figure 2–1, with the "martini-glass" shape showing the net price of the factory product. Suppose all consumers switch to

spring-loaded sneakers, decreasing walking time per round-trip mile from 1/12 hours to 1/18 hours.

a. The slope of the net-price curve changes to _____ and the width of the market area changes from 16 miles (8 on each side) to _____ miles (_____ on each side).

b. Using Figure 2–2 as a starting point, the spring-loaded sneakers change the number of factories in the 48-mile region from _____ to _____, each with a market area _____ miles wide.

5. Innovation and Market Areas

Consider the example shown in Figure 2–1. Depict graphically the effects of the following sequence of changes on the "martini glass" figure and the width of a factory's market area. The changes are cumulative.

a. An innovation in production that doubles labor productivity in factories will [shorten, lengthen, not change] the stem of the martini glass from _____ loaves to _____ loaves because. . . .

b. The width of the market area increases from 16 miles (8 miles on each side) to _____ miles (_____ on each side) because. . . .

c. An innovation in transportation that doubles consumer travel speeds will [decrease, increase, not change] the slope of the martini glass from _____ loaf per mile to _____ loaf per mile.

d. The width of the market area increases from _____ miles (_____ on each side) to _____ (_____ on each side) miles because. . . .

6. Matter Transmitter in a Factory City

Consider a region with a single factory city that developed as a result of scale economies in production. Consider the effects of a new matter transmitter, which can instantly transport goods (but not people) from the factory to any consumer up to 12 miles away, with a zero marginal cost of transport. The hourly cost of the transmitter is one loaf of bread. Using Figure 2–1 as a starting point, show the effects of the matter transmitter on the market area of the factory. Is "martini glass" still an apt descriptor of the figure? If not, what is a better descriptor?

7. Singing and the Internet

Consider a region where households produce and consume two products, bread and live musical performances. All workers are equally productive at producing bread and music. The production of bread is subject to constant returns to scale, with one hour required to produce each loaf. In an hour, a single person can produce one unit of music for herself. A choral group of 20 people working for an hour (on practice and performance) can produce one unit of music for an audience of 80 people. Assume that the opportunity cost of actually listening to the music is zero. The travel cost for music consumers is 1/8 hour per round-trip mile.

a. The equilibrium price for choral music—the price paid by each person who listens—is _____ loaves because. . . .

b. Use a martini-glass figure like Figure 2–1 to show the equilibrium market area of the choral group. The stem of the martini glass is _____ and the slope is _____ per mile, so the market area is _____ on each side because. . . .

c. Suppose choral music becomes available on the Internet, and the provider charges 1/2 loaf per song. Use a martini-glass figure like Figure 2–1 to show the new market area. The market area is _____ miles on each side because. . . .

8. Catapult in Retireland

In Retireland, no one commutes to work, and everyone consumes a single good (food), which is imported from another region and can be purchased (one meal at a time) from the nearest vending machine. Alternatively, food can be delivered by a distant catapult, capable of flinging a meal through a food slot on the customer's roof. The price of a delivered catapult meal is $6 and the price of a vending-machine meal is $2. The travel cost for consumers is $0.04 per round-trip meter ($0.02 per meter traveled).

a. Use a martini-glass figure like Figure 2–1 to show the equilibrium market area of the vending machine. The stem of the martini glass is _____ and the slope is _____ per meter, and the horizontal curve showing the cost of catapult meals is at _____. The market area is _____ meters on each side because. . . .

b. Arrows up, down, or horizontal: As the distance to the nearest vending machine decreases, the price of land _____ and population density _____.

9. Performance City

Consider a region where households produce and consume musical performances, either live performances or free recorded performances transmitted over the Internet into homes. After accounting for price, the utility from a live performance is three times the utility from recorded music. There are scale economies in the production of live music, which are provided at a performance center at the center of the region.

a. Using Figure 2–1 as an inspiration, depict graphically a method for determining the market area of the performance center. Hint: The vertical axis measures utility rather than price.

b. Suppose the demand for live performance is income-elastic, and the demand for recorded performances is independent of income. Use a second graph to show the effect of an increase in consumer income on the market area of the performance center.

10. Diesel Cost and Market Areas

Consider Figure 2–3, which shows the market areas of sugar-beet processing plants. Suppose that an increase in the price of diesel fuel doubles the cost of hauling sugar beets from farms to processing plants. Modify Figure 2–3 to show the effects of higher transport costs.

a. The slope of the net-price curve changes from _____ per mile to _____ per mile.

b. If the number of processing plants remains at three, the net price received by the farm that is most distant from its closest processing plant changes from _____ per ton to _____ per ton.

11. Beer and Wine

Consider the locations of breweries and wineries.

a. Most breweries locate close to their customers (far from their primary input sources) because. . . .

b. Most wineries locate close to their input sources (far from their consumers) because. . . .

c. Consider a nation that is 120 miles wide. Beer consumers are uniformly distributed throughout the nation, while grapes are uniformly distributed through the western region of the nation. There will be two evenly spaced wineries and two evenly spaced breweries. The wineries will locate at mile _____ and mile _____, splitting the _____. The breweries will locate at mile _____ and mile _____, splitting the _____.

12. Consider a region where the self-sufficient wage is constant at $4. Suppose the payoff from innovation in a city with population n (measured in thousands) is $\pi(n) = 2 + n^{1/2} - (n/10)$.

a. Suppose a group of 1 (thousand) workers form a city. Will other workers have an incentive to join the cluster?

b. Suppose a group of 9 (thousand) workers form a city. Will other workers have an incentive to join the cluster?

c. Compute the (stable) equilibrium size of the innovation city.

REFERENCES AND ADDITIONAL READING

1. Bartlett, Randall. *The Crisis of American Cities*. Armonk, NY: M.E. Sharp, 1998.
2. Combes, P., and H. Overman. "The Spatial Distribution of Economic Activities in the European Union." Chapter 64 in *Handbook of Regional and Urban Economics 4: Cities and Geography,* eds. V. Henderson and J. F. Thisse. Amsterdam: Elsevier, 2004.
3. Davis, Kingsley. "Urbanization." In *The Urban Economy,* ed. Harold Hochman. New York: W. W. Norton, 1976.
4. Ellison, Glen, and Edward Glaeser. "The Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration?" *American Economic Review* 89 (1999), pp. 311–316.
5. Fujita, Mashisa, and Jacques-Francois Thisse. *Economics of Agglomeration*. Cambridge: Cambridge University Press, 2002.
6. Hohenberg, Paul M., and Lynn H. Lees. *The Making of Urban Europe 1000–1950*. Cambridge, MA: Harvard University Press, 1985.
7. Holmes, T., and J. Stevens. "Spatial Distribution of Economic Activities in North America." Chapter 63 in *Handbook of Regional and Urban Economics 4: Cities and Geography,* ed. V. Henderson and J. F. Thisse. Amsterdam: Elsevier, 2004.
8. Kim, Sukkoo. "Regions Resources, and Economic Geography: Sources of U.S. Regional Comparative Advantage, 1880–1987." *Regional Science and Urban Economics* 29 (1999), pp. 1–32.

9. Mumford, Lewis. *The City in History*. New York: Harcourt Brace Jovanovich, 1961.
10. O'Sullivan, Arthur. "The First Cities." In *A Companion to Urban Economics,* ed. Richard Arnott and Daniel McMillen. Boston: Blackwell, 2005.

Appendix: Location Decisions of Transfer-Oriented Firms

This brief appendix explores the location decisions of *transfer-oriented* firms. For a transfer-oriented firm, the dominant location factor is the cost of transporting inputs and outputs. The firm chooses the location that minimizes total transport costs, defined as the sum of procurement and distribution costs. *Procurement cost* is the cost of transporting raw materials from the input source to the production facility. *Distribution cost* is the cost of transporting the firm's output from the production facility to the output market.

The classic model of a transfer-oriented firm has four assumptions that make transportation cost the dominant location variable.

- **Single transferable output.** The firm produces a fixed quantity of a single product, which is transported from the production facility to an output market.
- **Single transferable input.** The firm may use several inputs, but only one input is transported from an input source to the firm's production facility. All other inputs are ubiquitous, meaning that they are available at all locations at the same price.
- **Fixed-factor proportions.** The firm produces its fixed quantity with fixed amounts of each input. In other words, the firm uses a single recipe to produce its good, regardless of the prices of its inputs. There is no factor substitution.
- **Fixed prices.** The firm is so small that it does not affect the prices of its inputs or its product.

Under these assumptions, the firm maximizes its profit by minimizing its transportation costs. The firm's profit equals total revenue (price times the quantity of output) less input costs and transport costs. Total revenue is the same at all locations because the firm sells a fixed quantity of output at a fixed price. Input costs are the same at all locations because the firm buys a fixed amount of each input at fixed prices. The only costs that vary across space are procurement costs (the costs of transporting the firm's transferable input) and distribution costs (the costs of transporting the firm's output). Therefore, the firm will choose the location that minimizes its total transport cost.

The firm's location choice is determined by the outcome of a tug-of-war. The firm is pulled toward its input source because the closer to the input source, the lower the firm's procurement costs. On the other side, the firm is pulled toward the market because proximity to the market reduces the firm's distribution costs.

RESOURCE-ORIENTED FIRMS

A resource-oriented firm is defined as a firm that has relatively high costs for transporting its input. Table 2A–1 shows the transport characteristics for a firm that produces baseball bats, using 10 tons of wood to produce three tons of bats. The firm is involved in a weight-losing activity: Its output is lighter than its transferable input because the firm shaves down logs to make bats.

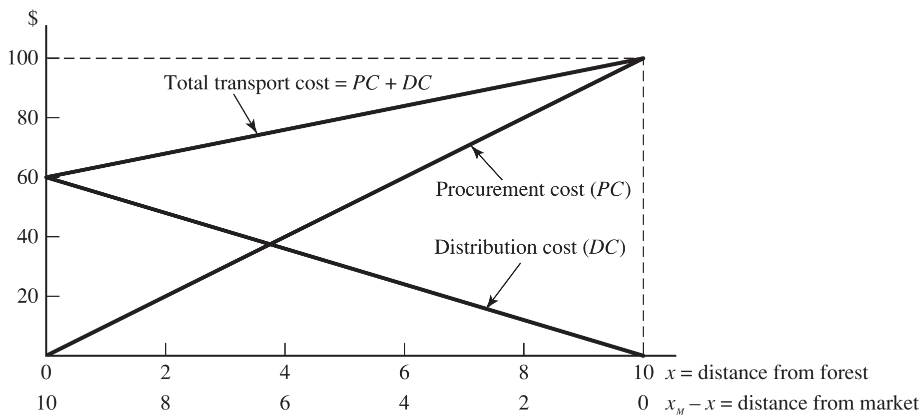
The outcome of the locational tug-of-war is determined by the monetary weights of the firm's inputs and outputs. The monetary weight of the input equals the physical weight of the input (10 tons) times the transportation rate ($1 per ton per mile), or $10 per mile. Similarly, the monetary weight of the output is three tons times $2, or $6 per mile. This firm is a resource-oriented firm because the monetary weight of its transferable input exceeds the monetary weight of its output. Although the unit cost of transporting output is higher (because finished bats must be packed carefully, but logs can be tossed onto a truck), the loss of weight in the production process generates a lower monetary weight for the output.

Figure 2A–1 shows the firm's transportation costs. Using x as the distance from the input source (the forest) to the production site (the factory), the firm's

TABLE 2A–1 Monetary Weights for a Resource-Oriented Firm

	Input (wood)	Output (bats)
Physical weight (tons)	10	3
Transport rate (cost per ton per mile)	$1	$2
Monetary weight (physical weight times rate)	$10	$6

FIGURE 2A–1 A Weight-Losing Firm Locates at Its Input Source



Total transport cost (procurement cost plus distribution cost) is minimized at the input source (the forest) because the monetary weight of the input ($10) exceeds the monetary weight of the output ($6).

procurement cost equals the monetary weight of the input (the physical weight w_i times the transport cost rate t_i) times the distance between forest and factory:

$$PC = w_i \cdot t_i \cdot x$$

The slope of the procurement-cost curve is the monetary weight of the input, so *PC* rises by $10 per mile, from zero at the forest to $100 at the market 10 miles away.

The distribution costs are computed in an analogous way. Using x_M as the fixed distance between the forest and the market (10 miles in our example), the firm's distribution cost is the monetary weight of the output (weight w_o times the transport cost rate t_o) times the distance from the factory to the market:

$$DC = w_o \cdot t_o \cdot (x_M - x)$$

The slope of the distribution-cost curve is the monetary weight of the output, so as we move from the forest toward the market, *DC* decreases by $6 per mile, from $60 at the forest (10 miles from the market) to zero at the market.

Total transport cost is the sum of procurement and distribution costs. In Figure 2A–1, total transport cost is minimized at the forest site at $60. Starting at any location except the forest, a one-mile move toward the forest would decrease procurement cost by $10 (the monetary weight of the wood) and increase distribution cost by $6 (the monetary weight of bats), for a net reduction of $4. The firm's total transport cost is minimized at the input source because the monetary weight of the input exceeds the monetary weight of the output. Some other examples of weight-losing firms that locate close to their input sources are beet-sugar factories, onion dehydrators, and ore processors.

Some firms are resource oriented because their inputs are relatively expensive to transport. For example, a canner produces one ton of canned fruit with roughly a ton of raw fruit. The firm's input is perishable and must be transported in refrigerated trucks, while its output can be transported less expensively on regular trucks. Because the cost of shipping a ton of raw fruit exceeds the cost of shipping a ton of canned fruit, the monetary weight of the input exceeds the monetary weight of the output, and the firm locates near its input source, a fruit farm. In general, a firm's input will be more expensive to ship if it is more bulky, perishable, fragile, or hazardous than the output.

There are many examples of industries that locate close to their transportable inputs (Ellison and Glaeser, 1999). The producers of soybean and vegetable oil are concentrated in Nebraska, North Dakota, and South Dakota, close to the farms that supply soybeans and corn. Milk and cheese producers are concentrated in South Dakota, Nebraska, and Montana, close to dairy farms. Sawmills and other wood processors are concentrated in Arkansas, Montana, and Idaho, close to vast timberlands.

MARKET-ORIENTED FIRMS

A market-oriented firm is defined as a firm that has relatively high costs for transporting its output to the market. Table 2A–2 shows the transport characteristics for a bottling firm that uses one ton of sugar and three tons of water (a ubiquitous input)

TABLE 2A–2 Monetary Weights for a Market-Oriented Firm

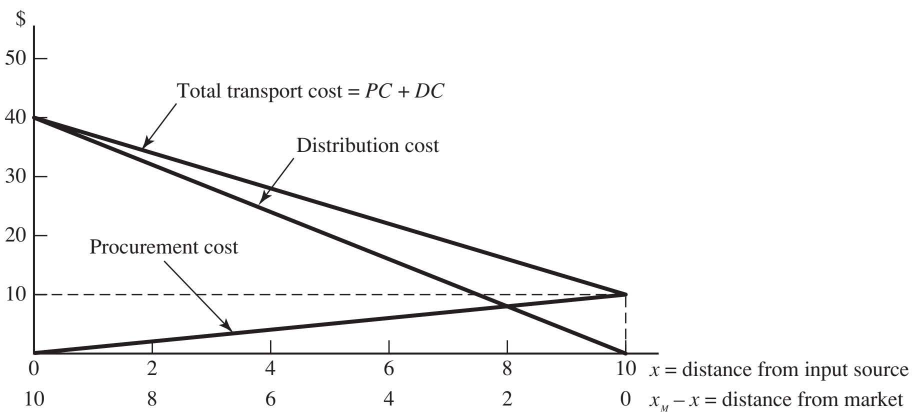
	Input (sugar)	Output (beverages)
Physical weight (tons)	1	4
Transport rate (cost per ton per mile)	\$1	\$1
Monetary weight (physical weight times rate)	\$1	\$4

to produce four tons of bottled beverages. The firm is involved in a weight-gaining activity in the sense that its output is heavier than its transferable input. The monetary weight of the output exceeds the monetary weight of the input, so this market-oriented firm will locate near its market.

As shown in Figure 2A–2, the firm's transport cost is minimized at the market. Because the monetary weight of the output exceeds the monetary weight of the input, a one-mile move away from the market increases the distribution cost by more than it decreases procurement cost. Specifically, such a move increases distribution cost by \$4, but decreases procurement cost by only \$1, for a net loss of \$3. For this weight-gaining activity, the tug-of-war between input source and market is won by the market because there is more physical weight on the market side.

Some firms are market oriented because their output is relatively expensive to transport. Output will be relatively costly to transport if it is bulky, perishable, fragile, or hazardous. The output of an automobile assembly firm (assembled cars) is more bulky than the inputs (e.g., rolls of wire, sheets of metal). The cost of shipping a ton of automobiles exceeds the cost of shipping a ton of component parts, so the monetary weight of the output exceeds the monetary weight of the inputs, pulling the firm toward the market. The output of a bakery is more perishable than its

FIGURE 2A–2 A Weight-Gaining Firm Locates at Its Output Market



Total transport cost is minimized at the market because the monetary weight of the output (\$4) exceeds the monetary weight of the transferable input (\$1). The weight-gaining firm locates at its output market.

inputs, pulling bakeries toward consumers. A weapons producer combines harmless inputs into a lethal output, and the firm locates near its output market to avoid transporting the hazardous (or fragile) output long distances. In general, when a firm's output is relatively bulky, perishable, fragile, or hazardous, the tug-of-war will be won by the market, not because the output is heavier, but because it is more expensive to transport.

THE PRINCIPLE OF MEDIAN LOCATION

The classic model of the transfer-oriented firm assumes that the firm has a single input source and a single market. For more complex cases involving multiple inputs or markets, we can use *the principle of median location* to predict where a firm will locate:

The median location minimizes total travel distance

The median location splits travel destinations into two equal halves, with half the destinations in one direction and half in the other direction.

We can illustrate this principle with the location decision of Ann, who makes and delivers pizzas to consumers along a highway. Under the following assumptions, her objective is to minimize her total delivery distance.

1. All inputs (labor, dough, toppings) are ubiquitous (available at all locations for the same price), so input transport costs are zero.
2. The price of pizzas is fixed, and each consumer along the highway demands one pizza per day.
3. Ann bears the delivery cost of $1 per mile traveled. Each pizza delivery requires a separate trip.

Figure 2A–3 shows the distribution of consumers along the highway. Distances are measured from the western end of the highway (point *W*). There are two customers at point *W*, eight customers at point *X* (one mile from *W*), one customer at *Y*, and 10 customers at *Z*.

Ann will minimize her total delivery distance at point *Y*, the median location. Point *Y* is the median because there are 10 customers to the west (at points *W* and *X*) and 10 customers to the east (at point *Z*). To show the superiority of the median location, suppose Ann starts at *Y*, and then moves one mile east to point *S*. The

FIGURE 2A–3 The Principle of Median Location

	W	X	Y	S	Z
Distance from *W*	0	1	2	3	9
Number of consumers	2	8	1	0	10

Total delivery cost is minimized at the median location (*Y*), defined as the location that splits travel destinations into two equal halves.

move reduces her travel distance to customers at point *Z* by 10 miles, but increases her travel distance to customers to the west (at points *W*, *X*, and *Y*) by 11 miles. Her total delivery distance increases because she moves closer to 10 customers but farther than 11 customers. In general, any move away from the median location will increase delivery distances for the majority of consumers, so the total delivery distance increases.

It is important to note that the distance between the consumers is irrelevant to the firm's location choice. For example, if the *Z* consumers were located 100 miles from *W* instead of nine miles from *W*, the median location would still be point *Y*. Total delivery distance would still be minimized (at a higher level, of course) at point *Y*.

The principle of median location provides another explanation of why large cities become larger. In Figure 2A–3, suppose the locations (*W*, *X*, *Y*, *Z*) are cities, with the populations of cities *W*, *X*, and *Y* (in millions) indicated by the numbers listed for customers. For example, the population of city *W* is 2 million. In addition, suppose city *Z*, at the end of the line, has a population of 12 million, making it the median location. In this case, firms will minimize total delivery cost by locating in the large city, so the large city will grow. A firm that started in city *Z* and then moved to the west would decrease delivery costs for a minority of its customers (11 million) while increasing delivery costs for a majority (12 million), so total delivery cost would increase. The lesson from this example is that the concentration of demand in large cities causes large cities to grow.

TRANSSHIPMENT POINTS AND PORT CITIES

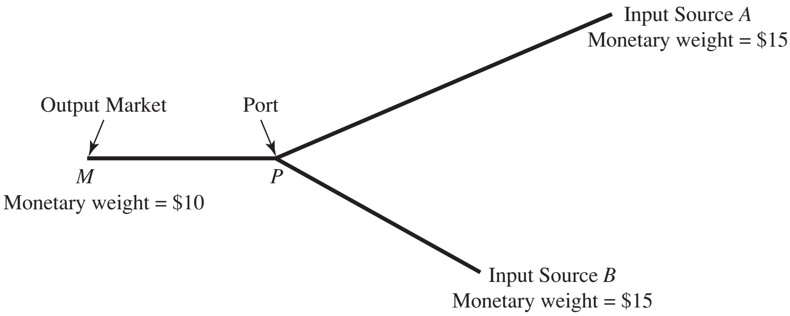
The principle of median location also explains why some industrial firms locate at transshipment points. A transshipment point is defined as a point at which a good is transferred from one transport mode to another. At a port, goods are transferred from trucks or trains to ships; at a railroad terminal, goods are transferred from trucks to the train.

Figure 2A–4 shows the location options for a sawmill. The firm harvests logs from locations *A* and *B*, processes the logs into lumber, and then sells the lumber in an overseas market at *M*. Highways connect points *A* and *B* to the port, and ships travel from the port to *M*. The sawmill is a weight-losing activity: The monetary weights of the inputs are \$15 for point *A* and \$15 for point *B*, and the monetary weight of the output is \$10.

Where will the firm locate its sawmill? Although there is no true median location, the port is the closest to a median location. If the firm starts at the port (*P*), it could move either toward one of its input sources or to its market.

- **Toward input source *A*.** A one-mile move from *P* toward point *A* will cause offsetting changes in the costs of transporting logs from the two input sources: the cost of logs from *A* would decrease, but the cost of logs from *B* would increase. At the same time, the cost of transporting output would increase by \$10. Given the offsetting changes in input transport costs and the increase in output

FIGURE 2A–4 Median Location and Ports



The firm locates its sawmill at the port (*P*) because it is the median transport location. A one-mile move from the port toward either input source would increase output transport costs by $10 while generating offsetting changes in the input transport costs for sources *A* and *B*. A one-mile move from the port toward the market would increase input transport costs by $30 but decrease output transport costs by only $10.

costs, the port location is superior to locations between *P* and *A*. The same argument applies for a move from *P* toward *B*.

- **To market (*M*).** Unless the firm wants to operate a floating sawmill, it would not move to points between the port and the overseas market at *M*. It could, however, move all the way to the market. A move from *P* to *M* would decrease output transport cost by $10 (the monetary weight of output) times the distance between *M* and *P*, and increase input transport cost by $30 (the monetary weight of the inputs) times the distance. Therefore, the port location is superior to the market location.

Although the sawmill is a weight-losing activity, it will locate at the port, not at one of its input sources. The port location is efficient because it provides a central collection point for the firm's inputs.

There are many examples of port cities that developed as a result of the location decisions of industrial firms. Seattle started in 1880 as a sawmill town: Firms harvested trees in western Washington, processed the logs in Seattle sawmills, and then shipped the wood products to other states and countries. Baltimore was the nation's first boomtown: Flour mills processed wheat from the surrounding agricultural areas for export to the West Indies. Buffalo was the midwestern center for flour mills, providing consumers in eastern cities with flour produced from midwestern wheat. Wheat was shipped from midwestern states across the Great Lakes to Buffalo, where it was processed into flour for shipment by rail to cities in the eastern United States. In contrast with Baltimore, which exported its output (flour) by ship, Buffalo imported its input (wheat) by ship.

CHAPTER 3

Why Do Firms Cluster?

People don't go there anymore. It's too crowded.

—Yogi Berra

If two firms compete for customers in a region, will they locate close together or far apart? It is natural to imagine that the two firms will split the region into two halves, giving each firm a local monopoly. That's what happened in the theoretical models of Chapter 2, and it happens for many firms in the real world. Yet all sorts of competing firms locate close to one another, including carpet producers in Georgia and television producers in Los Angeles. Why?

This chapter explores agglomeration economies, the economic forces that cause firms to locate close to one another in clusters. The forces acting on firms in a single industry together are called localization economies, indicating that they are "local" to a particular industry. For example, firms in the software industry cluster in Silicon Valley. When agglomeration economies cross industry boundaries, they are called urbanization economies. The idea is that the presence of firms in one industry attracts firms in other industries. For example, the corporate headquarters of different industries cluster in cities. Urbanization economies lead to the development of large, diverse cities. As we'll see, localization and urbanization economies have common roots.

Before we explore the reasons for localization economies, it will be useful to look at some facts on industry clusters in the United States. Table 3–1 shows the facts on employment clusters for six industries. In the production of aircraft engines, the four metropolitan areas listed (Hartford, Phoenix, Cincinnati, and Indianapolis) contain nearly half of U.S. employment in the industry. Firms producing biopharmaceuticals cluster in New York, Chicago, Philadelphia, and San Francisco, while firms producing software cluster in Seattle, the San Francisco Bay Area, and Boston. Among the small metropolitan areas with relatively large clusters of software employment are Austin, Texas, and Raleigh, North Carolina. Bloomington, Indiana, has over a fifth of national employment in the production of elevators and moving stairs, while Los Angeles has over two-fifths of national employment in video production and distribution.

TABLE 3–1 Select Industrial Clusters in U.S. Metropolitan Areas, 2004

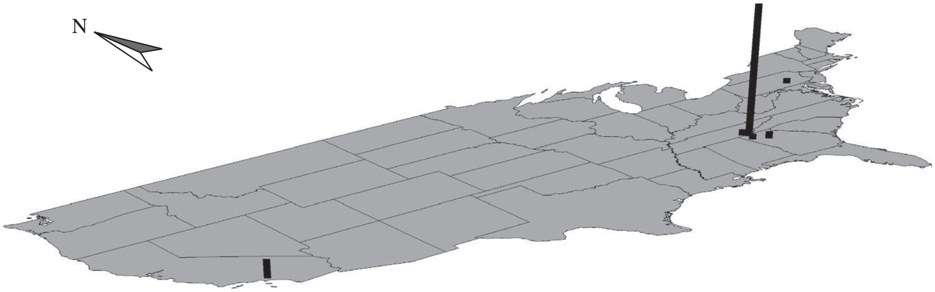
Product	Metropolitan Area	2004 Employment	Nationwide Employment (%)
Aircraft engines	Hartford, CT	15,619	22.67
	Phoenix, AZ	7,500	10.89
	Cincinnati, OH	6,957	10.10
	Indianapolis, IN	4,045	5.87
Biopharmaceutical products	New York, NY	51,604	27.21
	Chicago, IL	19,754	10.42
	Philadelphia, PA	11,383	6.00
	San Francisco, CA	10,706	5.65
Computer software	Seattle, WA	36,454	11.10
	San Francisco, CA	31,353	9.54
	San Jose, CA	29,221	8.89
	Boston, MA	23,415	7.13
Elevators and moving stairways	Bloomington, IN	1,750	20.03
	New York, NY	1,170	13.39
Financial services	New York, NY	427,296	12.97
	Chicago, IL	151,499	4.60
	Los Angeles, CA	142,337	4.32
	Boston, MA	133,342	4.05
Video production and distribution	Los Angeles, CA	161,561	44.00
	San Francisco, CA	28,394	7.73
	New York, NY	27,541	7.50

Source: Author's calculations based on data from Cluster Mapping Project, Harvard Business School.

Of course, not all industry clusters occur because of agglomeration economies. We saw in the previous chapter that beet-sugar production facilities are concentrated in beet-growing areas. Similarly, employment in the tobacco-products industry is concentrated in tobacco-growing areas: North Carolina has about 31 percent of national employment in the industry. For the hospitality and tourism industry, two clusters occur in cities with legalized gambling (Las Vegas and Atlantic City).

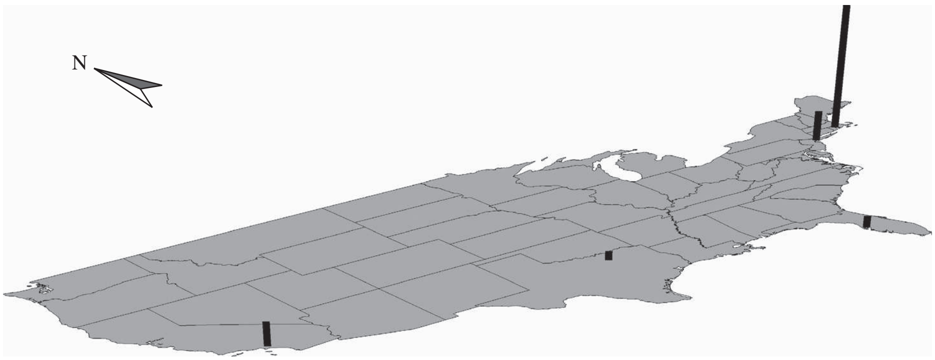
Maps 3–1 and 3–2 show the locations of job clusters of two industries. Each vertical bar shows the number of jobs in a particular industry in a specific metropolitan area. In Map 3–1, the bars show the job clusters for the carpet and rug industry, which is concentrated in the area around Dalton, GA, home to almost 17,000 jobs in the industry, or 41 percent of the industry's nationwide employment. There are smaller clusters nearby in Atlanta and Chattanooga, and more distant clusters of 2,300 jobs in Los Angeles and 750 jobs in Harrisburg, PA. Map 3–2 shows the job clusters for the production of costume jewelry. The bars show employment clusters for the costume-jewelry industry, which is concentrated in Providence, RI (4,100 jobs, or 55 percent of national employment), with smaller concentrations in New York, Los Angeles, Tampa, and Dallas. For additional maps of employment clusters, visit the Web site of the book.

MAP 3–1 Job Clusters: Carpets and Rugs



The bars show employment in the production of carpets and rugs, with 16,790 jobs in Dalton, GA, and smaller clusters in Los Angeles; Atlanta; Chattanooga, TN; Harrisburg, PA; and Rome, GA.

MAP 3–2 Job Clusters: Costume Jewelry



The bars show employment in the production of costume jewelry, with 4,100 jobs in Providence, RI, and smaller clusters in Los Angeles; New York; Tampa, FL; and Dallas, TX.

SHARING INTERMEDIATE INPUTS

Some competing firms locate close to one another to share a firm that supplies an intermediate input. The conventional list of production inputs includes labor, raw materials, and capital (machines, equipment, structures), but usually ignores intermediate inputs. An intermediate input is something one firm produces that a second firm uses as an input in its production process. For example, buttons produced by one firm are used as inputs by a dressmaking firm. The classic example of a cluster motivated by sharing an intermediate input is a cluster of dressmakers around a buttonmaker (Vernon, 1972).

Dresses and Buttons

Consider the production of high-fashion dresses. The demand for dresses is subject to the whims of fashion, so the dressmaking firms must be small and nimble, ready to respond quickly to changes in fashion. The varying demand for dresses causes varying demands for intermediate inputs such as buttons. A dressmaker's demand for buttons changes from month to month, not in the quantity demanded, but in the type of buttons demanded. One month the dressmaker might use square blue buttons with a smooth finish and the next month round pink buttons with a rough finish.

Consider next the production of buttons, the intermediate input. The production technology for buttons is summarized in three assumptions. The first is one of the axioms of urban economics:

Production is subject to economies of scale

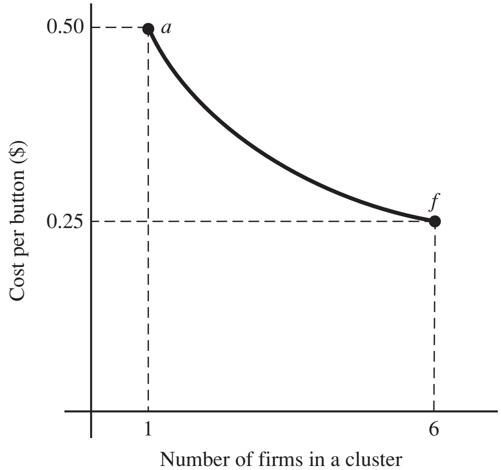
Because button producers use indivisible inputs and specialized labor, the cost per button decreases as the quantity increases. The scale economies are large relative to the button demand of an individual dressmaker, so dressmakers won't produce their own buttons but will buy them as intermediate inputs from button producers. There are two other assumptions in the dress-button model:

- **Face time.** A button for a high-fashion dress is not a standardized input that can be ordered from a catalog or a Web site, but requires interaction between dressmaker and buttonmaker to design and produce the perfect button for the dress of the month. The face time means that a dressmaker must be located close to its button supplier.
- **Modification cost.** Once a dressmaker buys a button from a buttonmaker, the dressmaker may incur a cost to modify the button to make a perfect match. For example, the dressmaker might have to shave the edges of a square button to make it a hexagon.

Figure 3–1 shows the average cost of buttons from the perspective of the dressmaker. Point *a* shows the cost for an isolated dressmaker, which has a relatively high button cost for two reasons. First, the buttonmaker produces for a single dressmaker, so output will be relatively low and the average cost (and price) of buttons will be relatively high. Second, the buttonmaker produces just one type of button (e.g., square buttons), so the dressmaker's modification costs will be relatively high. When the dress of the month calls for square buttons, modifications won't be necessary, but in all the other months, the dressmaker incurs a modification cost.

A dressmaker in a cluster has lower button costs for two reasons. First, a cluster of several dressmakers will generate sufficient button demand to allow buttonmakers to exploit scale economies, leading to lower button prices. Second, the larger total demand for buttons will allow buttonmakers to specialize in different varieties of buttons, reducing the modification costs of dressmakers. In a cluster, a dressmaker might be able to choose from buttons that are squares, hexagons, or triangles. In Figure 3–1, the average cost (and price) of buttons drops from $0.50 for an isolated

FIGURE 3–1 Clustering and the Average Cost of Intermediate Inputs



An isolated firm has a relatively high unit cost of buttons (point *a*). As the number of dressmakers in a cluster increases, the unit cost of buttons decreases because the firms generate sufficient demand to realize scale economies in button production and can support a wider variety of buttons.

firm (point *a*) to $0.25 for a six-firm cluster (point *f*). The lower cost provides an incentive for dressmakers to cluster to share a buttonmaker.

High-Technology Firms

The lessons from the button–dressmaker story apply to other industries. Firms producing high-technology products face rapidly changing demand for their cutting-edge products. The small, innovative firms share the suppliers of intermediate inputs, such as electronic components, and cluster to get the face time required to match components and new products. Innovative high-technology firms also share firms that provide product-testing services and locate close enough to quickly tap the facilities.

Intermediate Inputs in the Movie Industry

The U.S. movie industry is concentrated in the area in and around Hollywood, CA. There are seven major studios and hundreds of independent movie producers. Both types of movie producers rely on other firms to provide all sorts of intermediate inputs, such as script writing, film processing and editing, orchestras, and set design and construction. The scale economies associated with producing these inputs are large relative to the demands of individual firms, so movie producers share the suppliers of intermediate inputs. The inputs are not standardized and require face-to-face collaboration in their design and production. The result is a cluster of movie producers and the suppliers of intermediate inputs.

The market for movie props provides an example of intermediate inputs. The objects used in film scenes include mundane items such as table lamps and chairs, special items such as castoff medical instruments and vintage cars, and signature props such as elf ear tips and Gryffindor scarves. Although the major studios have their own internal props departments, most independent producers get their props from firms known as "prop houses." In the Hollywood area, there are three clusters of prop houses. Set decorators and dressers go from one prop house to another looking for the perfect object for the set, and the prop houses cluster to facilitate this comparison shopping.

SELF-REINFORCING EFFECTS CAUSE INDUSTRY CLUSTERS

So far we have seen that clustering is beneficial because it allows firms to take advantage of agglomeration economies from input sharing. What about the costs? In this part of the chapter, we will use an example of the clustering of movie producers to explore the costs and benefits of clustering. When agglomeration economies are strong enough to offset the cost of clustering, firms will form industry clusters, causing the development of specialized cities.

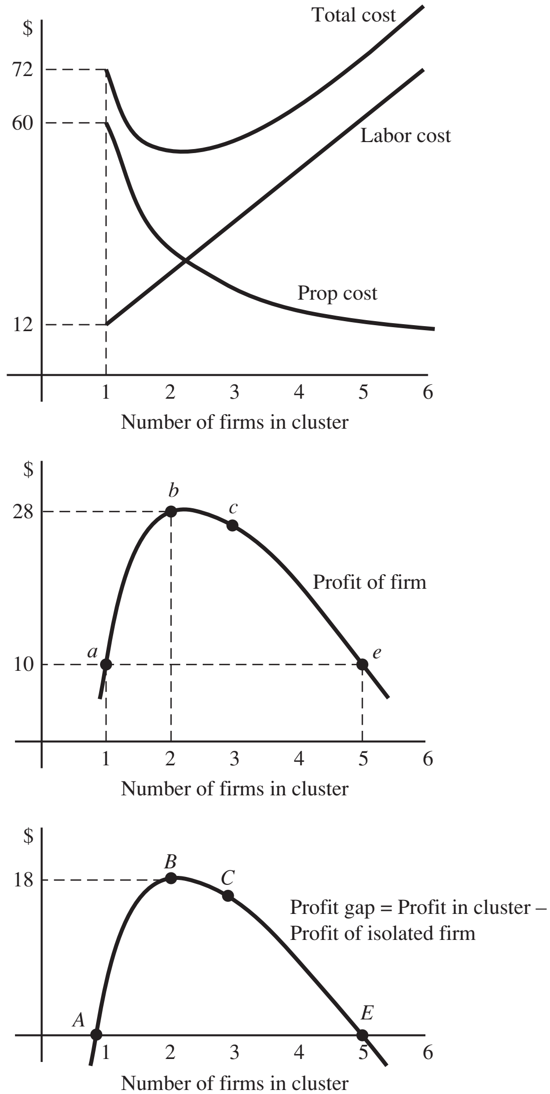
The Benefits and Costs of Clustering

Consider the location decisions of movie producers. Suppose the scale economies in providing movie props are large relative to the demand of an individual producer. As a result, movie producers won't run their own prop departments, but will instead purchase this intermediate input from prop houses. Locating in a cluster allows the producers to share prop houses and benefit from a lower price for props.

Figure 3–2 shows the trade-offs associated with clustering. In the upper panel, the negatively sloped curve shows the prop cost of the typical movie producer, which decreases as the number of producers in the movie cluster increases and the average cost—and price—of props decreases. The positively sloped curve shows the labor cost of the typical movie producer. The larger the number of firms in the cluster, the greater the competition for labor, and thus the higher the wages of movie workers and the higher the labor costs. The U-shaped curve shows the firm's total cost, equal to the sum of prop and labor costs. Going from one to two firms, the savings in prop cost dominate the increase in labor cost, and the total-cost curve reaches its minimum with two movie producers. Beyond that point, the increase in labor cost dominates, generating a positively sloped total-cost curve.

The middle panel of Figure 3–2 shows the profit of a typical movie producer for different numbers of producers in the cluster. Assume that the revenue of the typical producer is constant at \$82, and the profit equals this fixed revenue minus prop costs and labor costs. The profit of an isolated producer (a one-firm cluster), shown by point *a*, is \$10 = \$82 in revenue minus \$60 in prop cost and \$12 in

FIGURE 3–2 Self-Reinforcing Effects and Clustering



The profit gap, equal to the profit for firm in a cluster, minus the profit of an isolated firm increases, then decreases, reflecting the trade-offs from lower prop costs and higher labor costs. The profit gap reaches zero with five firms in the cluster, the equilibrium number.

labor cost. As the number of firms in the cluster increases, profit increases, then decreases. The inverted U reflects decreasing, then increasing total cost. With five producers in the cluster, the profit per producer again equals the profit of the isolated producer (point *e*).

The Profit Gap and the Size of the Cluster

The lower panel of Figure 3–2 shows the gap between the profit of a movie producer in a cluster and the profit of an isolated producer ($10). The profit gap is of course zero with a one-firm cluster (point *A*). The profit gap grows to $18 (point *B*, with 2 firms), then shrinks. In a five-firm cluster, the profit gap is again zero (point *E*).

How many movie producers will locate in the cluster? Suppose we have a large number of movie producers, and intially each is isolated, earning a profit of $10 (point *a* in the middle panel of Figure 3–2). Will this dispersed outcome persist? Suppose a single movie producer relocates next to another, forming a two-firm cluster. As shown in Figure 3–2, each firm in the cluster will earn $28 (point *b*), or $18 more than an isolated firm (point *B*). This higher profit gives the remaining isolated firms an incentive to relocate to the cluster. The third firm in the cluster will earn $26 in the cluster compared to $10 in isolation—a $16 gap (point *C*). Firms will continue to join the cluster as long as the profit gap is positive, that is, as long as the cluster location is more profitable than the isolated location. In the stable equilibrium, there are five firms in the cluster (point *E*). At this point, each firm in the cluster earns $10, the same as an isolated firm.

The agglomeration economies from sharing an intermediate input supplier generate self-reinforcing changes. Recall the second axiom of urban economics:

Self-reinforcing changes generate extreme outcomes

In this case, movie producers that compete for labor don't disperse to minimize labor costs, but instead cluster to realize agglomeration economies. In this example, rising labor costs generate diseconomies of scale that limit clustering, but the same logic applies with other diseconomies, such as rising land costs or rising transport costs for inputs and outputs.

SHARING A LABOR POOL

What do the producers of television programs and the producers of computer software have in common? Every year, dozens of new television programs are aired, and only a few are hits. In the rapidly changing software industry, hundreds of new products are introduced every year, and only a few succeed. For an individual firm in either industry, this year's new product—television program or computer program—may be wildly successful, and next year's may be a dud. In this environment of rapidly changing demand, unsuccessful firms will be firing workers at the same time that successful firms are hiring them. A cluster of firms facilitates the transfer of workers from unsuccessful firms to successful ones.

The key notion of sharing a labor pool is that the boom-bust process occurs at the level of the firm, not the industry. Suppose the total demand for output in an industry is constant over time, but the demand facing an individual firm varies from year to year. For example, the number of slots for television programs is fixed, so the success of one television firm (a hit) comes at the expense of another (a canceled dud). Similarly, the success of one firm's encryption software comes at the expense of other firms that introduce similar products.

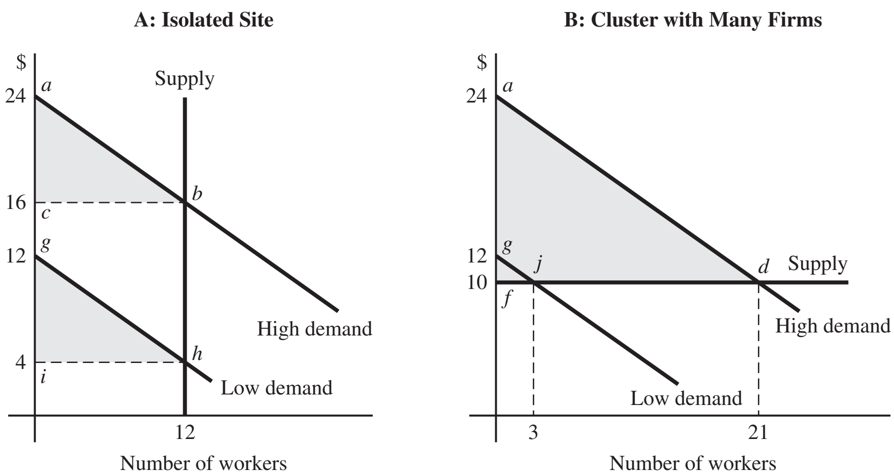
In this part of the chapter, we develop a formal model of labor pooling. The total demand at the industry level is constant, but the demand for each firm varies from year to year. For each firm, there are two possibilities—high demand or low demand—and each outcome is equally likely. As we'll see, there is an incentive for firms in such an industry to cluster to share a pool of workers.

The Isolated Firm

Consider first the situation for an isolated firm outside any industry cluster. The isolated firm doesn't face any competition for labor within its town, and to simplify matters, we assume that labor supply in the isolated site is perfectly inelastic, fixed at 12 workers. This means that wages will rise and fall with the demand for the firm's product.

When the demand for the firm's product is high, so is the firm's demand for labor. In Panel A of Figure 3–3, the high-demand equilibrium is shown by the intersection of the upper demand curve and the vertical supply curve at point *b*, generating a wage of $16. When demand for the firm's product is low, so is its demand for labor and the equilibrium wage ($4 at point *h*). To summarize, the isolated firm hires the same number of workers during high and low demand but pays a lower wage when demand is low.

FIGURE 3–3 Clustering to Share a Labor Pool



In an isolated site, the firm faces a perfectly inelastic supply of labor (12 workers). The firm hires the same number of workers during high demand and low demand but pays a higher wage during high demand.

In a cluster, the firm faces a perfectly elastic supply of labor, and the wage is fixed at $10. The firm hires 21 workers during high demand but only three workers during low demand.

Locating in a Cluster

The key difference between an isolated site and a cluster concerns the competition for labor and the variability of wages. Workers in the cluster can choose from a large number of firms. For every successful firm hiring workers, there is an unsuccessful firm firing them. Therefore, the total demand for labor in the cluster is constant, and so is the equilibrium wage.

Workers are mobile between the isolated site and the cluster, and in equilibrium they will be indifferent between the two locations. Recall the first axiom of urban economics:

Prices adjust to generate locational equilibrium

At the isolated site, the wage is uncertain, being either \$16 during high demand or \$4 during low demand. The two outcomes are equally likely, so the expected wage (the sum of the probabilities times the wages) is \$10:

$$\text{Expected wage} = \frac{1}{2} \cdot \$16 + \frac{1}{2} \cdot \$4 = \$10$$

To make workers indifferent between the two sites, the certain (constant) wage in the cluster must be \$10.

Panel B in Figure 3–3 shows the outcomes in the cluster. An individual firm can hire as many workers as it wants at the market wage. The typical firm hires 21 workers when demand is high (point *d*), but only three workers when demand is low (point *j*). When the demand for a firm's product goes from high to low, the firm fires 18 workers at the same time that another firm in the cluster is hiring 18 workers as its demand goes from low to high.

Expected Profits Are Higher in the Cluster

Expected profits will be higher in the cluster. To see why, consider what happens when a firm moves from the isolated site to a cluster and then experiences one year of high demand, followed by one year of low demand.

- **Good news when demand is high.** The move to the cluster cuts the wage (from \$16 to \$10) and allows the firm to hire more workers (21 instead of 12), generating higher profit in the cluster.
- **Bad news when demand is low.** The move to the cluster increases the wage (from \$4 to \$10), generating lower profit in the cluster.

Which is larger, the good news with high demand, or the bad news with low demand?

The good news will dominate the bad news because a firm in the cluster responds to changes in the demand for its product. When demand is high, the firm takes advantage of the lower wages (a \$6 gap) in the cluster by hiring more workers (21). When demand is low, a firm in the cluster cushions the blow of low demand by hiring fewer workers (only three). Because the firm changes its workforce when the demand for its product changes, the good news will be large relative to the bad news, and profit will be higher in the cluster.

Another way to show that profit is higher in the cluster is to compute the expected profits at the two sites. As shown in Section 3 of "Tools of Microeconomics" (the Appendix to the book), the labor-demand curve shows the marginal benefit of labor, the value of output produced by the marginal worker. A firm's profit from hiring a worker equals the difference between the worker's marginal benefit and the wage, and a firm's profit from its entire workforce is shown by the gap between the labor demand curve and the horizontal wage line. In Panel A of Figure 3–3, triangle *abc* shows the profit for an isolated firm when demand is high ($48), and triangle *ghi* shows the profit when demand is low ($48). In Panel B, the profit with high demand is shown by triangle *adf* ($147), and the profit with low demand is shown by triangle *gjf* ($3). So if the two outcomes are equally likely, the expected profit in the cluster is $75 (the average of $147 and $3), compared to $48 in the isolated site.

Labor Pooling in the Movie Industry

The U.S. movie industry, concentrated in the area in and around Hollywood, CA, provides an example of the benefits of labor pooling. One segment of the labor market includes workers involved in the craft and technical side of the industry. These workers move periodically from one producer to another as projects come and go, and rely on an "economy of favors," building and maintaining personal relationships to keep informed about potential jobs and ease the moves from one firm to another. The same phenomenon occurs for creative workers (actors, directors, writers) as they move between firms to work on different projects. In a cluster of movie producers, firms draw from a common labor pool, facilitating the flow of workers between firms.

There are a number of mechanisms that facilitate coordination in the local labor market and improve the flow of workers between firms. Intermediaries such as agents, casting directors, and talent managers match labor demanders and suppliers. Dozens of worker associations, including the Production Assistants Association and the Stuntmen's Association, provide useful information and training programs to their members. Colleges and universities in the area have professional programs that train students in the production of film and television. These coordinating mechanisms improve the efficiency of the labor market and help maintain the competitive advantages of the movie-industry cluster.

LABOR MATCHING

In a typical economic model of a labor market, we assume that workers and firms are matched perfectly. Each firm can hire workers who have precisely the skills the firm requires. In the real world, things are not so tidy. Workers and firms are not always perfectly matched, and mismatches require costly worker training. As we'll see, a large city can improve the matching of workers and firms in the untidy real world, decreasing training costs and increasing productivity.

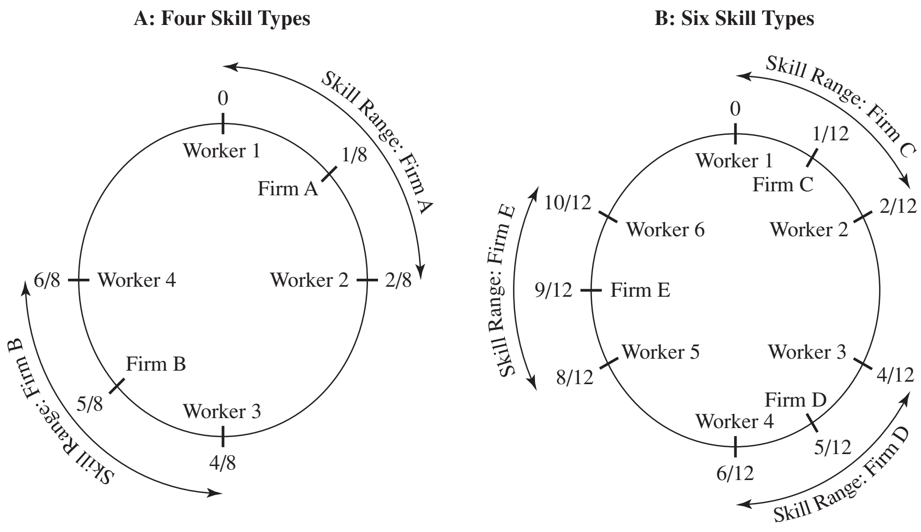
As an illustration of the labor matching problem, consider a set of software firms that hire computer programmers. Programmers have different skill sets, depending on their facility with different programming languages (e.g., C, C++, Java) and their experience with different programming tasks (e.g., graphics, number crunching, artificial intelligence, operating systems, e-commerce). Although some programmers are more productive than others, what matters for the matching model is that they have different skill sets. A firm enters the market with a particular skill requirement and hires workers who provide the best skill matches.

A Model of Labor Matching

Helsley and Strange (1990) developed a formal model of labor matching. The model uses several key assumptions about workers and firms.

- **Variation in worker skills.** Each worker has a unique skill described by a position or "address" on a circle with a one-unit circumference. In Panel A of Figure 3–4, there are four workers, and their skills are evenly spaced on the circle. The address of a worker is the distance between her skill position and the north pole of the circle. The addresses of the four workers are {0, 2/8, 4/8, 6/8}.

FIGURE 3–4 Skills Matching



With four skill types, worker addresses are {0, 2/8, 4/8, 6/8}. There are two workers per firm, so two firms will enter with skill requirements {1/8, 5/8}, and the mismatch per worker is 1/8.

With six skill types, worker addresses are {0, 2/12, 4/12, 6/12, 8/12, 10/12}. There are two workers per firm, so three firms will enter the market with skill requirements {1/12, 5/12, 9/12}, and the mismatch per worker is 1/12.

- **Firm entry.** Each firm enters the market by picking a product to produce and an associated skill requirement. In Panel A of Figure 3–4, one firm enters with skill requirement S = 1/8, and a second enters with S = 5/8.
- **Training costs.** Workers incur the cost associated with closing the gap between the worker's skill and the skills required by a firm.
- **Competition for workers.** Each firm offers a wage payable to any worker who meets its skill requirement, and each worker accepts the offer with the highest net wage, which is equal to the wage minus the training cost required to close the skills gap.

The next two assumptions of the matching model are related to the axioms of urban economics:

Production is subject to economies of scale

Because of scale economies in production, each firm will hire more than one worker. This is important because in the absence of scale economies, each firm would hire just one worker, and each worker would be perfectly matched with a firm. To simplify matters, we will assume that scale economies require each firm to hire two workers. The final assumption is that entry is unrestricted, so firms will continue to enter the market until economic profit is zero.

Competition generates zero economic profit

In the labor-matching model, entry involves picking a skill requirement and hiring workers with closely matched skills. In other words, each firm gets a local monopsony (single buyer) in the skill interval surrounding its skill requirement.

Panel A of Figure 3–4 shows the equilibrium with four skill types and two firms. The equilibrium mismatch per worker is 1/8. For example, the workers at S = 0 and S = 2/8 work in a firm with S = 1/8, so each worker has a skills gap of 1/8. Each firm pays a gross wage equal to the value of output produced by a perfectly matched worker. The net wage earned by a worker equals the gross wage minus the training cost:

$$\text{Net wage} = \text{Gross wage} - \text{Skills gap} \cdot \text{Unit training cost}$$

Suppose the gross wage is \$12 and the unit training cost is \$24. In the equilibrium shown in Panel A of Figure 3–4, the skill gap is 1/8, so the net wage is

$$\text{Net wage} = \$12 - \frac{1}{8} \cdot \$24 = \$9$$

Agglomeration Economies: More Workers Implies Better Matches

What happens to skills matching as an urban economy grows? We can represent an increase in the size of the workforce by increasing the number of workers on the unit circle. This increases the density of workers with respect to skills but doesn't change the range of skills. As we'll see, more workers means better skill matches and higher net wages.

TABLE 3–2 Number of Workers, Skills Gap, and Net Wage

Number of Workers	Skills Gap	Training Cost	Net Wage
4	1/8	\$24/8 = \$3	\$12 – \$3 = \$9
6	1/12	\$24/12 = \$2	\$12 – \$2 = \$10
12	1/24	\$24/24 = \$1	\$12 – \$1 = \$11

Panel B of Figure 3–4 shows the effects of increasing the number of workers from four to six. Each firm still hires two workers, so three firms will enter the market. In Panel B of Figure 3–4, the six workers are equally spaced, with skill addresses {0, 2/12, 4/12, 6/12, 8/12, 10/12}. The three firms enter the market with skill requirements {1/12, 5/12, 9/12}, so the mismatch per worker drops to 1/12. For example, workers at skill addresses 0 and 2/12 are hired by the firm at address 1/12, so each worker has a mismatch of 1/12. Workers incur lower training cost, so the net wage increases to \$10:

$$\text{Net wage} = \$12 - \frac{1}{12}\$24 = \$10$$

In general, an increase in the number of workers decreases mismatches and training costs, increasing the net wage. This is shown in Table 3–2 for up to 12 workers.

What are the implications of skill matching for the clustering of firms and urban development? The presence of a large workforce attracts firms that compete for workers, generating better skill matches and higher net wages for workers. The higher net wage provides an incentive for workers to live in large numbers in cities, so the attraction between firms and workers is mutual. Both firms and workers benefit from better skill matching.

KNOWLEDGE SPILLOVERS

A fourth agglomeration economy comes from sharing knowledge among firms in an industry. As Marshall (1920) explained,

> When an industry has chosen a locality for itself, it is likely to stay there for long; so great are the advantages which people following the same skilled trade get from near neighborhood to one another. The mysteries of the trade become no mysteries; but are as it were in the air, and children learn many of them unconsciously. Good work is appreciated; inventions and improvements in machinery, in processes and the general organization of the business have their merits promptly discussed; if one man starts a new idea, it is taken up by others and combined with suggestions of their own; and thus it becomes the source of new ideas.

There is ample evidence that knowledge spillovers cause firm clustering. Dumais, Ellison, and Glaeser (2002) show that knowledge spillovers increase the number of new plant births, with the largest effect on industries that employ college graduates. Their results suggest that knowledge spillovers are important in determining the locations of firms in idea-oriented industries. Rosenthal and Strange (2001) show that the most innovative industries are more likely to form clusters. They also show that knowledge spillovers are highly localized, petering out over a distance of a few miles.

There is also evidence that knowledge spillovers are more important for industries with small, competitive firms. A recent study compared two clusters of the electronics industry, California's Silicon Valley and Route 128 near Boston (Saxenian, 1994). Knowledge spillovers are more important in Silicon Valley because its network of specialized companies generates an atmosphere of collaboration, experimentation, and shared knowledge. In contrast, the firms in the Route-128 cluster are less interdependent so there are fewer knowledge spillovers.

EVIDENCE OF LOCALIZATION ECONOMIES

A large volume of economics literature examines the magnitude of localization economies. In searching for evidence of localization economies, researchers focus on the effects of industry concentration on (1) worker productivity, (2) the number of new production plants (plant births), and (3) growth in industry employment. If there are localization economies, we expect industry clusters to generate higher productivity, more births, and more rapid employment growth.

Consider first the effect of concentration on worker productivity. Henderson (1986) estimates the elasticity of output per worker with respect to industry output, defined as the percentage change in output per worker divided by the percentage change in industry output. For the electrical machinery industry, the elasticity is 0.05, meaning that a 10 percent increase in the output of the industry increases output per worker by 0.50 percent. The elasticity estimates for other U.S. industries range from 0.02 for the pulp and paper industry to 0.11 for the petroleum industry.

Mun and Hutchison (1995) use data from Toronto to estimate agglomeration economies in the office sector. They estimate a productivity elasticity of 0.27, suggesting that localization economies are more powerful in the office sector than in the manufacturing sector. The productivity effects are larger for growth in central locations and are localized.

Consider next the implications of industry concentrations for the location of new production facilities. Carlton (1983) examines the location choices of firms in three industries: plastics products, electronic transmitting equipment, and electronic components. His estimated elasticity of firm births with respect to industry output is 0.43: A 10 percent increase in industry output increases the number of births by 4.3 percent. More recently, Head, Reis, and Swenson (1995) show that Japanese corporations locate their new plants close to other Japanese plants in the same industry. Rosenthal and Strange (2003) show that firm births are more numerous in locations close to concentrations of employment in the same industry.

Consider next the effects of industry concentration on employment growth. Henderson, Kuncoro, and Turner (1995) show that growth in mature industries is more rapid in areas that start with large concentrations of the industry. Rosenthal and Strange (2003) compute this localization effect for six industries, including computer software. A zip code area that starts out with 1,000 more software jobs than another zip code area experiences a larger increase in software employment—about 12 more jobs. On average, the localization effect peters out at a rate of about 50 percent per mile. The rapid attenuation of the localization economies explains the local in "localization economies."

URBANIZATION ECONOMIES

So far in this chapter, we have considered agglomeration economies experienced within a particular industry, also known as localization economies. These localization economies generate clusters of firms producing the same product. In contrast, urbanization economies—defined as agglomeration economies that cross industry boundaries—cause firms of different industries to locate close to one another. The result is the development of large, diverse cities. The four agglomeration economies that generate localization economies also generate urbanization economies.

Sharing, Pooling, and Matching

Consider first the notion of input sharing. Although some intermediate inputs such as buttons are specific to an industry, others are shared by firms in different industries. For example, most industries use business services such as banking, accounting, building maintenance, and insurance. Similarly, firms in different industries share hotels and firms providing transportation services. In addition, firms share public infrastructure such as highways, transit systems, ports, and universities. By sharing these intermediate inputs, firms in larger cities pay lower prices and tap a wider variety of inputs.

Another source of agglomeration economies is labor pooling. Recall that labor pooling is beneficial when the product and labor demand per firm varies while total industry demand remains constant. A cluster of firms in the same industry facilitates the movement of workers from firing firms to hiring firms. Labor pooling generates urbanization economies when demand varies across industries, with some industries expanding while others decline.

Consider next the benefits of labor matching. Recall that an increase in a city's workforce increases the density of worker skills, reducing the mismatches between workers' skills and firms' skill requirements. Because some skill requirements are common to multiple industries, the benefits of labor matching cross industry boundaries. For example, firms in many industries require computer programmers, and firms in these industries benefit from producing in a city with a high density of programmers.

Corporate Headquarters and Functional Specialization

Corporations locate their headquarters in cities to exploit urbanization economies. Corporate executives and managers perform a variety of tasks—developing marketing campaigns, picking locations for new plants, and fending off lawsuits—and draw on other firms to accomplish these tasks. Corporate expenditures on outsourced legal, accounting, and advertising services are equivalent to about two-thirds of their wage bill (Aarland, Davis, Henderson, Ono, 2003).

TABLE 3–3 Increase in Functional Specialization of Metropolitan Areas

	Percentage Gap between Metropolitan Ratio of Management to Production Workers and the National Ratio		
Population	1950	1970	1990
5–20 million	+10.2	+22.1	+39.0
1.5 to 5 million	+0.30	+11.0	+25.7
75,000 to 250,000	−2.1	−7.9	−20.7
67,000 to 75,000	−4.0	−31.7	−49.5

Source: Gilles Duranton and Diego Puga. "From Sectoral to Functional Specialization," *Journal of Urban Economics* 57 (2005), pp. 343–70.

Corporations cluster to share firms that provide business services. For example, given the large economies of scale in producing advertising campaigns, corporations cluster to share advertising firms, and they get specialized marketing campaigns at a lower cost. Similarly, corporations are attracted by the large concentrations of firms providing financial and business services in midtown Manhattan, the Loop in Chicago, and the financial district of San Francisco.

In the last several decades, there has been a fundamental shift in the specialization of cities. Large cities have become increasingly specialized in managerial functions, while smaller cities have become more specialized in production. Duranton and Puga (2005) compute the ratio of managerial workers to production workers for the nation as a whole and for different metropolitan areas. Table 3–3 shows the percentage differences between the national ratio and the metropolitan ratio for metropolitan areas of different sizes. For example, in the largest metropolitan areas in 1950, the metropolitan ratio was 10.2 percent higher than the national ratio, indicating a slight specialization in managerial functions. At the other extreme, for the smallest areas, the metropolitan ratio was 4.0 percent lower than the national ratio, indicating slight specialization in production.

Over the 40-year period shown in Table 3–3, there was a dramatic change in specialization. By 1990, the ratio for largest cities was 39 percent higher than the national ratio, indicating substantial specialization in managerial functions. At the other extreme, the ratio for the smallest cities was nearly 50 percent lower than the national ratio, indicating a high degree of specialization in production. These changes in specialization were caused by decreases in the cost of managing production facilities from afar. Firms are better equipped to operate multiplant firms from headquarters in large cities where agglomeration economies generate lower production costs. The most important cost reductions have come from innovations in telecommunications, in particular the development of duplicators (photocopiers, fax machines, and e-mail) that have facilitated the rapid transmission of information and reduced the cost of coordination.

Knowledge Spillovers

The essential feature of knowledge spillovers is that physical proximity facilitates the exchange of knowledge between people, leading to new ideas. The ideas lead to new products as well as new ways to produce old products. Some knowledge spillovers occur within an industry, but knowledge spillovers often cross industry boundaries. A city that produces a wide variety of products is fertile ground for applying ideas refined in the design and production of one product to new products.

Carlino and Hunt (2009) study the factors that determine the incidence of patents across metropolitan areas. After adjusting the raw number of patents to incorporate their relative importance (reflected in the number of times a patent is cited in other patents), they computed the elasticities of patent intensity with respect to a number of variables, including the following.

- **Employment density (jobs per square mile)**. The overall elasticity is 0.22: a 10 percent increase in employment density increases patent intensity by about 2.22 percent. There are diminishing returns to density: the positive relationship levels off at an employment density of about 2,200 jobs per square mile.
- **Total employment**. The overall elasticity is 0.52: a 10 percent increase in total employment increases patent intensity by about 5.2 percent. There are diminishing returns to total employment: the positive relationship levels off at a metropolitan population of about 1.8 million.
- **Human capital (share of workforce with a college degree)**. The elasticity is 1.05: A 10 percent increase in the share of the population with a college degree increases patent intensity by 10.5 percent.
- **Establishment size**. The elasticity is −1.4: a 10 percent increase in the average size decreases patent intensity by 14 percent. It appears that people in cities with relatively competitive environments are more inventive.

The authors also document the substantial variation in patent intensity across metropolitan areas. The average patent intensity is 2.0, and the values ranging from 0.07 (in McAllen, TX) to 17 (in San Jose, CA). Following San Jose in the rankings are Rochester, NY; Trenton, NJ; Ann Arbor, MI; Austin, TX; Wilmington, DE; Raleigh-Durham, NC; Boston, MA; and San Francisco, CA.

Evidence of Urbanization Economies

There have been many studies of urbanization economies. The general conclusion is that the elasticity of productivity with respect to population is in the range 0.03 to 0.08 (Rosenthal and Strange, 2004). In other words, a doubling of population increases output per worker by between 3 percent and 8 percent. Two studies (Glaeser, Kallal, Scheinkman, and Schleifer, 1992, and Henderson, Kuncoro, and Turner, 1995) suggest that diversity promotes employment growth, especially in new and innovative industries. Hanson (2001) concludes that long-run industry growth is higher in cities with a wider variety of industries, suggesting that diversity promotes growth.

OTHER BENEFITS OF URBAN SIZE

The urbanization economies discussed so far—input sharing, labor pooling, skills matching, and knowledge spillovers—generate higher productivity and lower production costs. In this part of the chapter, we'll consider three other advantages associated with a larger urban economy: better employment opportunities for families, a better learning environment for workers, and better social opportunities.

These advantages of size increase the relative attractiveness of large cities and increase the supply of labor to big cities. How does that contribute to the clustering of firms in cities? Recall the first axiom of urban economics:

Prices adjust to generate locational equilibrium

An increase in the relative attractiveness of a big city decreases the wage that workers are willing to accept to live and work in the city, generating lower production costs for firms. This is similar to the Dullsville versus Coolsville example in Chapter 1: A city that has superior opportunities for family employment, learning, and social interactions has lower wages, everything else being equal.

Joint Labor Supply

Most families have two workers, but are tied to a single residential location. In other words, families must confront the problem of joint labor supply. If the skills of the two workers are suited to different industries, the family will be attracted to locations with a mix of industries. Therefore, the joint supply of labor encourages firms in different industries to cluster. The role of cities in resolving the issue of joint labor supply has a long history. In the 1800s, mining and metal-processing firms (employing men) located close to textile firms (employing women), and each industry benefited from the presence of the other. More recently, "power couples" (defined as a pair of college graduates) are concentrated in large cities, where they are more likely to find good employment matches for both workers.

Learning Opportunities

Another benefit of urban size comes from the greater learning opportunities in cities. Human capital is defined as the knowledge and skills acquired by workers in formal education, work experience, and social interaction. Human capital can be increased through learning by imitation, that is, observing other workers and imitating the most productive workers. A larger city provides a wider variety of role models for workers so it attracts workers looking for learning opportunities.

The evidence for urban learning comes from data on the wages earned by workers who migrate to cities (Glaeser, 1999). Wages are higher in cities, reflecting the higher productivity of urban workers. But when a worker migrates from a rural area, she doesn't earn the higher urban wage immediately. Instead she experiences rising wages over time as learning increases her productivity. When a worker leaves the city, her wage does not drop back to the wage she earned prior to coming to the city. Instead, the higher productivity resulting from urban learning leads to a higher wage outside the city. In other words, the benefits of urban learning translate into higher wages everywhere.

Social Opportunities

A third benefit of city size comes from social interactions. An implicit assumption of the backyard-production model in Chapter 2 is that people do not value social interaction. Of course, people enjoy interacting with one another, and a larger city provides more opportunities for social interactions.

To think about the social dimension of cities, recall the labor-matching model. Suppose we replace labor skills with social interests: People have different hobbies, conversational topics, and social activities. In addition, suppose we replace firms seeking good skills matches with people seeking a network of friends with similar interests. In a model of social-interest matching, a larger city will generate better interest matches, with each network (like each firm) achieving a tighter range of social interests. Some people live in cities to take advantage of better opportunities for social-interest matching.

To illustrate the notion of social benefits of large cities, suppose you want to form a book club to discuss your favorite book, *Giles Goat Boy* (by John Barth). In a small town, you may be the only person who has read the book. In contrast, thousands of people in the typical large city have read the book and perhaps a dozen will be eager to discuss the masterpiece. A quick Internet search reveals that larger cities have more book clubs on a wider variety of topics, consistent with the notion that bigger cities provide better social matches.

SUMMARY

Firms cluster to exploit agglomeration economies, including localization economies at the industry level and urbanization economies at the city level. Here are the main points of the chapter:

1. Firms may cluster to share a supplier of an intermediate input if the input is subject to relatively large scale economies and requires face time for its design and production.
2. Firms may cluster to share a labor pool if the variation in product demand is greater at the firm level than at the industry level.
3. Larger cities provide better skill matches, leading to higher productivity and wages.
4. People and firms are attracted to cities because they facilitate knowledge spillovers, learning, and social opportunities.
5. Agglomeration economies cause self-reinforcing changes in location: The movement of one firm to a city increases the incentive for other firms to move to the city.

APPLYING THE CONCEPTS

For exercises that have blanks (________), fill each blank with a single word or number. For exercises with ellipses (…), complete the statement with as many words as necessary. For exercises with words in square brackets ([increase, decrease]), circle one of the words.

1. **Attention Kmart Shoppers**
   Most of the dresses sold in the United States are produced in large factories that are dispersed, not concentrated. Reconcile this fact with the text discussion of localization economies that cause dressmakers to cluster.
2. **Labor Pooling: What's Fixed and Variable?**
   Consider the model of labor pooling, with each firm locating either in an isolated site or in a cluster with other firms. Fill the blanks with "fixed" or "variable."
   a. In an isolated site, the wage is _______ and the firm's workforce is _______ because. . . .
   b. In a cluster, the wage is _______ and the firm's workforce is _______ because. . . .
   c. Illustrate with two graphs, one for the isolated site and one for the cluster.
3. **Trade-offs with Clustering for Labor Pooling**
   Consider the model of labor pooling, with each firm locating either in an isolated site or in a cluster with other firms. Suppose that good times (high demand) and bad times (low demand) are equally likely. The table shows wages and workforces in different times and locations.

	Isolated		Cluster	
	Wage	Workforce	Wage	Workforce
Good times (high demand)	$40	50	$30	60
Bad times (low demand)	$20	50	$30	40

   a. Use a figure like Figure 3–2 to illustrate the situation.
   b. During good times, the benefit of being in the cluster as opposed to being isolated is _______, computed as. . . .
   c. During bad times, the cost of being in the cluster as opposed to being isolated is _______, computed as. . . .
   d. The benefit exceeds the cost because a firm in a cluster. . . .
4. **Mr. Mullet's Carnival**
   Mr. Mullet runs a traveling carnival that hires local workers in each city it visits. The demand for carnival activities is uncertain, with low or high demand equally likely in any given city. At the end of the year, Mr. Mullet reviews his financial records and discovers some puzzling differences between his experiences in small and large cities.
   i. He always paid the same wage in large cities ($9), but paid different wages in small cities ($6 or $12).
   ii. He always hired the same quantity of labor in small cities (20 workers), but different quantities in big cities (10 or 30 workers).
   a. Using Figure 3–3 as a model, illustrate with two graphs, one for the typical small city and one for the typical big city. Assume that the demand curves

for labor are linear and parallel, with vertical intercepts of \$18 (high demand) and \$12 (low demand).

b. In the typical big city with high demand, profit is ________ computed as. . . .
c. In the typical big city with low demand, profit is ________ computed as. . . .
d. In the typical small city with high demand, profit is ________ computed as. . . .
e. In the typical small city with low demand, profit is ________ computed as. . . .
f. The expected profit is ________ in a big city, compared to ________ in a small city.

5. Number of Workers and Net Wages

Using Table 3–2 as a starting point, suppose the gross wage is \$36 and the unit training cost is \$48. Complete the following table.

Number of Workers	Skills Gap	Training Cost	Net Wage
4	________	________	________
8	________	________	________
24	________	________	________

6. Models on the Color Wheel

Consider the model-management industry, with firms that supply human models for advertisements. Workers (models) vary in skin tone along the color wheel, which can be divided into 12 colors. Firms enter the market with a specific skin tone requirement for their models. If a model's skin tone does not match the firm's tone requirement, the model incurs a makeup cost to close the gap, with a cost of \$3 for each unit of color shift. For example, to go from color #2 to color #4, the cost is \$6. Given the scale economies in model management, each firm manages three models. The gross wage is \$20.

a. Smallville has six models, equally spaced on the color wheel at 12:00, 2:00, 4:00, and so on. There will be ________ firms in the city, with firm A at 12:00 and the other firm or firms at ________.
b. Illustrate with a graph like Figure 3–4.
c. For the typical firm, the average mismatch is ________ skin tones and the average makeup cost is \$ ________. The average net wage after makeup costs is ________, computed as. . . .
d. Bigburg has twice as many models as Smallville. It will have ________ firms. Its average makeup costs will be [lower, higher] and its average net wage will be [higher, lower].
e. Complete the following table.

Number of Models	Color Gap	Makeup Cost	Net Wage
6	________	________	________
12	________	________	________

7. Advertising and Corporate Clusters

Consider corporations that use advertising firms to develop marketing campaigns. Each corporation buys one campaign per year, and the cost per campaign is \$120/n, where n = the number of corporations in the cluster (and campaigns per year). The cost of labor per firm is $30 · n$. A corporation's profit equals its total revenue of $200 minus the sum of its marketing and labor costs. There are two location options: an isolated site ($n = 1$) or a cluster with up to five corporations.

a. Use a graph like Figure 3–2 to show the profit gap (profit in cluster – profit in isolation) for one through five corporations.

b. If initially all corporations are isolated and then one joins another to form a two-corporation cluster, other firms [will, won't] have an incentive to join the cluster because….

c. In the long-run equilibrium, there will be a cluster of ________ corporations, each of which will earn a profit of ________, differing from the profit of an isolated site by ________.

8. Agglomeration Economies and Auto Row

Chapter 1 uses Auto Row as an example of self-reinforcing changes that lead to extreme outcomes. Consider a city with three isolated automobile dealers, each of which has three buyers per day. The profit per car sold is $1,000. A two-dealer cluster would get six times as many buyers (18), and a three-dealer cluster would get 12 times as many buyers (36).

a. Use a graph like Figure 3–2 to show the profit gap (the profit for a firm in a cluster minus the profit for an isolated firm) for one, two, and three dealers.

b. If initially all dealers are isolated and then one joins another to form a two-dealer cluster, other firms [will, won't] have an incentive to join the cluster because. . . .

9. Personal and Pet Grooming in the Minimall

Suppose that personal grooming and pet grooming are complementary products. Betty Beehive could move her beauty shop from an isolated location to a minimall that also contains Peter's pet-grooming shop. If she moves, she will attract some of Peter's customers and her pre-rent profit will increase by $180. Her current rent is $100, compared to $300 in the minimall.

a. Betty [will, won't] make the move because….

b. Betty's presence in the minimall would increase Peter's profit by $100. If you were the manager of the minimall, with the power to set the rent of each tenant, what would you do?

10. Diversify the Economy?

According to the conventional wisdom concerning urban economic development, a city should develop a diverse economy with a large number of industries. Evaluate the merits of this advice in light of the empirical evidence concerning the magnitudes of localization and urbanization economies.

REFERENCES AND ADDITIONAL READING

1. Aarland, K., J. C. Davis, J. Henderson, and Y. Ono, *Spatial Organization of Firms: The Decision to Split Production and Administration*. Providence, RI: Brown University Press, 2003.
2. Audretsch, David, and Maryann Feldman. "Knowledge Spillovers and the Geography of Innovation." Chapter 61 in *Handbook of Regional and Urban*

Economics 4: Cities and Geography, eds. Vernon Henderson and Jacques-Francois Thisse. Amsterdam: Elsevier, 2004.

3. Carlino, Gerald, and Robert Hunt. "What Explains the Quantity and Quality of Local Inventive Activity?" *Brookings-Wharton Papers on Urban Affairs 2009*, pp. 65–124.
4. Carlton, D.W. "The Location and Employment Choices of New Firms." *Review of Economics and Statistics* 65 (1983), pp. 440–49.
5. Dumais, Guy, Glen Ellison, and Edward Glaeser. "Geographic Concentration as a Dynamic Process." *Review of Economics and Statistics* 84 (2002), pp. 193–204.
6. Duranton, Gilles, and Diego Puga. "Micro-foundations of Urban Agglomeration Economies." Chapter 48 in *Handbook of Regional and Urban Economics 4: Cities and Geography,* eds. Vernon Henderson and Jacques-Francois Thisse. Amsterdam: Elsevier, 2004.
7. Duranton, Gilles, and Diego Puga. "From Sectoral to Functional Specialization." *Journal of Urban Economics* 57 (2005), pp. 343–70.
8. Fujita, Mashisa, and Jacques-Francois Thisse. *Economics of Agglomeration*. Cambridge: Cambridge University Press, 2002.
9. Glaeser, Edward. "Learning in Cities." *Journal of Urban Economics* 46 (1999), pp. 254–277.
10. Glaeser, Edward L., Hedi D. Kallal, Jose A. Scheinkman, and Andrei Shleifer. "Growth in Cities." *Journal of Political Economy* 100 (1992), pp. 1126–52.
11. Hanson, Gordon. "Scale Economies and the Geographic Concentration of Industry." *Journal of Economic Geography* 1 (2001), pp. 255–76.
12. Harvard Business School, Cluster Mapping Project. http://data.isc.hbs.edu/isc/.
13. Head, K., J. Ries, and D. Swenson. "Agglomeration Benefits and Location Choice." *Journal of International Economics* 38 (1995), pp. 223–48.
14. Helsley, R., and W. Strange. "Matching and Agglomeration Economies in a System of Cities." *Regional Science and Urban Economics* 20 (1990), pp. 189–212.
15. Henderson, J. V. "Efficiency of Resource Usage and City Size." *Journal of Urban Economics* 19 (1986), pp. 47–90.
16. Henderson, J. V., Kuncoro, A., and M. Turner. "Industrial Development and Cities." *Journal of Political Economy* 103 (1995), pp. 1067–81.
17. Jacobs, Jane. *The Economy of Cities*. New York: Random House, 1969.
18. Marshall, Alfred. *Principles of Economics.* London: Macmillan, 1920, p. 352.
19. Mun, Seil, and Bruce G. Hutchinson. "Empirical Analysis of Office Rent and Agglomeration Economies: A Case Study of Toronto." *Journal of Regional Science* 35 (1995), pp. 437–55.
20. Rosenthal Stuart and William Strange. "The Micro-Empirics of Agglomeration." Chapter 1 in *A Companion to Urban Economics,* eds. R. Arnott and D. McMillen. London: Blackwell, 2005.
21. Rosenthal, S. S., and W. C. Strange. "The Determinants of Agglomeration," *Journal of Urban Economics* 50 (2001), pp. 191–229.
22. Rosenthal, S. S., and W. C. Strange. "Geography, Industrial Organization, and Agglomeration," *Review of Economics and Statistics* 85 (2003), pp. 377–93.

23. Rosenthal, Stuart, and William Strange. “Evidence on the Nature and Sources of Agglomeration Economies.” Chapter 49 in *Handbook of Regional and Urban Economics 4: Cities and Geography,* eds. Vernon Henderson and Jacques-Francois Thisse. Amsterdam: Elsevier, 2004.
24. Saxenian, Annalee. *Regional Advantage: Culture and Competition in Silicon Valley and Route 128.* Cambridge, MA: Harvard University Press, 1994.
25. Scott, Allen J. *On Hollywood: The Place, the Industry*. Princeton, NJ: Princeton University Press, 2005.
26. Vernon, Raymond. “External Economies.” In *Readings in Urban Economics,* eds. M. Edel and J. Rothenberg. New York: Macmillan, 1972.

CHAPTER 4

City Size

There is no need to worry about mere size. Sir Isaac Newton was very much smaller than a hippopotamus, but we do not on that account value him less.

—Bertrand Russell

New York, the largest urban area in the United States, has a population of more than 18 million, while the smallest urban area (Andrews, Texas) has a population of about 13,000. As shown in Table 4–1 (page 71), there are a few very large cities, a moderate number of medium-size cities, and many small cities. In this chapter, we'll explore the economic forces responsible for the development of cities of different size. We'll also explore why cities differ in their economic scope—from highly specialized cities to diverse cities.

UTILITY AND CITY SIZE

The previous chapter explained how agglomeration economies cause firms to cluster. As we'll see in this chapter, these agglomeration economies increase productivity and wages, so workers in larger cities earn higher wages. As a city grows, the benefits of higher wages are at least partly offset by several undesirable features of larger cities, including longer commuting times, greater density, and more congestion and pollution. Given these trade-offs, the key question is:

> *How does an increase in city size (population) affect the utility of the typical worker?*

Benefits and Costs of Bigger Cities

We are interested in the relationship between city size and the utility level of the typical worker. Consider a city where production occurs at a single point, and workers commute from a residential area to the production center. We'll start with the benefits of a larger city, then turn to the costs.

TABLE 4–1 Size Distribution of Urban Areas, 2000

Population Range	Number of Urban Areas
Greater than 10 million	2
5 million to 10 million	4
1 million to 5 million	43
100,000 to 1 million	324
Less than 100,000	549

TABLE 4–2 Utility and City Size

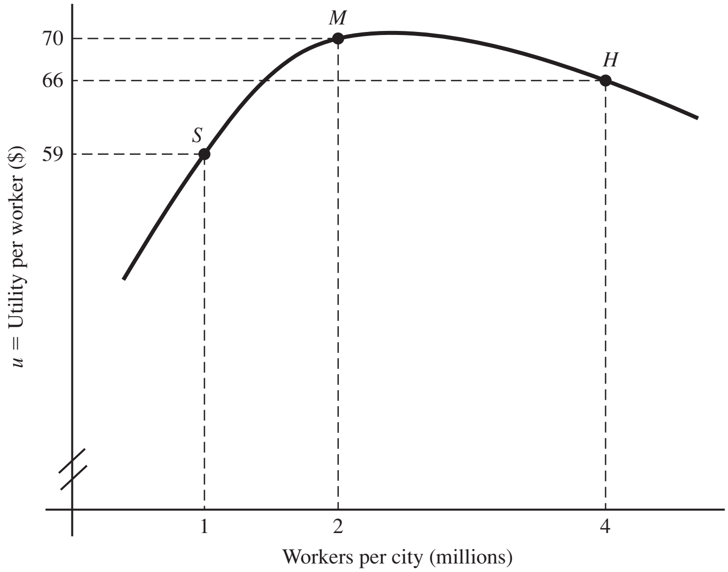
Workforce (millions)	Wage	Labor Income	Commuting Cost	Utility
1	$ 8	$64	$ 5	$59
2	$10	$80	$10	$70
4	$11	$88	$22	$66

As we saw in the previous chapter, agglomeration economies—from input sharing, labor pooling, skills matching, and knowledge spillovers—increase labor productivity. In a competitive labor market, competition among firms ensures that wages reflect labor productivity so wages are higher in larger cities. Table 4–2 provides a simple example of the relationship between wages and city size. As shown in the second column, the wage increases at a decreasing rate, reflecting the assumption that agglomeration economies diminish as the city grows: Labor productivity increases with the size of the workforce, but at a diminishing rate.

To simplify matters, let's assume that the only cost of population growth is longer commuting time. Suppose commuting time comes at the expense of leisure time, and we can place a dollar figure on the value of leisure time lost to commuting. As shown in the fourth column of Table 4–2, the cost of commuting increases with city size. A doubling of the workforce from 1 million to 2 million doubles the commuting cost per worker from $5 per day to $10. Doubling the workforce again more than doubles commuting cost.

The last column of Table 4–2 shows the utility level of the typical worker. For now, we'll define utility as income minus the value of leisure time lost from commuting. We assume that each person works an eight-hour day, so income is eight times the wage. Moving from a city of 1 million to a city of 2 million, the increase in labor income (from higher productivity) is large relative to the increase in commuting cost, so utility increases from $59 to $70. In Figure 4–1, this is shown as a move from point *S* to point *M* on the utility curve. In other words, agglomeration economies are stronger than the diseconomies of scale associated with commuting, so utility increases. In contrast, an increase in the workforce from 2 to 4 million decreases utility because agglomeration economies are weaker than the diseconomies from commuting. In Figure 4–1 the city's utility curve reaches its peak with a utility level of $70 at a workforce of 2 million.

FIGURE 4–1 Utility and City Size



An increase in city size increases wages because of agglomeration economies and increases commuting costs. As long as agglomeration economies are stronger, utility increases with city size. When agglomeration economies are weaker than the diseconomies from commuting, utility decreases as city size increases.

Locational Equilibrium, Land Rent, and Utility within a City

So far we have ignored the location decisions of workers within a particular city. Consider a city with a workforce of 2 million (shown by point *M* in Figure 4–1), where workers commute from different residential locations to the city's production center. Workers differ in their commuting costs. Recall the first axiom of urban economics:

Prices adjust to achieve locational equilibrium

In this case, the price of residential land will adjust to make workers indifferent among all residential locations.

Table 4–3 (page 73) shows how differences in commuting generate differences in land rent. Let's compare two workers, one who lives near the production center and thus incurs no commuting cost, and a second who lives 5 miles from the center and has a commuting cost of $10. If the worker at the center has a rent of $25, the rent for the other worker must be $10 lower, or $15. The same logic suggests that a worker who lives 10 miles from the center will pay only $5 for land. Workers are indifferent between the three residential locations because the differences in commuting cost are exactly offset by differences in land rent.

TABLE 4–3 Commuting, Land Rent, and Utility within a City

A Commute Distance	B Commute Cost	C Land Rent Paid	D Labor Income	E Rental Income	F Utility
0	0	$25	$80	$15	$70
5 miles	$10	$15	$80	$15	$70
10 miles	$20	$ 5	$80	$15	$70

Who gets the rent from land? Suppose that workers own land, and to keep things simple, assume that land rent is shared equally among the city's workers. The average rent is $15 (the rent paid by the worker living five miles from the center), and as shown in column E of Table 4–3, each worker earns $15 of rental income to supplement labor income. For the average worker (five miles from the center), land rent paid equals the rental income received. In contrast, for the worker living at the center, the rent paid exceeds rental income, while the opposite is true for the worker living 10 miles from the center.

The last column of Table 4–3 shows the utility for workers at different locations in the city. We can define utility as

$$\text{Utility} = \text{Labor income} + \text{rental income} - \text{commute cost} - \text{rent paid}$$

Each worker earns the same $95 of total income (labor plus rental income). For locational equilibrium, the differences in commuting costs are exactly offset by differences in land rent paid, so workers living at different locations reach the same utility level, $70.

A SYSTEM OF CITIES

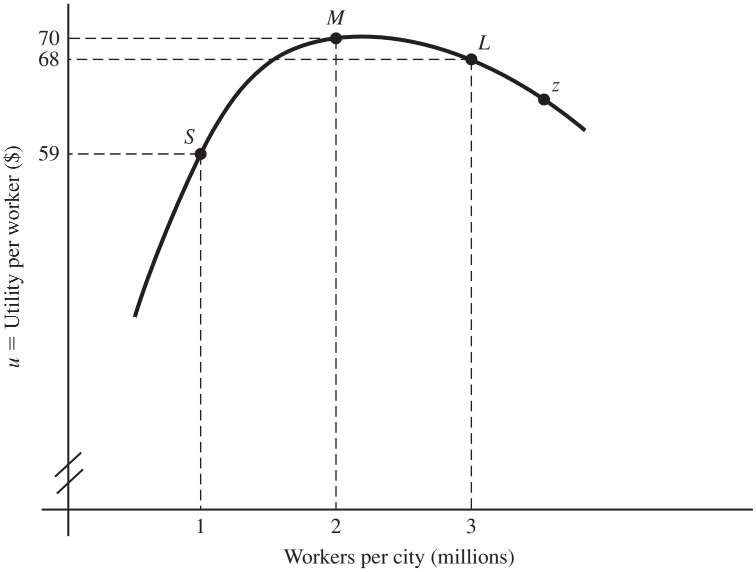
We can use the utility curve to explore how a region's workforce can be distributed among its cities. The issue is whether the region will have a large number of small cities, a small number of large cities, or something between the two extremes. Consider a region with a total urban workforce of 6 million and three possible configurations:

- Six cities {A, B, C, D, E, F}, each with a workforce of 1 million.
- Three cities {D, E, F}, each with a workforce of 2 million.
- Two cities {E, F}, each with a workforce of 3 million.

Cities Are Not Too Small

We can use Figure 4–2 to explore the feasibility of the alternative configurations. Consider first the six-city outcome, with each city housing a workforce of 1 million. As shown by point *S*, utility in each of the six cities is $59. Is this a stable equilibrium, or will workers have an incentive to move from one city to another?

FIGURE 4–2 Cities May Be Too Large But Not Too Small



The utility curve reaches its maximum with 2 million workers in a city (point *M*), so a region with 6 million workers will maximize utility with three cities, each with 2 million workers. The outcome with six small cities (point *S*) is unstable because the utility curve is positively sloped. The outcome with two big cities (point *L*) is stable because the utility curve is negatively sloped.

To demonstrate the instability of the six-city outcome, imagine that a group of workers moved from city A to city D. The workforce of city D will grow, causing the city to move upward along the positively sloped portion of its utility curve, leading to a higher utility level, for example, $60. At the same time, the workforce of city A will shrink, moving the city downward along its utility curve to a lower utility level, for example, $58. In other words, the movement of workers from city A to city D opens a utility gap of $2, which will encourage more workers to move from A to D.

Because the utility curve is positively sloped near point *S*, migration is self-reinforcing. The more workers who migrate, the larger the utility gap between the two cities and the greater the incentive to relocate. For example, if the workforce of city D grows to 1.2 million and the workforce of city A shrinks to 0.8 million, utility in city D will be about $61, compared to about $57 in city A. Recall the second axiom of urban economics:

Self-reinforcing effects generate extreme outcomes

The extreme outcome is that everyone will relocate from city A to city D, so city A will disappear.

The same logic of self-reinforcing migration applies to other cities in the region. Suppose that cities A, B, and C shrink and eventually disappear, while D, E, and F grow, each eventually doubling in size. In this case, we reach point *M* in Figure 4–2, with each of the region's three cities housing a workforce of 2 million and reaching a utility level of $70. The three-city outcome happens to be the optimum, where utility is maximized.

Cities May Be Too Large

What happens if we start out with a small number of large cities? Suppose we start with two large cities, each with a workforce of 3 million. In Figure 4–2, the starting point in each city is shown by point *L*. Each city (E, F) has a workforce of 3 million and reaches a utility level of $68. For all workers in the region, utility is lower than the maximum level. Is the two-city outcome a stable equilibrium?

To show why this is a stable equilibrium, consider the effects of migration from city E to city F. The workforce of city F will grow, so the city will move downward along the negatively sloped portion of the utility curve (from point *L* toward point *z*) and reach a lower utility level (e.g., $67). At the same time, the workforce of city E will shrink, moving the city upward along the utility curve to a higher utility level (e.g., $69). In other words, the migration of workers opens a utility gap of $2, but utility is higher in the smaller city, not the larger one.

In this case, migration is self-correcting, not self-reinforcing, and will be reversed. The migrant workers who migrated reach a lower utility level in their new city, so they'll regret the move and perhaps return to their original city. Alternatively, other workers in the larger city now have lower utility than is available in the smaller city, giving them an incentive to move, effectively swapping places with the original migrants. In either case, migration will be reversed, restoring the original workforces and utility levels in the two cities.

Why is the situation with small cities unstable, while the situation with large cities is stable? The reason is shown by the utility curve. With small cities, the utility curve is positively sloped because agglomeration economies are stronger than the diseconomies of scale resulting from commuting. The utility of the migrants increases because their new city is larger and more efficient, and the utility of people left behind decreases because they now live in a smaller and less efficient city. In contrast, when we start with large cities, the utility curve is negatively sloped because agglomeration economies are weaker than commuting diseconomies. The utility of migrants decreases because their new city is larger but less efficient, and the utility of workers left behind increases because they now live in a smaller, more efficient city.

The general lesson from this discussion is that cities tend to be too large rather than too small. The too-small outcome occurs when a region has at least one city on the positively sloped portion of the utility curve. This triggers self-reinforcing migration that eliminates some of the small cities and causes the others to grow. These self-reinforcing changes do not occur when cities are too large, so inefficiently large cities persist.

SPECIALIZED AND DIVERSE CITIES

Do cities specialize in a narrow set of economic activities, or do they generalize, producing a diverse mix of products? It turns out that the "or" in this question is misplaced: The typical region contains a wide variety of cities, from highly specialized cites to highly diverse ones (Henderson, 1988). The specialized cities develop because of localization economies, while the diverse cities develop because of urbanization economies.

In fact, specialized and diverse cities are actually complementary, serving different roles in a market economy. Many firms start their lives in a diverse city and eventually relocate to a specialized city. Diverse cities foster new ideas and experimentation so they serve as laboratories for innovative firms. Once a firm settles on a product design and production process, production is likely to be more efficient in a specialized city that fully exploits localization economies. In other words, diverse cities foster innovation, while specialized cities facilitate efficient production.

A Model of Laboratory Cities

We can use a model developed by Duranton and Puga (2001) to explore the role of cities in innovation and production. Consider a firm that is looking for the ideal production process for a new product. By experimenting with different processes, the firm will eventually find the ideal one. At that point, the firm will switch to mass production and start earning a profit. Where should the firm experiment—in a diverse city or a specialized city?

Consider first a firm that experiments in a diverse city until it discovers the ideal process and then relocates to a specialized city. An experiment entails producing a prototype of the firm's new product with a particular production process. Suppose there are six potential production processes, and the experimenting firm can observe other firms using these processes in the diverse city, and then imitate these other firms to produce a prototype. In addition, suppose that, on average, it takes the firm three years to discover the ideal process. Once the entrepreneur discovers the ideal process, the firm will move to a specialized city and start making a profit.

The alternative scenario is to experiment—search for the ideal process—in the region's specialized cities. As usual, there are some trade-offs:

- **Good news: lower prototype cost.** The cost of producing a given prototype will be lower in a specialized city because each city has the specialized inputs for that production process.
- **Bad news: higher moving cost.** The search for the ideal process requires the firm to move from one specialized city to another. On average, a firm adopting this strategy will have three moves, compared to a single move for the firm that experiments in the diverse city. If moving costs are large relative to the savings in prototype costs, profit will be lower when the firm experiments in specialized cities.

This model of laboratory cities shows the roles of diverse and specialized cities in the product cycle. A diverse city has a rich variety of products and production processes, providing fertile ground for new ideas about how to produce new products. Once a firm finds its ideal production process, the benefits of being in a diverse environment diminish, so the firm relocates to a specialized city, where localization economies generate lower production cost.

Example: The Radio Industry in New York

Vernon (1972) identifies the radio industry in New York as the classic example of an industry that developed in a diversified city:

> In the 1920s that industry had all the earmarks of an activity whose establishments were heavily dependent on external [agglomeration] economies, speed, and personal contact. Its technology was unsettled and changing rapidly; its production methods were untried; its market was uncertain. Accordingly, at that stage, producers were typically small in size, numerous, agile, nervous, heavily reliant upon subcontractors and suppliers. Mortality in the industry was high. In those circumstances, the attraction of an urban area like the New York Metropolitan Region was especially strong.

New York was attractive because it provided a wide variety of intermediate inputs and a large and diverse workforce. The area also provided production knowledge—embodied in a wide variety of production processes—that proved useful in developing a production process for the radio.

Vernon explains why the radio industry eventually left the New York metropolitan area:

> A decade or two later, however, the technology of the industry had settled down. Production methods were standardized and sets were being turned out in long runs. Now, the critical competitive questions had become transport and labor costs, rather than product design. The small firm faded from the picture and large assembly plants appeared at lower-range locations more centrally placed for national markets.

When a product reaches its mature stage, with a settled design and established production process, producers have less to gain from diversified cities, and they can relocate to places with lower production costs because of localization economies, lower wages, or lower land rent.

Evidence of Laboratory Cities

Duranton and Puga (2001) provide evidence for the notion that diversified cities serve as laboratories for firms in innovative industries. Using data from firms in France, they show that among firms that change locations, over 7 in 10 relocated from a diverse city to a specialized city. As firms mature, they relocate from diverse cities with urbanization economies to specialized cities with localization economies.

The most innovative industries have the highest frequency of relocations from diverse to specialized cities. For example, the frequency is 93 percent for research and development, 88 percent for pharmaceuticals and cosmetics, and 82 percent for information technology. The frequency is high for other industries, including business services, printing and publishing, aerospace equipment, and electronic equipment. In contrast, the frequency of movement to specialized cities is relatively low for less innovative sectors such as furniture, food, beverages, clothing, and leather.

Much of the recent work on urbanization economies focuses on how city diversity affects plant births and employment growth. Rosenthal and Strange (2004) summarize the most recent studies and conclude that diversity encourages both births and growth in employment, especially in high-technology industries.

DIFFERENCES IN CITY SIZE

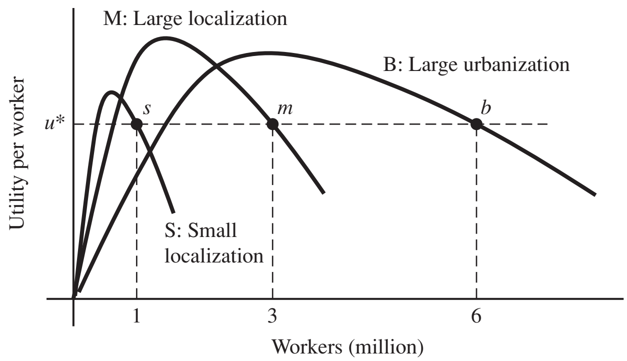
Table 4–1 shows the wide range of city sizes in the United States. Other countries have a similar pattern of differences in city sizes. In this part of the chapter, we'll explore the roles of localization and urbanization economies in determining city size. We'll also explore the role of consumer goods.

Differences in Localization and Urbanization Economies

Figure 4–3 shows utility curves for three types of cities in a regional economy. The curve on the left applies to an industry for which localization economies are exhausted with a relatively small workforce. In this case, the diseconomies of commuting quickly overwhelm agglomeration economies, so the optimum city size is relatively small. The middle utility curve is for a specialized city with larger localization economies and thus a larger optimum size. Finally, the utility curve on the right is for a city that experiences large urbanization economies, generating a large optimum size.

Locational equilibrium requires that workers in the region are indifferent between the three cities, meaning that workers in the three cities must achieve the same

FIGURE 4–3 Differences in City Sizes from Differences in Agglomeration Economies



City S has small localization economies and its optimum population is smaller than that of city M with its large localization economies. City B has large urbanization economies and a large population. The set of points $\{s, m, b\}$ shows a possible equilibrium, with all residents achieving the utility level u^* and populations of 1 million (city S), 3 million (city M), and 6 million (city B) adding up to the regional population of 10 million.

utility level. Suppose the region has a total of 10 million workers. In Figure 4–3, points *s*, *m*, and *b* show one possible equilibrium. The utility level, u^* is the same in all three cities, and the city workforces (1 million in the small city, 3 million in the medium-size city, and 6 million in the big city) add up to the regional workforce (10 million). As we saw earlier in the chapter, this is a stable equilibrium because each city is on the negatively sloped portion of its utility curve.

Local Goods and City Size

So far our discussion of city size has focused on employment in industries subject to localization and urbanization economies. In other words, we have ignored the consumer side of the urban economy. We can distinguish between employment in industries that export their output to people outside the city and those that sell their products locally to residents of the city. For example, most of the cars produced in Detroit are sold to people in other cities, while most of the donuts produced in Detroit are sold to city residents, as are most haircuts and groceries. Total employment in a city is the sum of export employment and local employment.

Some local products are available in all cities, large and small. If the per-capita demand for a product is large relative to the scale economies associated with producing it, even a small city will generate sufficient demand to support at least one firm. For example, it takes just a few thousand people to support a barber, so even a small city will have at least one barber. Similarly, a pizzeria can be supported by a few thousand people, so even a small city will have pizzerias and pizza workers. Of course, a larger city has more hair to cut and more people to feed, so it will have more barbers and pizza tossers. In fact, we expect the number of barbers and pizza tossers to increase proportionately with city size.

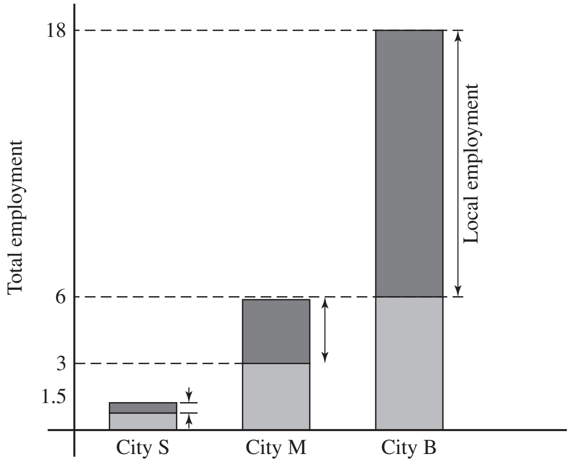
Some local products are available only in large cities. If the per-capita demand for a product is low relative to the scale economies in production, it will take a large city to generate enough demand to support a firm. For example, the per-capita demand for opera is relatively small, so it may take a million people to support an opera company. As a result, we will find opera companies in large cities but not small ones. Similarly, the per-capita demand for brain surgery is low relative to the scale economies in production, so brain surgeons operate in large cities.

Larger cities have a wider variety of consumer products. In a large city, consumers can get everything available in small cities (pizzas and haircuts) as well as products not available in small cities (opera and brain surgery). In fact, people in small cities travel to large cities to buy products that are not available locally. In contrast, consumers in large cities can buy almost anything they want locally so there is little reason to travel to smaller cities.

Local Employment Amplifies Size Differences

Figure 4–4 (page 80) shows the implications of local employment for the size of cities with different levels of export employment. Suppose that in a city with export employment of 1 million, each export job supports half a job in local industry. If so,

FIGURE 4–4 The Introduction of Local Goods Amplifies Differences in City Size



Introducing local consumer amplifies the differences in population arising from differences in export employment. The population of the small city increases by half, while the population of the medium city doubles and the population of the large city triples.

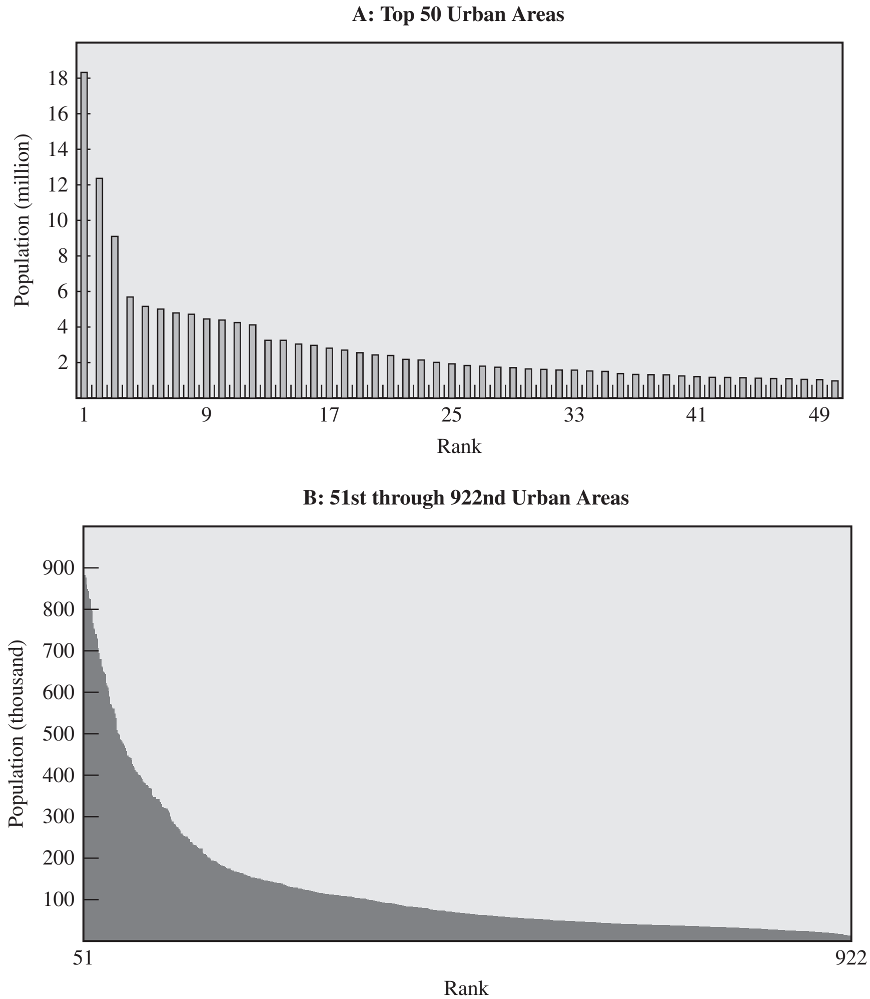
total employment in the city is 1.5 times its export employment, or 1.5 million. We know that a larger city can support a wider range of consumer goods. Suppose that in a city with export employment of 3 million, each export job supports one local job. As shown in Figure 4–4, total employment in such a city is 6 million (3 million export jobs plus 3 million local jobs). Finally, suppose that in a city with 6 million export jobs, each export job supports two local jobs. In this case, total employment is 18 million (6 million export jobs plus 12 million local jobs).

As shown in Figure 4–4, local employment amplifies differences in population. Total employment increases by half in the small city (1 to 1.5 million), while it doubles in the medium-sized city (from 3 to 6 million) and triples in the large city (from 6 to 18 million). After incorporating local employment, the largest city has 12 times the total employment of the small city, up from six times before considering local employment. This occurs because the large city has a larger consumer base and can support a wider variety of products.

THE SIZE DISTRIBUTION OF CITIES

Figure 4–5 shows the size distribution of cities split into two graphs, one for the top 50 urban areas and a second for the rest of the urban areas. The figure is drawn with the new census definitions of urban areas, with metropolitan areas (urban areas with populations exceeding 50,000) and micropolitan areas (smaller urban areas).

FIGURE 4–5 Size Distribution of U.S. Urban Areas, 2000



The Rank-Size Rule

Geographers and economists have estimated the relationship between city rank and size. One possibility is that the relationship follows the rank-size rule:

Rank times population is constant across cities

In other words, if the largest city (rank 1) has a population of 24 million, the second largest city will have a population of 12 million ($12 \cdot 2 = 24$), while the third largest will have a population of 8 million ($8 \cdot 3 = 24$), and so on.

Nitsche (2005) analyzes the results of 29 studies of the rank-size relationship, with data from countries around the world. The hypothesized relationship is

$$Rank = \frac{C}{N^b}$$

where C is a constant, N is population, and the exponent b is to be estimated from the data on rank and population. If $b = 1.0$, the rank-size rule holds. In the studies considered by Nitsche, two-thirds of the estimates of b are between 0.80 to 1.20, and the median estimate is 1.09. This is consistent with earlier cross-country studies that generate estimates of b in the range 1.11 to 1.13. In other words, the urban population is more evenly distributed across cities than would be predicted by the rank-size rule.

An important qualification of this conclusion is that many of the studies use political definitions of cities rather than the economic definition. A political city is defined by boundaries that separate political jurisdictions. In contrast, the economic definition of a city ignores political boundaries and includes in a city's population all the people who are economically involved in a particular urban economy. In practical terms, the economic city is typically defined as a metropolitan area (e.g., the San Francisco Bay Area) that includes the central (political) city along with all the surrounding communities. For the studies of the rank-size rule that use economic cities rather than political cities, the median estimate for b is 1.02, which is much closer to the rank-size result.

Urban Giants: The Puzzle of Large Primary Cities

In many developing countries, the largest city has a relatively large share of population. Table 4–4 shows the populations and the national population shares for the largest cities in several countries. As a point of reference, the New York metropolitan area has only 6.5 percent of the population of the United States. In contrast, the other metropolitan areas listed in the table have national shares between 11 percent and 39 percent.

Economists have developed several models to explain the large concentrations of population in primary cities. One theory is based on the idea that large economies of scale in trade encourage the development of a single large trading city rather than several smaller ones. For example, the substantial investment in a port facility

TABLE 4–4 Population of Largest Cities as Share of National Population

Metropolitan Area	Population	Share of National Population
Tokyo	19,037,361	15.76%
Mexico City	16,465,487	20.97
São Paulo	15,538,682	11.46
Buenos Aires	10,759,291	35.47
Santiago, Chile	4,227,049	34.87
Montevideo, Uruguay	1,157,450	39.36

encourages nations to designate a primary port city. English trade grew rapidly between 1520 and 1670, a result of military victories against the Spanish, improved shipping technology, and the expansion of markets in Asia and the Americas. The population of London increased during this period from 55,000 to 475,000, and its share of England's population increased from 2 percent to 10 percent. In Argentina, between 1887 and 1914 exports quadrupled, and the population of Buenos Aires—the principal trading city—increased by 1.1 million (265 percent).

Transportation infrastructure plays a role in urban concentration. In many developing countries, a disproportionate share of investment in roads and telecommunication facilities occurs in and around the capital city. The relatively low investment in infrastructure outside the capital area generates high transport costs within the country and encourages the development of large primary cities.

What is the role of politics in the development of large primary cities? Ades and Glaeser (1995) suggest that nations run by dictators have larger primary cities than democracies. One way for a dictator to stay in power is to take resources from the hinterland (areas outside the capital city) and transfer these resources to the people who are most likely to overthrow him—people in the capital city. If a dictator pays off local agitators, the capital city will grow as some people migrate to the city to get the payoffs and others migrate for jobs in local industry that are supported by the bribes. Based on a study of cities in 85 countries, Ades and Glaeser conclude that capital cities in countries with dictatorships are 45 percent larger than capital cities in other countries.

The experience of Rome illustrates the role of redistribution policies on urban concentration. In the period 130–50 B.C., the population of Rome increased from 375,000 to 1 million, making Rome more than twice as large as any city up to that point in history. Military successes during this period extended the empire into Gaul and the eastern provinces of Asia, providing a large hinterland from which to extract resources. The rulers responded to political unrest in Rome by distributing free grain to the residents and staging the infamous—and very expensive—Roman circuses.

SUMMARY

This chapter explains why cities come in different sizes. Here are the main points of the chapter:

1. The utility curve shows the trade-offs from an increase in population: Agglomeration economies increase productivity and wages, but diseconomies of scale from increased commuting costs reduce utility.
2. Cities are unlikely to be too small because such an outcome is not a stable equilibrium: Migration is self-reinforcing because the growing city becomes more productive while the shrinking city becomes less productive.
3. Cities may be too large because such an outcome is a stable equilibrium: Migration is self-correcting because the shrinking city becomes more productive while the growing city becomes less productive.

4. Local employment amplifies differences in workforces and population across cities.
5. A diverse city fosters experimentation and leads to innovations in product design and production.
6. The rank-size rule provides a rough approximation to the size distribution of cities.

APPLYING THE CONCEPTS

For exercises that have blanks (_____), fill each blank with a single word or number. For exercises with ellipses (. . .), complete the statement with as many words as necessary. For exercises with words in square brackets ([increase, decrease]), circle one of the words.

1. **Migration to Portland**
   In the last several months, the number of cars driving in Portland with Washington State license plates has increased dramatically. Suppose this reflects migration from Seattle to Portland. Use two utility curves, one for Portland and a second for Seattle, to represent this migration. Assume that both cities are on the negatively sloped portions of their utility curves.
   a. Label the current position for Portland (population 2 million) with "P" and the current position for Seattle (population 3 million) with "S." Use arrows to indicate the direction of movement along each utility curve.
   b. In Portland, migration [decreases, increases] utility because _____ of scale from _____ dominate _____ of scale from __________.
   c. Label the long-run equilibrium point for Portland as "Q" and the long-run equilibrium point for Seattle as "T."
2. **Heli-Segways for Workers**
   Consider a region with 6 million workers and two cities (A and B), each with an initial workforce of 3 million (the utility-maximizing workforce). Suppose the heli-segway (a flying personal transporter) replaces the automobile, cutting commuting cost and increasing the utility-maximizing urban workforce per city to 5 million. The positively sloped part of the new utility curve is steeper than the negatively sloped part.
   a. Show the effect of the heli-segway on the urban utility curve in the typical city.
   b. A workforce of 3 million [is, isn't] a stable equilibrium because. . . .
   c. Using the new utility curves, show the changes in the workforces of two cities, one that grows and one that shrinks. Label the initial point *i* and indicate the directions of the changes with arrows.
   d. The new equilibrium number of cities is _____ and each city has a population of _____.
3. **A Free Circus and City Size**
   Consider a region with a fixed population of 2 million. The urban utility curve reaches the maximum utility with 1 million people, and initially there are two

cities, R and S, each with 1 million people. The positively sloped portion of the urban utility curve is steeper than the negatively sloped portion. Suppose a dictator in city R starts to provide free circuses, financed by coersive transfer payments from people outside the region.

a. Show the effect of the free circuses on the urban utility curve in city R. Assume that the the utility-maximizing population doesn't change.

b. People will migrate from city _____ to city _____ because. . . .

c. Arrows up or down: Migration _____ utility in city R and _____ utility in city S.

d. Show the new equilibrium distribution of population betweeen the two cities. The population in city R is _____ and the population in city S is _____

4. Formation of New Cities

Consider a region with a workforce of 12 million in a single city. The urban utility curve reaches its maximum with 3 million workers and includes the following combinations (W = workers; U = utility in $):

W	1	2	3	4	5	6	7	8	9	10	11	12
U	32	56	70	65	60	55	50	45	40	35	30	25

Suppose the government establishes a new city with 1 million workers, leaving 11 million workers in the old city. Assume that the number of cities remains at two.

a. Immediately following the establishment of the new city, the utility in the small new city is _____ and the utility in the large old city is _____

b. On the utility curve, mark the position of the new city with "N" and the position of the old city (immediately following the formation of the new city) with "D." Use arrows to indicate the direction of movement for each city.

c. In the long-run equilibrium, the workforce of the new city = _____ with utility = _____; the workforce of the old city = _____ with utility = _____

5. New Cities Boom or Bust?

Consider a region with a workforce of 12 million in a single city. The urban utility curve reaches its maximum with 6 million workers and includes the following combinations (W = workers; U = utility in $):

W	1	2	3	4	5	6	7	8	9	10	11	12
U	23	33	43	48	52	55	50	45	40	35	30	25

Suppose the government establishes a new city with 1 million workers, leaving 11 million workers in the old city. Assume that the number of cities remains at two.

a. Immediately following the establishment of the new city, the utility in the small new city is _____ and the utility in the large old city is _____, so the new city will [shrink, grow] and the old city will [shrink, grow].

b. In the long-run equilibrium, the workforce in the old city = _____.

c. A new city will be viable in the long run if the starting workforce is at least _____ because. . . .

6. Rural versus Urban Development

In the initial equilibrium, a nation's workforce of 24 million is divided equally between a city and a rural area. The initial common utility level is \$50. The urban utility curve has the conventional hump shape, with a peak at 8 million workers. In the workforce range 8–12 million, the slope of the utility curve is –\$2 per million workers. The rural utility curve is subject to mild increasing returns to scale. In the workforce range 12–16 million, the slope is \$1 per million workers. Suppose the nation invests in rural infrastructure, increasing rural productivity and shifting the rural utility curve upward by \$3.

a. Depict graphically the effects of this infrastructure investment on the two sectors.

b. Workers will migrate from _____ to _____ because. . . .

c. Migration [increases, decreases] utility in the rural area and [increases, decreases] utility in the city, with a [larger, smaller] change in the [city, rural area].

d. Show the new equilibrium distribution of the workforce between the rural area and the city. In the new equilibrium, the workforce of the city is _____; the workforce of the rural area is _____; the common utility level is _____.

7. Size Distribution Example

Consider a region with three industries subject to localization economies, each of which employs 300 workers. There is a single cluster for industry L (all firms in the industry locate in a single city labeled "L"), three clusters for industry M (firms are divided equally between three cities, labeled "M1," "M2," "M3"), and five clusters for industry S (firms are divided equally between five cities, labeled "S1," "S2," . . . , "S5"). There are no urbanization economies. In a city with 300 export jobs, there are three local jobs for each export job; in a city with 100 export jobs, there are two local jobs for each export job; in a city with 60 export jobs, there is one local job for each export job.

a. Complete the following table.

	Export Workers	Local Workers	Total
City S	_________	_________	_________
City M	_________	_________	_________
City L	_________	_________	_________

b. Compared to the smallest city, the largest city has _____ times the export employment but _____ times the total employment. The difference in total employment is relatively large because. . . .

8. Specialized Services in Large Cities

Complete the statement: Large urban areas provide specialized cultural, legal, medical, financial, and other services that are not available in small urban areas because these specialized services are characterized by. . . .

9. **Boring Capital Cities**
   Some people claim that state capitals (Sacramento, CA; Salem, OR; Olympia, WA) are boring in the sense that they have a smaller variety of goods and services than other cities of equal size. Check a map and then provide an explanation why these cities could be boring.
10. **One City Size**
   Consider a region with two export products (gloves and socks) and two local goods (tattoos and manicures). The production of each export good is subject to localization economies, so each city specializes in one export good. According to Mr. Wizard, "If my two assumptions (one for export products and one for local goods) are correct, all the cities in the region will be the same size." Assume that Mr. Wizard's logic is correct. List his assumptions and explain why together they imply the region's cities will be the same size.

REFERENCES AND ADDITIONAL READING

1. Ades, Alberto F., and Edward L. Glaeser. "Trade and Circuses: Explaining Urban Giants." *Quarterly Journal of Economics* (1995), pp. 195–227.
2. Abdel-Rahman, H., and A. Anas. "Theories of Systems of Cities." Chapter 52 in *Handbook of Regional and Urban Economics 4: Cities and Geography,* eds. V. Henderson and J. F. Thisse. Amsterdam: Elsevier, 2004.
3. Audretsch, D., and M. Feldman. "Knowledge Spillovers and the Geography of Innovation," Chapter 61 in *Handbook of Regional and Urban Economics 4: Cities and Geography,* eds. V. Henderson and J. F. Thisse. Amsterdam: Elsevier, 2004.
4. Duranton, G., and D. Puga. "Nursery Cities: Urban Diversity, Process Innovation, and the Life Cycle of Products." *American Economic Review* 91 (5), 1454–77 (2001).
5. Gabaix, X., and Y. Ioannides. "Evolution of City Size Distributions." Chapter 53 in *Handbook of Regional and Urban Economics 4: Cities and Geography,* eds. V. Henderson and J. F. Thisse. Amsterdam: Elsevier, 2004.
6. Glaeser, E. "Learning in Cities." *Journal of Urban Economics* 46 (1999), pp. 254–77.
7. Henderson, V. *Urban Development: Theory, Fact, and Illusion*. Oxford: Oxford University Press, 1988.
8. Nitsche, V. "Zipf Zipped." *Journal of Urban Economics* 57 (2005), pp. 86–100.
9. Rosenthal, S., and W. Strange, "Evidence on the Nature and Sources of Agglomeration Economies." Chapter 49 in *Handbook of Regional and Urban Economics 4: Cities and Geography,* eds. V. Henderson and J. F. Thisse. Amsterdam: Elsevier, 2004.
10. U.S. Department of Transportation. *Summary of Travel Trends: 2001 National Household Travel Survey.* Washington D.C., December 2004.
11. Vernon, Raymond. "External Economies." In *Readings in Urban Economics,* ed. M. Edel and J. Rothenberg. New York: Macmillan, 1972.

Appendix: Central Place Theory

Central place theory, which was developed by Christaller (translated in 1966) and refined by Losch (translated in 1954), shows how the location patterns of different industries combine to form a regional system of cities, with a small number of large cities and a large number of small cities.

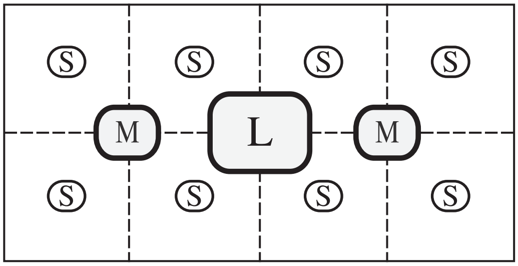
The starting point for central place theory is the observation that the market areas of individual firms vary across industries. If the economies of scale in production are large relative to per-capita demand for the product, each firm will need a relatively large area to sell a large enough quantity to exhaust its scale economies. For example, the per-capita demand for brain surgery is low, and the specialized equipment used in brain surgery generates large economies of scale. As a result, a brain-surgery center draws patients from a large area and locates in a large city. In contrast, the per-capita demand for haircuts is high relative to the scale economies in haircutting, so the market area of a barber shop is small and even small towns have at least one barber shop.

Consider a region with three consumer products: books, pizzas, and jewelry. The region has the following characteristics:

1. **Population density.** The initial distribution of population is uniform. The total population of the region is 80,000.
2. **No shopping externalities.** Shopping externalities normally occur with complementary goods (one-stop shopping) and imperfect substitutes (comparison shopping). The simple central place model assumes that there are no shopping externalities.
3. **Ubiquitous inputs.** All inputs are available at all locations at the same prices.
4. **Uniform demand.** For each product, per-capita demand is the same throughout the region.
5. **Number of stores.** The three goods have different per-capita demands and scale economies:
   - **a. Jewelry.** Scale economies are large relative to per-capita demand. Every jewelry store requires a population of 80,000, so a single jeweler will serve the entire region.
   - **b. Books.** Scale economies are moderate relative to per-capita demand. Each bookstore requires a population of 20,000, so there will be four bookstores in the region.
   - **c. Pizza.** Scale economies are small relative to per-capita demand. Every pizza parlor requires a population of 5,000, so there will be 16 pizza parlors in the region.

In the central place model, firms base their location decisions exclusively on access to their consumers. Because production costs are the same at all locations (inputs are ubiquitous), the jeweler will minimize its total costs by minimizing its travel costs. Because population density is uniform, travel costs are minimized at

FIGURE 4A–1 The Central Place Hierarchy



City L	M Cities	S Cities
1 Jeweler	1 Book store	1 Pizzeria
2 Book stores	2 Pizzerias	
4 Pizzerias		

There are 11 cities in the region, one large city (L), two medium-sized cities (M), and eight small cities (S). The larger the city, the greater variety of goods sold.

the center of the region and the jeweler will locate there. A city will develop around the jewelry store. Jewelry workers will locate near the store to economize on commuting costs. The population density near the jeweler will increase, generating a city (a place of relatively high density) at the center of the region. In Figure 4A–1, a city develops at point *L*.

The bookstores will carve up the region into market areas, causing the development of additional cities. If the region's population density were uniform, bookstores would carve out four equal market areas. However, because there is a city surrounding the jeweler in the center of the region, there will be enough demand to support more than one bookstore in city L. Suppose city L, along with the surrounding area, has enough people to support two bookstores. The two other bookstores will split the rest of the region into two market areas. In Figure 4A–1, two more cities develop at the locations marked with an *M*.

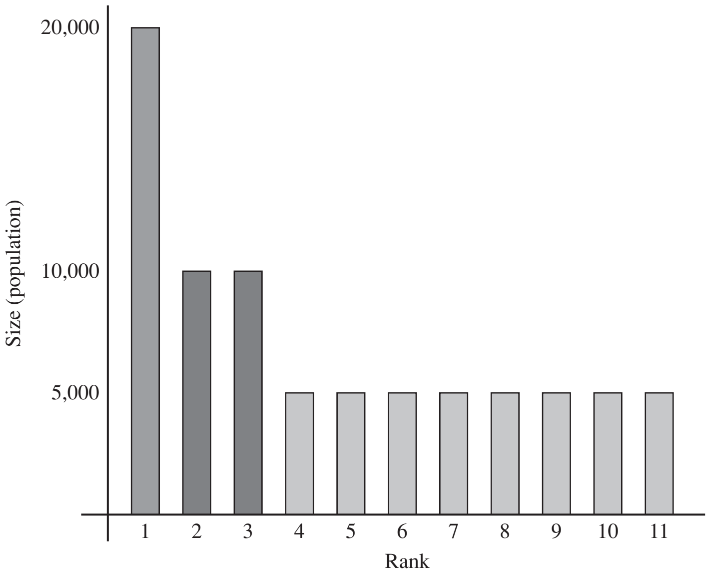
The pizza parlors will also carve up the region into market areas, causing the development of more cities. Because the population density is higher in the cities that develop around the jewelry store and the bookstores, there will be more than one pizza parlor in city L and the two M cities. Suppose L has a population large enough to support four pizza parlors, and each of the M cities has a population large enough to support two pizza parlors. A total of eight pizza parlors will locate in cities L and M, and the remaining eight pizza parlors will divide the rest of the region into eight market areas. As shown in Figure 4A–1, the eight additional cities (marked with "S") sell pizza.

The rectangular region has a total of 11 cities. The large city at the center of the region sells jewelry, books, and pizza. City L has a population of 20,000, meaning that it is large enough to support four pizza parlors (5,000 people per pizza parlor). The city sells books to consumers from the four surrounding S cities, so the total number of book consumers is 40,000 (20,000 from L and 5,000 each from four S cities), enough to support two bookstores. The two medium-sized cities sell books and pizza. Each of the M cities has a population of 10,000, meaning that each city is large enough to support two pizza parlors. Each city sells books to consumers from two nearby S cities, so the total number of book consumers in each M city is 20,000 (10,000 from M and 5,000 each from two S cities), enough to support one bookstore per M city. Each of the S cities has a population of 5,000, meaning that each city can support one pizza parlor.

Figure 4A–2 shows the size distribution of cities in the region. The vertical axis measures city size (population), and the horizontal axis measures the rank of the city. The largest city (L) has a population of 20,000; the second and third largest cities (M cities) have populations of 10,000; and the 4th through the 11th largest cities have populations of 5,000.

The simple central place model generates a hierarchical system of cities. There are three distinct types of cities: L (high order), M (medium order), and S (low order). The larger the city, the greater the variety of goods sold. Each city imports goods from higher-order cities and exports goods to lower-order cities. Cities of the same order do not interact. For example, an M city imports jewelry from L and exports books to S cities, but does not interact with the other M city. Similarly, an S city imports jewelry and books from larger cities, but does not trade with other S cities. The system of cities is hierarchical in the sense that there are distinct types of cities and distinct patterns of trade dominance.

FIGURE 4A–2 The Size Distribution of Cities with a Simple Central Place Model



The simple central place model generates one large city (L) with a population of 20,000, two medium-sized cities, each with a population of 10,000, and eight small cities, each with a population of 5,000.

The simple central place model provides some important insights into how the market-area decisions of firms combine to generate the urban hierarchy.

1. **Diversity and scale economies.** The region's cities differ in size and scope. This diversity occurs because different scale economies relative to per-capita demand generate different sized market areas. In contrast, if the three goods had the same scale economies relative to per-capita demand, the region would have 16 jewelers, 16 bookstores, and 16 pizza parlors. The market areas of the three goods would coincide, so the region would have 16 identical cities, each of which would provide all three goods.
2. **Large means few.** The region has a small number of large cities and a large number of small cities. A large city provides more goods than a smaller city, and the extra goods provided by a large city are those goods that are subject to relatively large scale economies. A small number of stores sell the goods subject to relatively large scale economies, so few cities can be large. In our example, city L is the largest city because that's where the single jewelry store locates.
3. **Shopping paths.** Consumers travel to bigger cities, not to smaller cities or cities of the same size. For example, consumers from an M city travel to L to buy jewelry, but do not travel to the other M city or to an S city to consume books or pizza. Instead, they buy these goods in their own city.

Although central place theory provides some important insights into the hierarchy of cities, its focus on consumer goods limits its applicability to real cities. For many firms, the location decision is based in part on the cost of local inputs (e.g., labor, raw materials, intermediate goods) and agglomeration economies—localization and urbanization economies. Central place theory ignores these other location factors, so it applies to only part of the urban economy.

REFERENCES

1. Christaller, Walter. *Central Places in Southern Germany,* trans. C. W. Baskin. Englewood, NJ: Prentice Hall, 1966.
2. Losch, August. *The Economics of Location.* New Haven, CT: Yale University Press, 1954.

CHAPTER 5

Urban Growth

An economic forecaster is like a cross-eyed javelin thrower: He doesn't win many accuracy contests, but he keeps the crowd's attention.

—Anonymous

In an urban economy, there are two sorts of growth. First, economic growth is defined as an increase in a city's average wage or per-capita income. Second, employment growth is defined as an increase in a city's total workforce. In this chapter, we explore the various sources of income and employment growth and look at the consequences of increases in a city's total employment. One of the key questions is Who benefits when total employment increases?

ECONOMIC GROWTH: INCREASE IN PER-CAPITA INCOME

Economic growth is defined as an increase in per-capita income. The traditional—nongeographical—sources of economic growth are as follows:

- **Capital deepening.** Physical capital includes all the objects made by humans to produce goods and services, such as machines, equipment, and buildings. Capital deepening is defined as an increase in the amount of capital per worker—it increases productivity and income because each worker works with more capital.
- **Increases in human capital.** A person's human capital includes the knowledge and skills acquired through education and experience. An increase in human capital increases productivity and income.
- **Technological progress.** Any idea that increases productivity—from a worker's commonsense idea about how to better organize production, to a scientist's invention of a faster microprocessor—is a form of technological progress. The resulting increase in productivity increases income per worker.

As we saw earlier in the book, the geographical perspective adds a fourth source of economic growth.

- **Agglomeration economies.** Physical proximity increases productivity through input sharing, labor pooling, labor matching, and knowledge spillovers.

Cities increase productivity and income because they bring the inputs to the production process together and facilitate face-to-face communication. According to Lucas (2001), cities are the engines of economic growth.

It is important to distinguish between a change in a city's income level and a change in its growth rate of income. Suppose a city's annual per-capita income increases from $20,000 to $21,000 and then remains at the higher level. This city has experienced an increase in its income level but no change in its long-term economic growth rate, which is zero. In contrast, suppose a city whose income was growing at a rate of 1 percent per year starts to grow at a rate of 3 percent per year. If the higher growth rate persists, the city has experienced an increase in its long-term economic growth rate. A city's economic growth rate is determined by the rate of capital deepening (how rapidly capital per worker increases each year), the rate of technological progress (how many new ideas are developed each year), and the rate of increase in human capital.

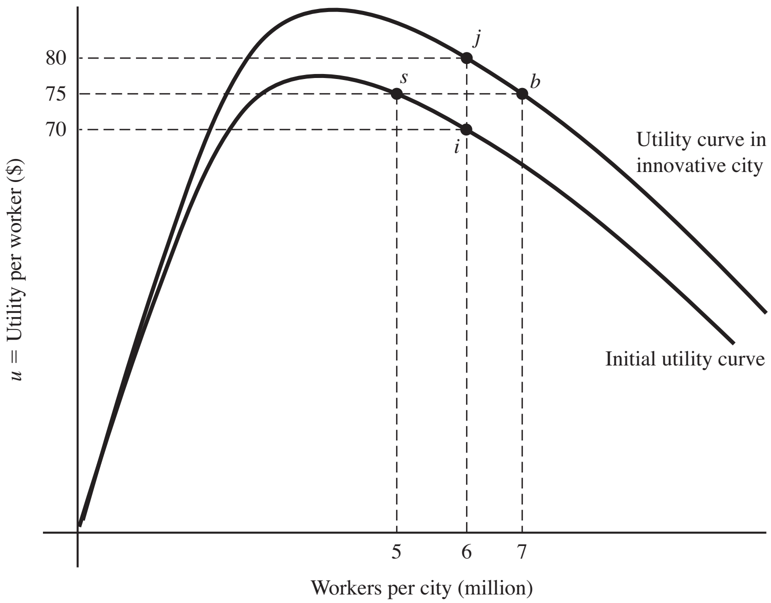
To illustrate the distinction between level and growth effects, consider the effects of an increase in human capital. Suppose the fraction of a city's population that completes college increases from 30 to 35 percent and remains at the higher level. If the resulting increase in productivity increases per-capita income from $20,000 to $21,000, we measure economic growth as a $1,000 increase in the city's income level. By itself, the increase in human capital does not affect the city's long-term growth rate. However, if better-educated people generate more and better ideas every year, the rate of technological progress will increase, leading to a higher long-term economic growth rate.

CITY-SPECIFIC INNOVATION AND INCOME

We can use the urban utility curve derived in Chapter 4 to show the connection between technological progress and per-capita income. Consider a region of 12 million workers and two cities that are initially identical. In Figure 5–1 (page 94), the initial utility curve has the familiar hump shape, reflecting the tension between agglomeration economies and diseconomies of scale (from rising commuting times; more congestion, noise, and pollution; and higher density). The two cities have the same initial utility curve, and the initial equilibrium is shown by point *i*. The region's workforce is split equally between two cities of 6 million workers, and the common utility level is $70 per worker.

Suppose one of the two cities experiences technological progress that increases worker productivity. In Figure 5–1, the city's utility curve shifts upward, with a higher productivity (and average income) for each level of the workforce. For example, with a workforce of 6 million in the innovative city, the income level would increase from $70 (point *i*) to $80 (point *j*). So in the absence of migration, utility in

FIGURE 5–1 Urban Economic Growth from Technological Progress



In the initial equilibrium shown by point *i*, a region's workforce is divided equally between two cities of 6 million workers. Innovation in one city shifts its utility curve upward, and in the absence of migration, the innovative city moves to point *j*. Migration to the innovative city generates points *b* (innovative city) and *s* (other city). The innovation increases utility in both cities and shifts population to the innovative city.

the innovative city would exceed utility in the other city by $10. In response to the utility gap, workers will migrate to the innovative city, and migration will continue until utility is equalized.

The new equilibrium is shown by points *s* and *b*. This is a locational equilibrium because both cities have the same utility level ($75), and the workforces in the two cities add up to the fixed regional population: The innovative city (shown by point *b*) gains 1 million workers, while the other city (shown by point *s*) loses the same number. Utility increases from $70 to $75 in both cities, meaning that workers in both cities benefit from innovation in one city. Workers in the other city benefit because the decrease in population causes the city to move upward along its negatively sloped utility curve to a higher utility level.

One of the lessons from Figure 5–1 is that the benefits of innovation in a single city spread to other cities in the region. Any initial gap in the utility levels of cities will be eliminated by labor migration to the city with the higher utility level, and migration will continue until the utility gap is eliminated. In our two-city region, the initial utility gap (shown by points *i* and *j*) is $10, and in equilibrium, each of the two cities experiences an increase in utility equal to half this initial gap ($5): Utility increases from $70 to $75.

In a larger region, the increase in per-capita utility would be smaller. If the region had 10 cities instead of two, there would be five times as many workers to share the benefits of the innovation. For example, the initial \$10 utility gap, spread over 10 cities, translates into a \$1 increase in per-capita utility.

Regionwide Innovation and Income

Consider next the effect of simultaneous innovation in both cities. Suppose the two cities experience the same innovation, and thus experience the same upward shift of the utility curve. In this case, both cities would move from point *i* to point *j*, and point *j* would be the new equilibrium. There would be no utility gap to overcome with migration because both cities would experience the same change in productivity. As a result, each city would maintain its workforce of 6 million workers.

We've seen that technological innovation—represented by an upward shift of the utility curve—increases the equilibrium utility and per-capita income throughout the region. The same logic applies to other sources of higher productivity—capital deepening, increases in human capital, and productivity boosts from localization and urbanization economies.

HUMAN CAPITAL AND ECONOMIC GROWTH

Urban economists have explored the effects of human capital on urban productivity and income. An increase in the education or job skills of a specific worker increases the worker's productivity, and competition among employers increases the wage to match the higher productivity. In addition, workers learn from one another by sharing knowledge—in both formal and informal settings—and a worker with more human capital has more knowledge to share and better communication skills. If better-educated workers generate more ideas, an increase in human capital also increases the rate of technological innovation. Glaeser, Scheinkman, and Shleifer (1995) show that cities with relatively high levels of human capital experienced relatively large increases in per-capita income over the period 1960–1990, suggesting a link between human capital and the rate of technological progress.

In recent decades, the share of metropolitan residents with college degrees has increased significantly. Between 1980 and 2000, the overall share for U.S. metropolitan areas increased from 0.17 to 0.23. There is substantial variation in the college share across cities, with a range of 0.11 to 0.44 in 2000. The cities with above-average shares experienced more rapid growth in the college share since 1990, so the variation across metropolitan areas has actually increased since then. For example, three cities that started the decade among the top seven cities in terms of college shares experienced the largest increases in the college share during the decade.

There is evidence that the largest beneficiaries of educational spillovers are less-skilled workers. One study estimated that a 1 percent increase in a city's share of college-educated workers increases the wage of high-school dropouts by

1.9 percent, while it increases the wage of high-school graduates by 1.6 percent and the wage of college graduates by 0.4 percent (Moretti, 2004). This reflects the general observation that urban economic growth tends to reduce income inequality (Wheeler, 2004).

A recent study of the biotechnology industry shows that physical proximity to top-notch "star" researchers is an important factor in the birth of biotechnology firms (Zucker, Darby, Brewer, 1998). The new biotechnology firms located close to scientists with specific human capital (those involved in the discovery of genetic sequences). Although many of the scientists were connected to universities and research centers, the key location factor was the human capital of the scientists, not the presence of a university or research center.

In less developed countries, secondary (high-school) education is an important factor in income growth. According to a recent study of Chinese cities (Mody and Wang, 1997), when enrollment in secondary education increases from 30 percent to 35 percent of the eligible population in a city, the growth rate of total output increases by 5 percentage points. This effect diminishes as the enrollment rate increases: An increase in the enrollment rate from 55 percent to 60 percent increases the growth rate by only 3 percentage points. The largest productivity boost from secondary education occurs in cities with relatively high levels of foreign investment, suggesting that foreign investment and human-capital investments are complementary inputs.

THE URBAN LABOR MARKET

We can use a model of the urban labor market to explore the market forces behind the equilibrium wages and total employment in a city. We assume that the metropolitan area is part of a larger regional economy and that households and firms move freely between cities in the region. The demand for labor comes from firms in the city, while supply comes from households living in the city. The model shows the effects of changes in demand and supply on the city's equilibrium wages and total employment.

Urban Labor Demand

As explained in the Appendix "Tools of Microeconomics," a labor-demand curve is also a marginal-benefit curve: it shows the marginal benefit of hiring an additional unit of labor. The marginal benefit of labor is the revenue generated by one additional unit of labor:

Marginal revenue product = Marginal product · price of output

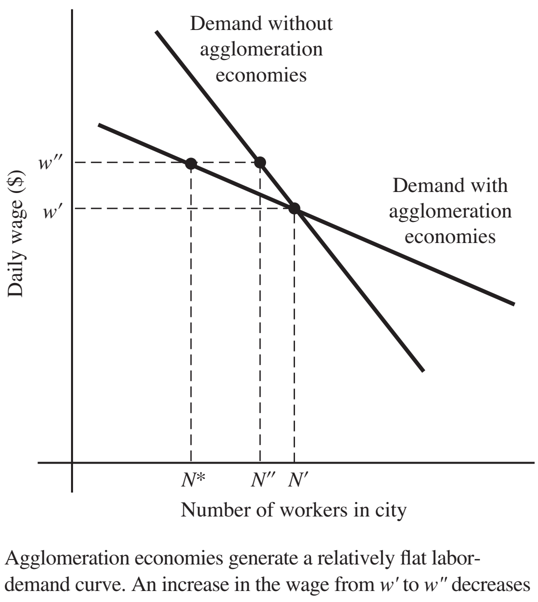
The marginal product (also called marginal physical product) is defined as the increase in output from one additional unit of labor. The labor-demand curve is negatively sloped because as the quantity of labor increases, the marginal product of labor decreases, decreasing the marginal revenue product (the marginal benefit). Firms naturally hire the most productive workers first, and as the quantity of labor increases, they are forced to hire progressively less productive workers. As the less productive workers are added to the city workforce, the marginal product of labor decreases, decreasing the marginal revenue product of labor.

As we saw in Chapter 3, agglomeration economies increase labor productivity. Therefore, as total employment in a city increases, the marginal product of labor increases, pushing up the marginal benefit of labor (the marginal revenue product). If we start with a demand curve that ignores agglomeration economies and then incorporate the productivity effects of agglomeration economies, the demand curve becomes flatter: agglomeration economies and the resulting productivity boost moderate the normal decline in productivity that occurs as the workforce expands. In other words, agglomeration economies generate a flatter labor-demand curve, with a more elastic demand for labor.

Figure 5–2 shows the implications of agglomeration economies for urban labor demand. The relatively steep demand curve is a conventional labor-demand curve that does not incorporate agglomeration economies. In contrast, the relatively flat demand curve incorporates agglomeration economies: as total employment increases, agglomeration economies boost productivity and moderate the normal decrease in marginal productivity that occurs as firms hire progressively less productive workers.

To illustrate the importance of agglomeration economies for the urban labor market, consider the effects of an increase in the market wage. In Figure 5–2,

FIGURE 5–2 Agglomeration Economies and Urban-Labor Demand



Agglomeration economies generate a relatively flat labor-demand curve. An increase in the wage from w' to w'' decreases the quantity demanded from N' to N'' in the absence of agglomeration economies. If the city is subject to agglomeration economies, the quantity of labor demanded decreases to N^*.

suppose the wage increases from w' to w''. In the absence of agglomeration economies, the citywide quantity of labor demanded will decrease from N' to N'' as firms shed workers whose marginal revenue product (marginal product times output price) now falls short of the higher wage. If the economy is subject to agglomeration economies, the quantity of labor demanded drops further, to N^*. As total employment decreases and the economy shrinks, agglomeration economies are lost (a result of reduced input sharing, labor pooling, labor matching, and knowledge spillovers), reducing labor productivity and the marginal revenue product of workers. Firms shed additional workers whose marginal revenue product drops below the new wage. To summarize, agglomeration economies amplify any reduction in the city's workforce because the loss of agglomeration economies makes workers less productive.

The same logic applies to a decrease in the wage. When the wage decreases, there will be more workers whose marginal productivity now exceeds the lower wage. As employment increases and the economy grows, agglomeration economies are gained (a result of increased input sharing, labor pooling, labor matching, and knowledge spillovers), increasing labor productivity and the marginal revenue product of workers. Firms add workers whose marginal revenue product rises above the new wage. Agglomeration economies amplify any increase in the city's workforce because the gain of agglomeration economies makes workers more productive.

We can also explain the slope of the labor-demand curve in terms of the effects of a change in the wage on the quantity of labor demanded. For a conventional demand curve, an increase in the wage has two effects.

1. **The substitution effect.** An increase in the city's wage causes firms to substitute other inputs (capital, land, materials) for the relatively expensive labor.
2. **The output effect.** An increase in the city's wage increases production costs, increasing the prices charged by the city's firms. Consumers respond by purchasing less output, so firms produce less and hire fewer workers.
   For the urban labor market, the presence of agglomeration economies adds a third effect of an increase in the wage.
3. **Agglomeration effect.** An increase in the wage and the resulting decrease in the quantity of labor demanded reduces agglomeration economies and decreases labor productivity, causing an additional decrease in the quantity of labor demanded.

The demand curve is negatively sloped because an increase in wages generates a substitution effect, an output effect, and an agglomeration effect.

Shifting the Urban Labor Demand Curve

What causes the demand curve to shift to the right or the left? The following factors determine the position of the curve:

1. **Demand for exports.** An increase in the demand for the city's exports increases export production and shifts the demand curve to the right: At every wage, more workers will be demanded.

2. **Labor productivity.** An increase in labor productivity decreases production costs, allowing firms to cut prices, increase output, and hire more workers. As we saw earlier in the chapter, labor productivity increases with capital deepening, technological progress, increases in human capital, and agglomeration economies.
3. **Business taxes.** An increase in business taxes (without a corresponding change in public services) increases production costs, which in turn increases prices and decreases the quantity produced and sold, ultimately decreasing the demand for labor.
4. **Industrial public services.** An increase in the quality of industrial public services (without a corresponding increase in taxes) decreases production costs and thus increases output and labor demand.
5. **Land-use policies.** Industrial firms require production sites that (*a*) are accessible to the intracity and intercity transportation networks and (*b*) have a full set of public services (water, sewerage, electricity). By coordinating its land-use and infrastructure policies to ensure an adequate supply of industrial land, a city can accommodate existing firms that want to expand their operations and new firms that want to locate in the city.

Export versus Local Employment and the Multiplier

We can divide production in the urban economy into two types, export and local. Export goods are sold to people outside the city. For example, steel producers sell most of their output to customers outside the city where steel is produced. In contrast, local goods are sold to people within the city. Most of the output of bakeries, bookstores, and pet salons is sold within the city. Total employment is the sum of export employment and local employment.

The two types of employment are related to one another through the multiplier process. Suppose a steel producer expands its operation by hiring 100 additional workers to produce goods for exports. These workers earn an income, and they spend part of it on local goods such as groceries, haircuts, and books. The firms producing these local goods hire more workers to produce additional output, so the increase in export employment leads to increases in local employment. These new local workers in turn spend part of their income on local goods, supporting additional local jobs. The spending and respending of income in the local economy supports local jobs, so the increase in total employment exceeds the initial increase in export employment.

How many additional local jobs are generated by the increase in export employment? To answer this question, policy makers examine the interactions between firms in an urban economy and estimate the employment multiplier, defined as the change in total employment per unit change in export employment. If the multiplier is 2.10, for example, a one-unit increase in export employment increases total employment directly by one export job, and indirectly by 1.10 local jobs, for a total effect of 2.10 jobs.

TABLE 5–1 Metropolitan Employment Multipliers

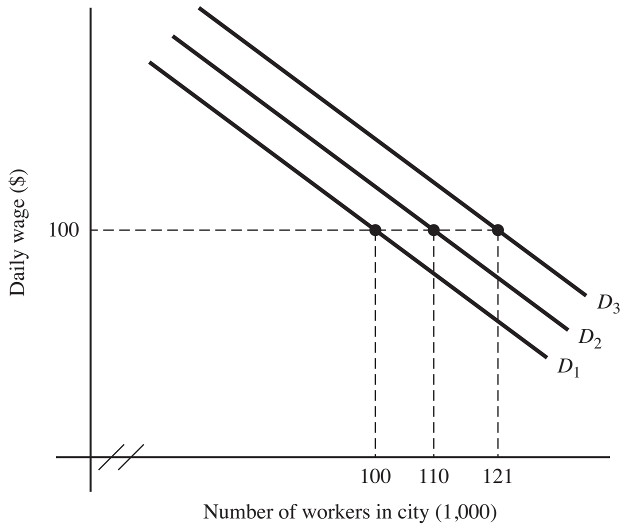
Industry	Portland Metropolitan Multiplier
Frozen food manufacturing	2.40
Wineries	2.74
Textile and fabric finishing mills	1.82
Carpet and rug mills	1.88
Footwear manufacturing	1.92
Envelope manufacturing	2.13
Photographic film and chemical manufacturing	2.53
Optical instrument and lens manufacturing	1.46
Fiber-optic cable manufacturing	2.71
Heavy-duty truck manufacturing	2.55
Motorcycle, bicycle, and parts manufacturing	1.92
Software publishers	2.17
Insurance carriers	2.49
Legal services	1.76
Architectural and engineering services	1.74
Custom computer programming services	1.58
Computer systems design services	2.21
Other computer-related services	1.60
Management consulting services	1.66
Environmental and other technical consulting	1.78
Scientific research and development services	1.51
Advertising and related services	1.67
Hospitals	2.13
Spectator sports	1.54
Independent artists, writers, and performers	2.77
Museums, historical sites, zoos, and parks	2.19

Source: ECONorthwest.

Table 5–1 shows the estimated multipliers for select industries for the Portland Metropolitan Area. The multipliers vary across industries, from a low of 1.46 for optical instruments to a high of 2.77 for independent artists. For the service sectors (legal, architecture, programming, computer, consulting, scientific, advertising), the multipliers are between 1.51 and 2.21. The average multiplier for the 423 industries for which multipliers are estimated is 2.13, indicating that on average, a one-unit increase in export employment increases total employment in the metropolitan area by 2.13 jobs.

Figure 5–3 shows the effects of an increase in export sales on a city's labor-demand curve. Suppose that an increase in exports increases the demand for export workers by 10,000. The city's demand curve will shift to the right from D_1 to D_2, and an additional 10,000 export workers would be demanded at a wage of \$100 per day. If the employment multiplier is 2.10, every export job supports 1.10 local jobs, so the demand curve shifts to the right by an additional 11,000 workers (from D_2 to D_3). Total labor demand increases by 21,000 (2.1 times the increase in the demand for export labor).

FIGURE 5–3 Direct and Multiplier Effects of an Increase in Export Employment



If export employment increases by 10,000, the labor-demand curve shifts to the right (D_1 to D_2) because of the direct effect (10,000 workers) and shifts further to the right (D_2 to D_3) because of the multiplier effect (11,000 additional local workers).

The Labor-Supply Curve

Consider next the supply side of the urban labor market. The supply curve is positively sloped, indicating that the higher the wage, the larger the number of workers in the city. We make two simplifying assumptions for the supply curve:

- **A fixed number of work hours per worker.** The empirical evidence on labor supply suggests that an increase in the wage has a negligible effect on the aggregate hours worked; some people work more and others work less, but on average, people work about the same number of hours.
- **A fixed labor-force participation rate.** We assume that a change in the wage does not change the fraction of the city's population in the workforce.

Given these two assumptions, an increase in the wage increases the supplied labor because more workers move to the city.

Why is the supply curve positively sloped? An increase in total employment in the city increases the total demand for housing and land, pulling up their prices. Recall the first axiom of urban economics:

Prices adjust to generate locational equilibrium

To ensure locational equilibrium in the labor market, a growing city must offer a higher wage to compensate workers for the higher cost of living. The elasticity of the cost of urban living with respect to the size of the labor force is 0.20 (Bartik 1991):

$$e(C, N) = \frac{\text{\% change in Cost of living}}{\text{\% change in Labor force}} = 0.20$$

For example, a 10 percent increase in the labor force increases the cost of living by about 2 percent. This means that to keep real wages constant, the elasticity of the wage with respect to the labor force must be 0.20:

$$e(W, N) = \frac{\text{\% change in Wage}}{\text{\% change in Labor force}} = 0.20$$

We can use these numbers to compute the elasticity of supply of labor, defined as the percentage change in the quantity of labor supplied divided by the percentage change in the wage. This is of course just the inverse of the elasticity of the wage with respect to the labor force.

$$e(N, W) = \frac{\text{\% change in Labor force}}{\text{\% change in Wage}} = 5.0$$

With an elasticity of labor supply of 5.0, a 2 percent increase in the wage increases the labor force by five times 2 percent, or 10 percent. This elasticity applies to an individual city and is much larger than the national labor-supply elasticity (close to zero) because there is less migration between nations than between cities.

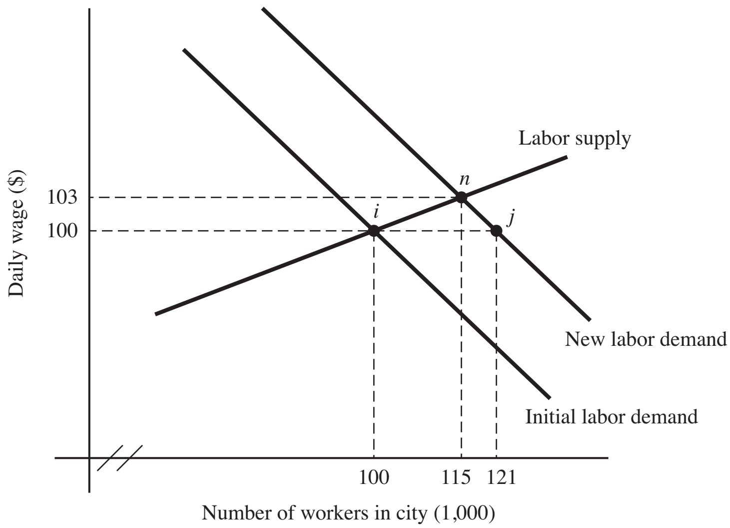
What causes the supply curve to shift to the right or left? The position of the supply curve is determined by the following factors:

1. **Amenities.** Anything that increases the relative attractiveness of the city (other than the wage) shifts the supply curve to the right. For example, an improvement in air or water quality causes migration that increases the supply of labor. Similarly, an increase in the variety of consumer goods (restaurants, entertainment) will increase the supply of labor.
2. **Disamenities.** Anything that decreases the relative attractiveness of a city decreases labor supply and shifts the supply curve to the left. For example, an increase in the crime rate causes people to flee the city, decreasing labor supply.
3. **Residential taxes.** An increase in residential taxes (without a corresponding change in public services) decreases the relative attractiveness of the city, causing out-migration that shifts the supply curve to the left.
4. **Residential public services.** An increase in the quality of residential public services (without a corresponding increase in taxes) increases the relative attractiveness of the city, causing in-migration that shifts the supply curve to the right.

Equilibrium Effects of Changes in Supply and Demand

Figure 5–4 shows the effects of an increase in export sales on the urban labor market. The labor-demand curve shifts to the right by 21,000 workers, reflecting the

FIGURE 5–4 Equilibrium Effects of an Increase in Export Employment



An increase in export employment shifts the demand curve to the right, reflecting both the direct and multiplier effects. The equilibrium moves from point *i* to point *n*, with an increase in the wage and total employment.

effect of an increase in 10,000 export jobs. As the population of the city increases, the prices of housing and land increase, requiring an increase in the wage to compensate workers for the higher cost of living. In other words, the city moves upward along its supply curve. The equilibrium wage rises from $100 per day to $103, and the equilibrium number of laborers increases from 100,000 to 115,000.

Figure 5–4 suggests that predicting the effects of an increase in export employment is tricky. The simple approach is to use the employment multiplier to predict the change in total employment from a projected change in export employment. In the numerical example, the predicted change in total employment from this method would be 21,000 (2.1 times 10,000). This approach tells us the horizontal shift of the demand curve, not the change in equilibrium employment. To accurately predict the change in total employment, one must also know the slopes of the supply and demand curves.

We can use two simple formulas to predict the effect of an increase in demand on a city's equilibrium wage and employment. The formula for the change in the equilibrium wage is

$$\text{Percentage change in equilibrium wage} = \frac{\text{Percentage change in Demand}}{E_d + E_s}$$

where the percentage change in demand is the percentage horizontal shift of the demand curve, E_s is the elasticity of supply, and E_d is the absolute value of the elasticity of demand. In the example depicted in Figure 5–4, the demand curve shifts horizontally by 21 percent (equal to 21,000/100,000). Suppose the absolute value of the demand elasticity is 2.0, and the supply elasticity is 5.0. Then the predicted wage change is

$$\text{Percentage change in equilibrium wage} = \frac{21\%}{2 + 5} = 3\%$$

The market moves upward along the supply curve, so we can use the supply elasticity to predict the change in quantity:

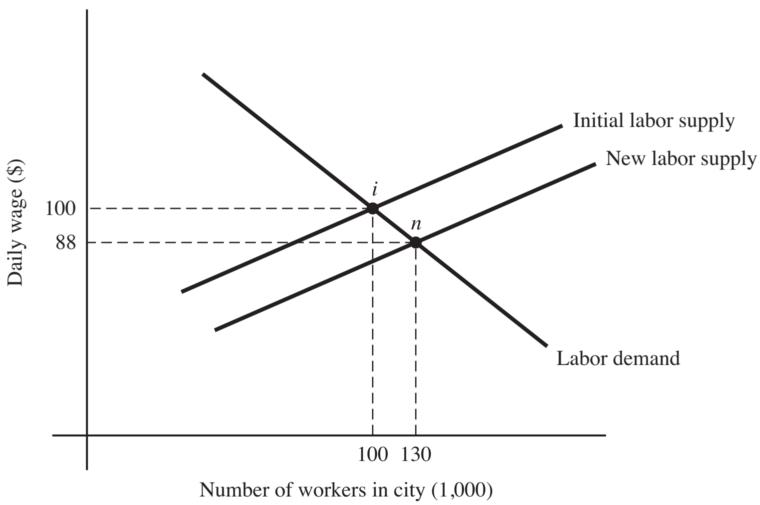
$$\%\text{ change in quantity of labor} = E_s \cdot \%\text{ change in Wage}$$

$$\%\text{ change in quantity of labor} = 5 \cdot 3\% = 15\%$$

In this case, a 21 percent increase in demand leads to a 3 percent increase in the wage and a 15 percent increase in total employment.

Figure 5–5 shows the effects of an increase in the supply of labor. Suppose the city improves its residential public services. For example, the city could improve its public-safety programs or alter its transportation system to decrease commuting costs. In Figure 5–5, the labor-supply curve shifts to the right: At each wage, more people are willing to work and live in the city. The shift of the supply curve increases equilibrium employment and decreases the equilibrium wage. Figure 5–5 is consistent with the empirical evidence provided by Eberts and Stone (1992) concerning the effects of improvements in local infrastructure on wages and total employment. Workers accept lower wages in cities that provide a superior mix of local public goods.

FIGURE 5–5 Equilibrium Effects of an Improvement in Public Services



An improvement in residential public services increases labor supply and shifts the supply curve to the right. The equilibrium moves from point *i* to point *n*, with a decrease in the wage and an increase in total employment.

EMPLOYMENT GROWTH AND DECLINE OF THE U.S. MANUFACTURING BELT

The model of the urban labor market provides some insights into the rise and then decline of the manufacturing belt in the Northeast and Great Lakes regions of the United States. The manufacturing belt developed in the second half of the 19th century. Innovations in production allowed firms to exploit scale economies, and many of the production processes required large volumes of relatively immobile resources (e.g., coal and iron ore). The manufacturing belt had a natural advantage in its access to these resources, so manufacturing was concentrated there. As late as 1947, the manufacturing belt contained 70 percent of the nation's manufacturing employment.

In the second half of the 20th century, manufacturing activity in the United States became more widely dispersed. In 1987, seven of the nation's nine regions had manufacturing employment shares within 2.4 percentage points of the national share of 17.6 percent. By the year 2000, the traditional manufacturing belt contained only about 40 percent of the nation's manufacturing employment, just above its share of total employment. An important factor in the dispersion of manufacturing was a general reduction in transport costs that reduced the natural advantage of the old manufacturing belt.

The dispersion of manufacturing decreased the demand for manufacturing workers throughout the old manufacturing belt. In some cities, the equilibrium total employment decreased, causing population losses. Among the cities that lost population over the period 1970–2000 were Detroit (7% loss), Cleveland (28% loss), and Pittsburgh (5% loss). In contrast, many cities grew despite the loss of manufacturing employment. Among the cities that gained population over the 30-year period were Boston (11% gain) and Minneapolis (50% gain).

The Role of Human Capital

What explains the different experiences of these cities? As documented by Glaeser (2009), the key factor is the stock of human capital. In declining cities such as Detroit, Cleveland, and Pittsburgh, the share of the workforce with college degrees was relatively low. In contrast, in growing cities such as Boston and Minneapolis, the college share was relatively high. In the last four decades, the demand for low-skilled manual labor (in manufacturing and other industries) decreased, while the demand for high-skilled thinking labor (in services such as finance, legal services, medical care) increased. The cities with relatively educated workforces were better equipped to make the transition to an economy with a greater share of high-skilled thinking jobs. In contrast, the cities with poorly educated workforces were ill equipped to cope with changing economic circumstances, so their economies suffered.

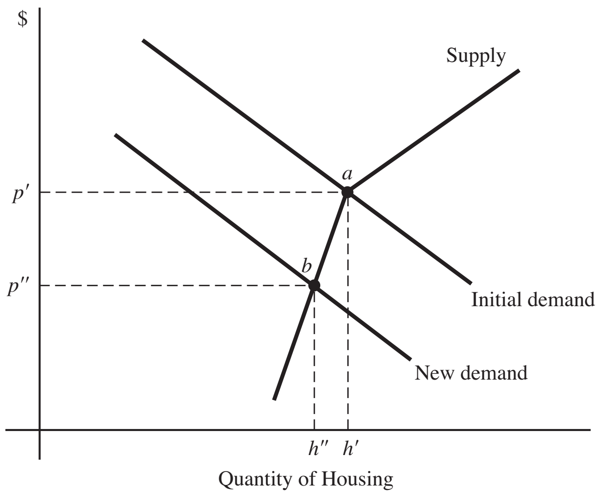
A number of studies have quantified the connection between human capital and urban growth. In the Glaeser study for the period 1980–2000, the estimated elasticity of population growth with respect to the share of the adult population with a college degree is 1.2: a 10 percent increase in the college share increases the population growth rate by 12 percent. Glaeser concludes that human capital is a most powerful variable in explaining the differences in employment and population growth across cities in the old manufacturing belt. For a broad sample of cities in the period 1990–2000, Wolman and colleagues (2008) show that a one-point increase in the share of the population with some college is associated with a 0.60 percentage-point increase in citywide employment.

Labor and Housing Markets in Shrinking Cities

Earlier in the chapter we saw that the urban labor-supply curve is positively sloped because larger cities have higher housing prices. To achieve locational equilibrium in the regional labor market for a given occupation, the real wage (the market wage divided by the cost of living) must be equal across cities. If a larger city has higher housing prices, it must have a higher market wage to offset a higher living cost.

Recent experiences in shrinking cities provide some important insights about the interactions between urban labor and housing markets. As we'll see later in the book, housing is unusual because it is durable: a well-maintained house can last for many decades. As a result, the supply of used housing is relatively inelastic in the downward direction. As shown in Figure 5–6, the supply curve is kinked at the initial market equilibrium quantity. An increase in price generates the conventional increase in quantity supplied, but a decrease in price generates a relatively small

FIGURE 5–6 Durable Housing and Kinked Supply Curve



Housing is durable, so the supply curve is kinked at the initial equilibrium (point *a*). A decrease in demand generates a relatively large decrease in price (from *p′* to *p″*) and a relatively small decrease in quantity (from *h′* to *h″*).

decrease in quantity supplied because housing is durable. As a result, a decrease in demand—resulting from a decrease in total employment and population—generates a relatively large decrease in price and a relatively small decrease in quantity.

The kinked supply curve has important implications for the regional labor market. As a city shrinks, locational equilibrium in the regional labor market (equal real wages across cities) is restored with a relatively large decrease in the price of housing and a relatively small decrease in the market wage. In other words, workers can be indifferent among cities with roughly the same wages but large differences in the price of housing. Comparing the typical declining city to the typical growing city (Gyorko, 2009), the declining city has a wage that is 2 percent lower ($14.49 versus $14.75) and a housing price that is 37 percent lower ($71,560 versus $112,540).

PUBLIC POLICY AND EQUILIBRIUM EMPLOYMENT

Public policy affects equilibrium employment in a city by shifting the labor-supply curve or the labor-demand curve. As explained earlier in the chapter, local government can shift the curves through its decisions concerning local education, public services, business infrastructure, and taxes. In this part of the chapter, we take a closer look at the effects of various public policies on the location decisions of firms—the demand side of the labor market. The question is whether a particular policy attracts firms or repels them.

Taxes and Firm Location Choices

In the last few decades, there have been dozens of studies examining the effect of local taxes on firm location choices and urban employment growth. Bartik (1991) draws some general conclusions from these studies. There is evidence that local taxes have a strong negative effect on employment growth: A high-tax city will grow at a slower rate than a low-tax city, *everything else being equal*. Of course, one of the items included in *everything else* is public services. The evidence suggests that if two cities have the same level of public services but different tax liabilities, the high-tax city will grow at a slower rate.

We can distinguish between two types of business location decisions, intercity decisions (choosing a city or metropolitan area) and intracity decisions (choosing a site within a city or metropolitan area). The elasticity of business activity with respect to tax liabilities is defined as the percentage change in business activity divided by the percentage change in tax liabilities.

- **Intercity location decisions.** The elasticity is between −0.10 and −0.60: A 10 percent increase in taxes in a particular metropolitan area decreases business activity in the metropolitan area by 1 percent to 6 percent.
- **Intracity location decisions.** The elasticity is between −1.0 and −3.0: If an individual municipality increases its taxes by 10 percent, business activity in the municipality decreases by 10 percent to 30 percent.

The elasticity for the intracity decision is larger because firms are more mobile within metropolitan areas than between them. The locations within a metropolitan area are better substitutes than locations in different metropolitan areas.

Two other results from recent empirical studies are worth noting. First, manufacturers are more sensitive than other firms to tax differences. This is sensible because manufacturers are oriented toward the national market and thus have a wider range of location options. Second, metropolitan areas with relatively high taxes on capital (in the form of taxes on business property) tend to repel capital-intensive industries and attract labor-intensive industries.

Public Services and Location Decisions

There is evidence that local public services have a strong positive effect on regional business growth. If two cities differ only in the quality of their local public services, the city with better public services will grow at a faster rate. Similarly, if a city improves its public services, it will grow faster, everything else (including taxes) being equal. The public services that have the largest positive effect on business growth are education and infrastructure.

How would simultaneous increases in taxes and spending on public services affect location choices and business activity? Studies by Helms (1985) and Munnell (1990) suggest that the effect of a tax increase depends on how the extra tax revenue is spent. If the extra revenue is spent on local public services (infrastructure, education, or public safety), the tax/expenditure program increases the relative attractiveness of the city and promotes employment growth. In contrast, if the extra tax revenue is spent on redistributional programs for the poor, the tax decreases the relative attractiveness of the jurisdiction and decreases the growth rate.

Subsidies and Incentive Programs

Many cities try to attract new firms by offering special subsidies. One approach is to lure firms with special tax abatements—for example, an exemption from paying property taxes for 10 years. Some cities loan money directly to developers, and others guarantee loans from private lenders. Some cities subsidize the provision of land and public services for new development. The city purchases a site, clears the land, builds roads and sewers, and then sells the site to a developer at a fraction of the cost of acquiring and developing the site.

Studies of economic-development programs suggest that they have relatively small effects. A study of economic-development policies of municipalities in the Detroit area shows that the programs have a positive effect on business activity in only 5 of 16 cases (Wassmer, 1994). For the other 11 cases, the programs have either no effect on business activity or a negative effect. Studies of enterprise zones (areas of a city where firms pay low tax rates, receive subsidies for worker training, and are exempted from local regulations) suggest that such zones are not very effective in luring firms (Boarnet and Bogart, 1996).

A study of property-tax abatement (Anderson and Wassmer, 1995) discusses how cities sometimes get into bidding to lure firms. The classic example is a firm that secretly decides to locate in a particular city but then asks an inferior city for a tax-abatement package. With the "bid" of the inferior city in hand, the firm then asks the preferred city to match the tax package. The preferred city matches the tax package, and the firm locates in that city, which it had intended to do even in the absence of any tax incentives.

Professional Sports, Stadiums, and Jobs

In 1997, billboards around the San Francisco metropolitan area read, "Build the Stadium. Create the Jobs." That was the slogan for a campaign to get citizens to approve $100 million of public money for a new football stadium for the San Francisco 49ers. Although the campaign failed in San Francisco, many cities have subsidized the construction of facilities for professional sports (Noll and Zimbalist, 1997). Between 1989 and 1997, 31 new stadiums were built, at an average cost of about $150 million. Are sports stadiums effective tools of economic development? Do they create jobs?

The logic behind the job-creation effects of sports stadiums is straightforward. A new stadium can be used to attract a professional sports team or to retain an existing team. Like other organizations, a professional team sells a product and hires workers, including athletes, groundskeepers, ticket takers, accountants, and media personnel. In addition, some of the money the team's employees earn is spent in the local economy, generating multiplier effects that increase employment in restaurants, dental offices, and hardware stores. How many additional jobs does a professional team generate?

Despite the hyperbole of stadium proponents, the job-creation effects of stadiums are modest. The stadium for the Arizona Diamondbacks cost $240 million but increased total employment in the area by only 340 jobs. This figure includes both the direct effect (people hired by the team) and the multiplier effect (local jobs). In other words, the cost per job was $705,882. Employment gains were modest for other host cities, with between 128 and 356 additional jobs in Denver, Kansas City, and San Diego. A comprehensive study of cities that host professional teams showed small positive effects in only one-quarter of cases (Baade and Sanderson, 1997). In about a fifth of the cases, the presence of a sports team actually decreased total employment.

Why are the employment effects of sports teams so small? Most of the money consumers spend on professional sports events comes at the expense of local goods such as movies and restaurant meals. When a sports team comes to town, a large fraction of the money spent on the team is diverted from local consumer products. For example, there may be more popcorn sellers in the stadium, but fewer popcorn sellers in movie theaters. Similarly, a sports event provides a different place to drink beer. To the extent that consumers switch from movies and other local goods to sport events, the employment effects of sports teams will be small.

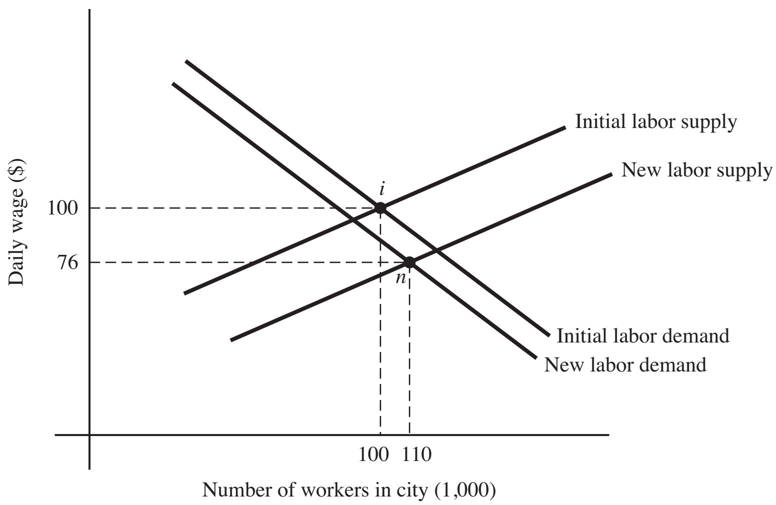
The real power of sports teams to increase employment comes from their ability to attract money from outside the metropolitan area. When someone travels from Providence to Boston to see the Red Sox, the $50 spent on tickets, souvenirs, and food adds money to the Boston economy—and subtracts it from the Providence economy. As a result, total spending and employment in the Boston economy increase. However, because most of the money spent on sporting events comes from local consumers, sports teams don't create many jobs.

Environmental Quality and Employment

Is there a trade-off between environmental quality and total employment? Consider a city with two industries, a polluting steel industry and a clean industry. Suppose the city imposes a pollution tax: Steel producers pay $100 for every ton of pollution they generate. The pollution tax affects both sides of the urban labor market as follows:

1. **Shift of demand curve.** The tax increases the production costs of steel producers, which increases the price of steel. Consumers respond by purchasing less steel, and the decrease in steel output decreases the demand for labor. In Figure 5–7, the labor-demand curve shifts to the left.
2. **Decrease in pollution.** The tax decreases air pollution for two reasons. First, steel producers will reduce pollution to decrease their pollution taxes (installing abatement equipment or changing their inputs or production process), so the volume of pollution generated per ton of steel drops. Second, the increase in the price of steel decreases total steel production.
3. **Shift of supply curve.** The improvement of the city's air quality increases the relative attractiveness of the city. People sensitive to air quality will move to the city, shifting the supply curve to the right.

FIGURE 5–7 The Equilibrium Effects of a Pollution Tax



A pollution tax increases production costs, decreasing the demand for labor. It also improves environmental quality, increasing the supply of labor. The equilibrium moves from point *i* to point *n*. In this example, the supply shift is large relative to the demand shift, so the equilibrium employment increases. The tax decreases the equilibrium wage.

Figure 5–7 shows one possible outcome of the pollution tax. Because supply increases and demand decreases, the program decreases the equilibrium wage from \$100 (point *i*) to \$76 (point *n*). The cleaner city has a lower wage, consistent with the first axiom of urban economics:

Prices adjust to generate locational equilibrium

To make workers indifferent among the cities in the regional economy, the wage must be lower in a city with a better environmental quality. In this example, the rightward shift of the supply curve is large relative to the leftward shift of the demand curve, so equilibrium employment increases from 100,000 to 110,000. The supply shift will be relatively large if households are very responsive to changes in environmental quality, meaning that a large number of households will migrate to the city as environmental quality improves.

How does the pollution tax affect the distribution of employment between the polluting industry and the clean industry? As the wage falls, the production costs of both industries decrease. For the steel industry, the decrease in the wage only partly offsets the pollution taxes, so its production cost increases and its workforce decreases. In contrast, the clean industry simply pays lower wages, so its production costs decrease and its workforce increases. In Figure 5–7, the increase in employment in the clean industry more than offsets the decrease in employment in the steel industry, so total employment increases.

Of course, the pollution tax could actually decrease total employment in the city. If households are not very responsive to improvements in environmental quality, the supply curve will shift by a relatively small amount, generating a relatively small wage reduction. In this case, the increase in employment in the clean industry will not be large enough to offset the decrease in employment in the steel industry, so total employment will decrease. In general, the policy will decrease total employment if the shift of the supply curve is small relative to the shift of the demand curve.

PROJECTING CHANGES IN TOTAL EMPLOYMENT

It is sometimes necessary to project future employment in a city. Cities use employment projections to plan public services such as roads and schools, and firms use employment projections to predict the future demand for their products. The projected change in total employment is computed as follows:

$$\text{Change in total employment} = \text{Change in export employment} \cdot \text{Employment multiplier}$$

As we saw in Table 5–1, multipliers are available for all sorts of industries. Armed with a set of multipliers and a projected change in export employment, policy makers and firms can project a city's future employment. Given the uncertainties associated with predicting future events, the projection of employment is more of an art than a science.

The employment-multiplier approach suffers from a number of problems that limit its applicability. As we saw earlier in the chapter, the approach projects the horizontal shift of the city's labor-demand curve, not the equilibrium change in employment. A second problem is that the approach focuses attention on jobs rather than per-capita income. A third problem is that the approach seems to suggest that a city's economic fate is in the hands of outsiders—the people who buy the exports. If exports determine a city's fate, why has the earth's economy grown without any exports?

WHO BENEFITS FROM INCREASED EMPLOYMENT?

This part of the chapter explores the benefits from an increase in employment, addressing two questions:

- How many of the new jobs are filled by newcomers and how many are filled by original residents of the city who would otherwise not be employed?
- How does an increase in total employment affect real income per capita?

Who Gets the New Jobs?

Bartik (1991) studied the effects of increases in employment on unemployment rates, labor-force participation rates, and migration rates in 89 metropolitan areas. His results suggest that if a city starts with 100,000 jobs, a 1 percent increase in employment (1,000 additional jobs) has the following effects:

- The unemployment rate (the number of unsuccessful job searchers divided by the workforce) decreases from 5.40 percent to 5.33 percent.
- The labor-force participation rate (the workforce divided by the number of adults) increases from 87.50 percent to 87.64 percent.
- The employment rate (jobs divided by the number of adults) increases from 82.78 percent to 82.97 percent.

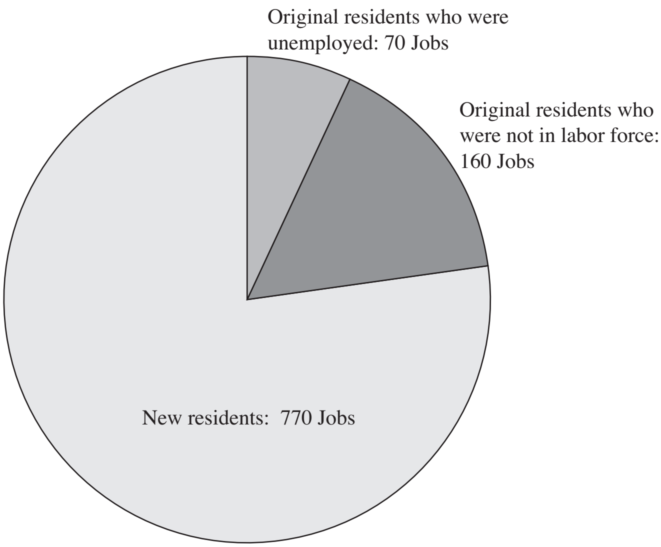
Figure 5–8 shows how the new jobs are divided between old and new residents. Newcomers fill 770 of the 1,000 jobs, leaving 230 jobs for the original residents. The 230 jobs filled by original residents are split between people who were unemployed (70 jobs) and people who were not in the labor force (160 jobs). The simple lesson from Figure 5–8 is that increases in employment cause in-migration and population growth, so a small fraction of the new jobs are filled by original residents.

Effects on Real Income per Capita

How does an increase in total employment affect a city's real income per capita? Income could increase in several ways:

1. **Increase in the real wage.** As explained earlier in the chapter, an increase in total employment causes offsetting changes in nominal wages and living costs, so the real wage for a given occupation will be unaffected.

FIGURE 5–8 Distribution of 1,000 New Jobs between Original Residents and Newcomers



Source: Timothy Bartik. *Who Benefits from State and Local Economic Development Policies?* Kalamazoo, MI: Upjohn Institute, 1991.

2. **Promotions.** Bartik (1991) shows that an increase in total employment hastens movement upward in the job hierarchy. An increase in the demand for labor causes firms to promote workers to higher-paying jobs more rapidly. The largest moves up the hierarchy are experienced by workers who are less educated, young, or black.
3. **Increase in the employment rate.** As explained earlier, an increase in total employment decreases the unemployment rate and increases the participation rate, so it increases the fraction of the working-age population that is employed.

Table 5–2 shows the combined effects of changes in real wages, occupational rank, unemployment rates, and participation rates. For the average household, a 1 percent increase in employment increases real income per capita by 0.40 percent. The most important factors behind higher income are the promotion effect (promotion to higher-paying jobs) and participation effect (higher labor-force participation). The elasticities are larger for households that experience relatively large promotion effects (less educated, young, or black).

TABLE 5–2 Effects of 1 Percent Increase in Total Employment on Real Income per Capita

	Average	Less Educated	Younger	Black
Percent increase in real income	0.40	0.47	0.41	0.49

Source: Bartik, Timothy J. *Who Benefits from State and Local Economic Development Policies?* Kalamazoo, MI: Upjohn Institute, 1991.

SUMMARY

This chapter explores the determinants of increases in urban income and employment. Here are the main points of the chapter:

1. An increase in per-capita income results from capital deepening, increases in human capital, technological progress, and agglomeration economies.
2. An increase in export employment increases local employment through the multiplier process.
3. The urban labor-supply curve is positively sloped because a larger city has higher housing prices, requiring firms to pay higher wages to compensate workers for higher living costs.
4. A large fraction of new jobs in a city are filled by newcomers, leaving few jobs for original residents.
5. An increase in total employment in a city increases real income per capita by (*a*) hastening the move up the job hierarchy and (*b*) increasing the labor-force participation rate.

APPLYING THE CONCEPTS

For exercises that have blanks (________), fill each blank with a single word or number. For exercises with ellipses (. . .), complete the statement with as many words as necessary. For exercises with words in square brackets ([increase, decrease]), circle one of the words.

1. **Innovation and Growth Numbers**
   Suppose a region's workforce of 14 million is initially split equally between two cities, *X* and *Y*. The urban utility curve peaks at 4 million workers, and beyond that point the slope is −$3 per million workers. The initial equilibrium utility level is $60. Suppose city *X* experiences technological innovation that shifts its utility curve upward by $12.
   a. Draw a pair of utility curves, one for *X* and one for *Y*, and label the positions immediately after the innovation (before any migration) as *x* for city *X* and *y* for city *Y*. Use arrows along the curves to indicate the migration that follows.
   b. In the new equilibrium, the utility level is ________ and the population of *X* is ________ million, while the population of *Y* is ________ million.
2. **Education Spillover Benefits**
   Consider a city where the initial wage of high-school dropouts is $10. Suppose the college share of the workforce increases by 2 percent. Use a demand-supply graph of the labor market for high-school dropouts to show the effects on the dropout wage. Use the numbers provided in the section "Human Capital and Economic Growth."
3. **Elasticity of Demand for Software Labor**
   Consider the computer software industry. Assume [i] labor is responsible for 80 percent of production costs, [ii] software is produced with fixed factor proportions (no capital-labor substitution), [iii] there are no agglomeration

economies, [iv] any change in production cost is passed on to consumers in a higher price, and [v] the price elasticity of demand for software is −1.50. Suppose the wage of software workers increases by 20 percent.

a. The price of software will increase by ________ percent, and the quantity of software demanded will ________ by ________ percent.

b. The quantity of software labor demanded will ________ by ________ percent.

c. The elasticity of demand for software labor is ________ , computed as. . . .

d. If assumption [i] is relaxed, the demand for software labor would be [more, less] elastic because. . . .

4. Labor Supply Elasticity: Island City versus Plains City

Island City is located on a small island, while Plains City is located at the center of a large, flat, featureless plain. Draw two labor-supply curves, one for each city.

a. The supply curve for Island City is [steeper/flatter] because. . . .

b. The elasticity of supply of labor in Island City is [higher, lower].

5. Predict Wages and Employment

In the city of Growville, the equilibrium employment is 100,000 workers, and the equilibrium wage is $100 per day. The elasticity of demand for labor is 1.0 (in absolute value) and the elasticity of supply of labor is 5.0. The employment multiplier is 2.0. Suppose the demand for labor used in the production of exports increases by 6,000 jobs.

a. Use a supply-demand graph of the urban labor market to show the effects of the increase in the demand for labor.

b. The equilibrium wage [increases, decreases] by ______ percent (to ________) computed as. . . .

c. The equilibrium employment [increases, decreases] by __________ percent (to __________ workers), computed as. . . .

6. Katrina Decreases Labor Supply

Consider the effects of a natural disaster like hurricane Katrina on a metropolitan economy. In the initial (prehurricane) equilibrium, total employment in the metropolitan area is 500,000 workers and the daily wage is $100. The price elasticity of supply of labor is 4.0 and the price elasticity of demand for labor is −1.0. Suppose the hurricane reduces labor supply (a horizontal shift of the supply curve) by 100,000 workers.

a. Use a supply-demand graph of the urban labor market to show the effects of the hurricane.

b. The equilibrium wage [increases, decreases] by ______ percent (to $ ______) computed as. . . .

c. The equilibrium employment [increases, decreases] by __________ percent (to ________ workers), computed as. . . .

d. The reduction in the equilibrium employment is [greater, less] than the initial decrease in labor supply because. . . .

7. Growth Control and Wages

Consider a city with an equilibrium wage of $80 per day, equilibrium employment of 100,000 jobs, and 100 million square feet of housing. The government's growth-control policy fixes the maximum total square footage in the city at its current level. New housing can be built, but every square foot of new housing requires that one square foot of old housing be retired from the market.

- **a.** Draw a graph with two labor-supply curves, a conventional supply curve and a second that represents labor supply under the city's growth-control policy.
- **b.** The supply curve under the growth-control policy is [steeper, flatter] because. . . .
- **c.** Add two labor-demand curves to your graph, an initial demand curve, and a second curve representing a 20 percent increase in labor demand.
- **d.** The increase in the demand for labor [increases, decreases, does not affect] the equilibrium wage. The wage change is [larger, smaller] under the growth-control policy because. . . .
- **e.** Under the growth-control policy, an increase in the demand for labor [increases, decreases, does not affect] equilibrium employment. The housing consumption per worker [increases, decreases, doesn't change] because housing prices are higher.

8. Effects of Environmental Policy

Consider a city that imposes a new pollution tax that increases the average cost (and price) of the good produced by a polluting industry by 4 percent and improves the city's environmental quality by 20 percent. The price elasticity of demand for the city's export goods is −1.50.

- **a.** The total output of the polluting industry will decrease by ________ percent, computed as. . . .
- **b.** The city's equilibrium wage will [increase, decrease] because. . . .
- **c.** The city's equilibrium employment will increase if the elasticity of the supply of labor with respect to environmental quality is relatively [large, small].
- **d.** Use a supply-demand graph of the urban labor market to illustrate your answer to (*c*).

9. Inefficient Environmental Policy

Consider a city where each polluting firm initially generates two tons of pollution. Half the polluters (type L) could cut back pollution at a cost of $4 per ton, and the other half (type H) could cut back at a cost of $30 per ton. The city is considering two alternative environmental policies:

- **i.** Pollution tax: Each firm would pay a tax of $5 for each unit of pollution.
- **ii.** Uniform-reduction: Each firm would be required to cut its pollution in half, to one ton.
- **a.** The tax policy is [more, less] efficient than the uniform-reduction policy because. . . .
- **b.** The pollution tax causes a [smaller, larger] shift of the city's labor-demand curve because. . . .
- **c.** The city is more likely to experience an increase in total employment under the ________ policy because. . . .
- **d.** Illustrate with two graphs, one for each policy.

10. **Economic Impact of a Football Team**
Consider the results of a consultant's report on the possible economic impacts of moving the Raiders (a professional football team) to Sacramento. The consultant estimated that the team would increase total spending in the Sacramento economy by \$61.6 million per year, computed as follows:

$$\text{Increase in spending} = \text{Spending per fan} \times \text{Attendance} \times \text{Average multiplier}$$

$$\$61.6 \text{ million} = \$40 \times 700{,}000 \times 2.20$$

a. The use of the city's average multiplier is based on two troublesome assumptions. List the assumptions and explain why they are troublesome.
b. According to Ms. Wizard, "If my assumptions are correct, total spending in the Sacramento economy would actually decrease." Assume that her logic is correct. List her two assumptions.

REFERENCES AND ADDITIONAL READING

1. Anderson, John E., and Robert W. Wassmer. "The Decision to 'Bid for Business': Municipal Behavior in Granting Property Tax Abatements." *Regional Science and Urban Economics* 25 (1995), pp. 739–57.
2. Baade, Robert A., and Allen R. Sanderson. "The Employment Effects of Teams and Sports Facilities." *Sports, Jobs and Taxes,* eds. Roger Noll and Andrew Zimbalist. Washington, D.C.: Brookings, 1997.
3. Bartik, Timothy J. *Who Benefits from State and Local Economic Development Policies?* Kalamazoo, MI: Upjohn Institute, 1991.
4. Black, Duncan, and Vernon Henderson. "A Theory of Urban Growth." *Journal of Political Economy* 107 (1999), pp. 252–84.
5. Boarnet, Marlon, and William T. Bogart. "Enterprise Zones and Employment: Evidence from New Jersey." *Journal of Urban Economics* 40 (1996), pp. 198–215.
6. Dowall, David. "An Evaluation of California's Enterprise Zone Programs." *Economic Development Quarterly* 10 (1996), pp. 352–68.
7. Eberts, Randall W., and Joe A. Stone. *Wage and Adjustment in Local Labor Markets*. Kalamazoo, MI: Upjohn Institute, 1992.
8. Glaeser, Edward, Jose Scheinkman, and Andrei Shleifer. "Economic Growth in a Cross-Section of Cities." *Journal of Monetary Economics* 36 (1995), pp. 117–43.
9. Glaeser, Edward. "Growth: The Death and Life of Cities." Chapter 2 in *Making Cities Work*, ed. Robert P. Inman. Princeton, NJ: Princeton University Press, 2009.
10. Gyourko, Joseph. "Housing: Urban Housing Markets." Chapter 5 in *Making Cities Work*, ed. Robert P. Inman. Princeton, NJ: Princeton University Press, 2009.
11. Helms, L. Jay. "The Effect of State and Local Taxes on Economic Growth: A Times Series-Cross Section Approach." *Review of Economics and Statistics* 68 (1985), pp. 574–82.
12. Lucas, Robert. "On the Mechanics of Economic Development." *Journal of Monetary Economics* 22 (1988), pp. 3–22.

13. Lucas, Robert. “Externalities and Cities.” *Review of Economic Dynamics* 4 (2001), pp. 245–74.
14. Mody, Ashoka, and Fang-Yi Wang. “Explaining Industrial Growth in Coastal China: Economic Reforms … and What Else?” *World Bank Economic Review* 11 (1997), pp. 293–325.
15. Moretti, Enrico. “Human Capital Externalities and Cities.” Chapter 51 in *Handbook of Regional and Urban Economics 4: Cities and Geography,* eds. V. Henderson and J. F. Thisse. Amsterdam: Elsevier, 2004.
16. Munnell, Alicia H. “How Does Public Infrastructure Affect Regional Economic Performance?” *New England Economic Review* (1990), pp. 11–33.
17. Noll, Roger G., and Andrew Zimbalist. “Build the Stadium—Create the Jobs!” *Sports, Jobs and Taxes,* eds. Roger Noll and Andrew Zimbalist. Washington D.C.: Brookings, 1997.
18. Papke, Leslie. “Tax Policy and Urban Development: Evidence from the Indiana Enterprise Zone Program.” *Journal of Public Economics* 54 (1994), pp. 37–49.
19. Papke, Leslie. “What Do We Know about Enterprise Zones?” In *Tax Policy and the Economy,* ed. James Poterba. Cambridge, MA: MIT Press, 1993.
20. Rauch, J. E. “Productivity Gains from Geographic Concentration of Human Capital: Evidence from the Cities.” *Journal of Urban Economics* 34 (1993), pp. 380–400.
21. Robak, Jennifer. “Wages, Rents, and the Quality of Life.” *Journal of Political Economy* 90 (1982), pp. 1257–78.
22. Wassmer, Robert W. “Can Local Incentives Alter a Metropolitan City’s Economic Development?” *Urban Studies* 31 (1994), pp. 1251–78.
23. Wheeler, Christopher H. “On the Distributional Aspects of Urban Growth.” *Journal of Urban Economics* 55 (2004), pp. 1371–97.
24. Wolman, Harold, Edward Hill, Pamela Blumenthal, and Kimberly Furdell. “Understanding Economically Distressed Cities,” in *Retooling for Growth: Building a 21st Century Economy in America’s Older Industrial Areas,* eds. Richard McGahey and Jennifer Vey. Washington DC: Brookings, 2008.
25. Zucker, L. G., M. R. Darby, and M. B. Brewer. “Intellectual Human Capital and the Birth of U.S. Biotechnology Enterprises.” *American Economic Review* 88 (1998), pp. 290–306.

Appendix: The Regional Context of Urban Growth

In this Appendix, we broaden our geographical perspective, looking at regions as parts of a national economy. The growth of cities is affected by economic forces at the regional level, and in turn affects those forces. We start with a discussion of the neoclassical model of regional development, then use the model to explain the general trend of regional concentration followed by dispersion.

THE NEOCLASSICAL MODEL

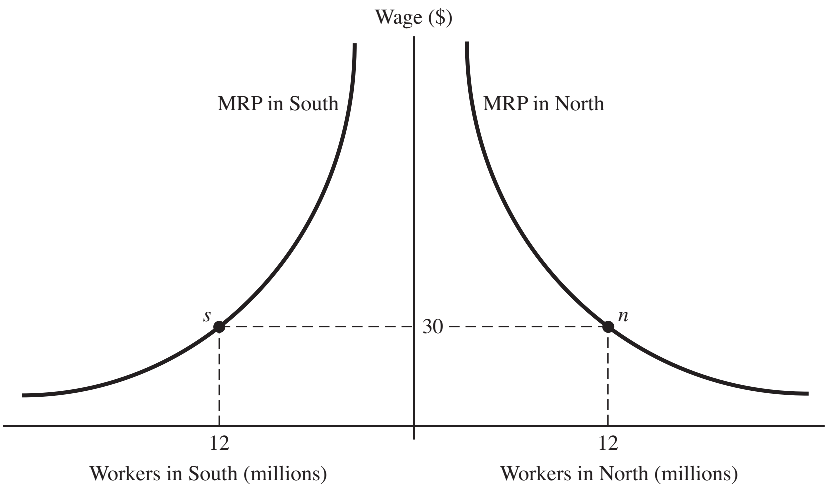
The neoclassical model of regional development focuses on the location decisions of workers, who are assumed to be perfectly mobile between two regions. In the simplest version of the model, the two regions have equal endowments of natural resources. In Figure 5A–1, the horizontal axis measures the number of workers in the regions. The two curves show the marginal revenue product of labor (MRP), which is equal to the marginal product of labor times the price of the export good. The MRP curves are negatively sloped, reflecting the assumption of diminishing marginal returns: As the workforce of a region expands, the marginal product of labor decreases, pulling down the MRP.

In this simple neoclassical model, the two regions are identical. Workers are perfectly mobile, and locational equilibrium requires that the two regions have the same wage. The two regions have the same MRP curve, so the only way to accommodate all the workers with a common wage is an equal division of workers between the two regions. In Figure 5A–1, the initial equilibrium is shown by points *s* (for South) and *n* (for North). At a wage of $30, the nation's population is split equally between the two regions, with 12 million workers in each.

Differences in Natural Advantage Cause Concentration

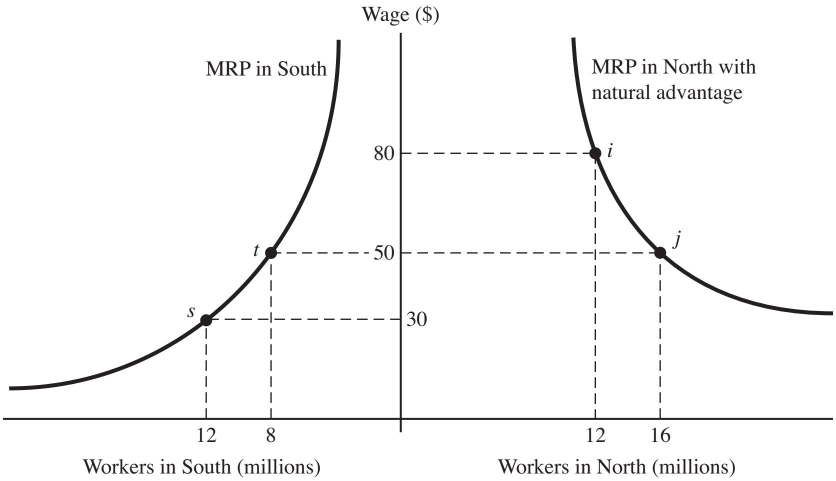
Consider the implications of differences in natural advantage between the two regions. For example, suppose the North has deposits of iron ore and coal, so the cost

FIGURE 5A–1 Neoclassical Model of Regional Development



Under the neoclassical model of regional convergence, diminishing returns generates a negatively sloped marginal revenue product (MRP) curve. With perfect mobility, wages are equal in the two regions ($30). In the absence of a natural advantage, the nation's workforce is split equally between the two regions (12 million in each).

FIGURE 5A–2 Natural Advantage and Size of the Regional Economy



The North has a natural advantage (e.g., access to raw materials), and has a higher MRP curve. Equilibrium occurs at points *t* (South) and *j* (North), with the same wage but a larger workforce in North.

of transporting raw materials is zero for firms in the North. In contrast, the South must import iron ore and coal to produce steel, incurring transport costs in the process.

The natural advantage of the North will result in migration that increases the size of the North economy. As shown in Figure 5A–2, the North's superior access to raw materials makes its workers more productive, generating a higher MRP curve. If we start with 12 million workers in each region, the MRP in the North (shown by point *i*) is $80, compared to only $30 in the South (point *s*). Workers are paid a wage equal to their MRP, and the higher wage in the North will cause workers to migrate there. As they migrate, we move downward along the North MRP curve and upward along the South MRP curve.

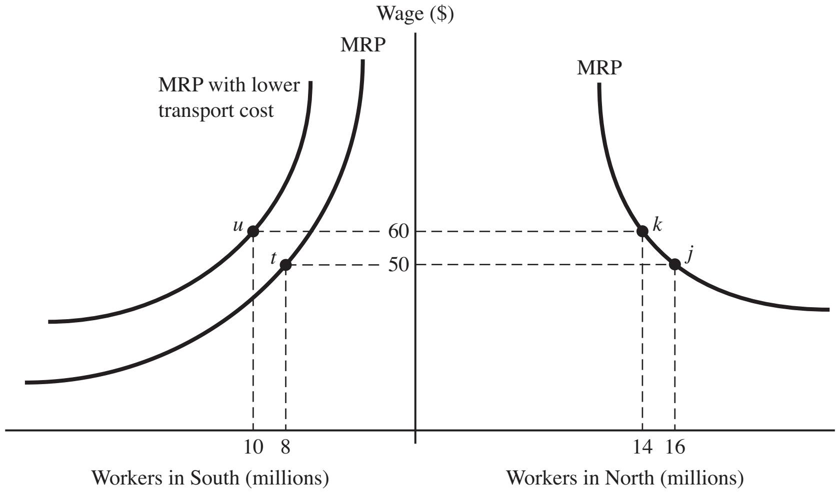
Locational equilibrium will be restored when the two regions have the same wage. In other words, migration will continue until we reach points *t* and *j*, with equal wages ($50) and a larger workforce in the North (16 million), the region with the natural advantage. To summarize, the region with a natural advantage has a larger economy.

A Decrease in Transport Costs Causes Regional Dispersion

How would a decrease in transportation cost affect the distribution of economic activity across regions? In our example, the North has a natural advantage because it bears no transport costs for iron ore and coal, but the South does. A decrease in transport costs will reduce the North's natural advantage and narrow the economic difference between the two regions.

Figure 5A–3 shows the effects of a decrease in transport costs. The productivity of South workers increases, shifting the South MRP curve upward and increasing the wage in the South. Workers will migrate to the South, and migration will continue until the wages are equalized. Equilibrium is restored at points *u* and *k*. The

FIGURE 5A–3 Decrease in Transport Cost Causes Regional Dispersion



A decrease in transport cost increases the productivity of labor of the region that imports raw materials (South). The upward shift of the South's MRP curve generates a new equilibrium with a higher common wage ($60, up from $50), and the South workforce grows (from 8 million to 10 million), while the workforce of the North shrinks (from 16 million to 14 million). Lower transport costs narrows the gap in the economic activity between the two regions.

equilibrium wage rises to $60, and the workforce in the South increases by 2 million workers at the expense of the North.

A decrease in transport costs reduces the differences in the regional economy because it reduces the natural advantage that caused the differences in the first place. As we've seen in earlier chapters, there is a long history of decreasing transport costs, and the neoclassical model predicts that differences across regions will diminish over time.

REGIONAL CONCENTRATION AND DISPERSION IN THE UNITED STATES

The economic history of the United States shows periods of regional concentration followed by dispersion (Kim, 1998). During the colonial period, the national economy was dominated by agriculture, extraction, and fishing. Regional specialization was based on natural comparative advantages generated by differences in soil, climate, and geography. Most nonagricultural products were produced in the home or by artisans in towns and cities.

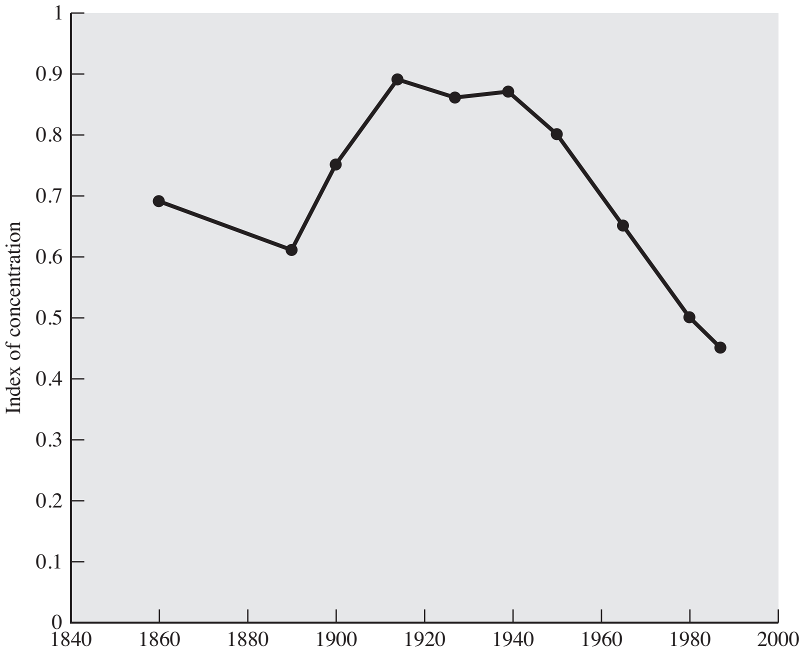
In the first half of the 19th century, production shifted from artisan shops to mechanized and nonmechanized factories. Among the products produced in factories were shoes, wagons, furniture, hats, paper, leather, and textiles. Factories were concentrated in the Northeast region, where, in 1840, about 36 percent of the labor force produced nonagricultural goods (compared to 21 percent for the nation and 9 percent for the South).

During the second half of the 19th century, a manufacturing belt developed in the Northeast and Great Lakes regions. Innovations in production allowed firms to exploit scale economies, and many of the production processes required large volumes of relatively immobile resources (e.g., coal and iron ore). The manufacturing belt had a natural advantage in its access to these resources, so manufacturing was concentrated there. As late as 1947, the manufacturing belt contained 70 percent of the nation's manufacturing employment. In 1954, the manufacturing industry employed about 28 percent of workers nationwide. In three of nine regions, the manufacturing share was well above the national average; in the remaining six, the share was well below the average.

In the second half of the 20th century, economic activity became more widely dispersed. In 1987, seven of the nation's nine regions had manufacturing employment shares within 2.4 percentage points of the national share of 17.6 percent. By the year 2000, the traditional manufacturing belt contained only about 40 percent of the nation's manufacturing employment, just above its share of total employment. An important factor in the dispersion of manufacturing was a general reduction in transport costs that reduced the natural advantage of the old manufacturing belt. In addition, producers switched to alternative raw materials as well as recycled inputs.

Figure 5A–4 shows the time trend in the concentration of manufacturing from 1860 to 1987. The index of concentration measures the differences in the mix of

FIGURE 5A–4 Regional Concentration of Manufacturing



Source: Kim, Sukkoo. "Economic Integration and Convergence: U.S. Regions, 1840–1987." *Journal of Economic History.* (1998), pp. 659–83.

economic activities between two regions. If two regions have the same mix (the same shares of employment in manufacturing, services, agriculture, and trade), the value of the index is zero. The maximum value is 2, indicating completely different economic mixes. Figure 5A–4 shows the average index for the nine regions of the United States. The index increases between 1890 and 1910, levels off for about 30 years, and then declines steadily.

The experience of the United States is consistent with the neoclassical model of regional development. The manufacturing belt developed because of natural advantage (access to material inputs such as coal and iron ore), and declined because the relative cost of transporting inputs decreased. The decrease in transport costs diminished the natural advantage that played a key role in the development of the manufacturing belt, causing manufacturing to disperse to other regions.

REFERENCES AND ADDITIONAL READING

1. Fujita, Masahisa, and Jacques-Francois Thisse. *Economics of Agglomeration*. Cambridge: Cambridge University Press, 2002.
2. Henderson, Vernon, and Jacques-Francois Thisse, eds., *Handbook of Regional and Urban Economics 4: Cities and Geography.* Amsterdam: Elsevier, 2004.
   - Chapter 63: Holmes, Thomas, and John Stevens. "The Spatial Distribution of Economic Activities in North America."
   - Chapter 64: Combes, Pierre-Philippe, and Henry Overman. "The Spatial Distribution of Economic Activities in the European Union."
   - Chapter 65: Fujita, Masahisa, Tomoya Mori, Vernon Henderson, and Yoshitsuga Kanemoto. "The Spatial Distribution of Economic Activities in Japan and China."
   - Chapter 66: Kim, Sukkoo, and Robert Margo. "Historical Perspectives on U.S. Economic Geography."
3. Kim, Sukkoo. "Economic Integration and Convergence: U.S. Regions, 1840–1987." *Journal of Economic History* (1998), pp. 659–83.
4. Kim, Sukkoo. "Regions Resources, and Economic Geography: Sources of U.S. Regional Comparative Advantage, 1880–1987." *Regional Science and Urban Economics* 29 (1999), pp. 1–32.

PART TWO

Land Rent and Land-Use Patterns

This part of the book examines the spatial structure of cities, exploring the market forces and government policies that determine land-use patterns within metropolitan areas. In Chapter 6, we divide the urban economy into three sectors—manufacturing, offices, and households—and see how much each sector is willing to pay for land in different parts of the city. Land usually goes to the highest bidder, so once we know how much each sector is willing to pay for land, we can predict what goes where. An appendix to Chapter 6 uses the economic choice model to explain why land-rent curves are convex rather than linear. In Chapter 7, we examine the actual land-use patterns in modern cities and see how things have changed in the last 100 years. In the heyday of the monocentric city, most jobs were close to the center. In modern cities, jobs are divided between central business districts, sub-centers, and "everywhere else." We'll explore the economic forces behind the spatial transformation of cities and discuss the causes and consequences of urban sprawl. An Appendix to Chapter 7 describes and applies a model of a monocentric city. Chapter 8 explores the economics of neighborhood choice, focusing on how decisions of where to live are affected by local public goods, schools, and crime. The chapter shows why we observe so much sorting of households with respect to income, educational level, and race. Chapter 9 discusses the role of local governments in the urban land market, exploring the market effects of zoning and growth controls.

CHAPTER 6

Urban Land Rent

The trouble with land is that they're not making it anymore.
—Will Rogers

Take a walk from the outskirts of a metropolitan area to the center, and you'll observe some curious changes along the way. Early in your trip, the price of land will increase slowly and sometimes decrease, but eventually the price will start to increase exponentially. As you approach the center, building heights will increase exponentially too, so buildings near the center will tower over buildings just a few blocks away. In this chapter, we explain why the price of land varies within cities and show the connection between expensive land and tall buildings.

This is the first of four chapters on the spatial structure of cities. In this chapter, we divide the urban economy into three sectors—manufacturing, offices, and households—and see how much each sector is willing to pay for land in different parts of the city. Land usually goes to the highest bidder, so once we know how much each sector is willing to pay for land, we can predict what goes where. In the next chapter, we'll look at the actual land-use patterns in modern cities and see how things have changed in the last 100 years. The third chapter on spatial structure explores the economics of neighborhoods, and the fourth explores the role of government in urban land use.

INTRODUCTION TO LAND RENT

It will be useful to define two terms, *land rent* and *market value*. Land rent is the periodic payment by a land user to a landowner. For example, a firm may pay $9,000 per month to use an empty lot as a parking lot. In contrast, the market value of land is the amount paid to become the land owner. In this book, the "price" of land is land rent, a periodic payment to a landowner. This is sensible because many other economic variables are expressed as periodic payments, including household income, firm profits, and interest payments.

The rent on a particular plot of land is determined by how much money can be earned by using the land. David Ricardo (1821) is credited with the idea that the price of agricultural land is determined by its fertility. Consider an agricultural

TABLE 6–1 Fertility and Land Rent

	Price of Corn	Quantity Produced	Total Revenue	Nonland Cost	WTP for Land	Bid Rent for Land
Low fertility	$10	2	$20	$15	$ 5	$ 5
High fertility	$10	4	$40	$15	$25	$25

county where farmers grow corn on two types of land, highly fertile and less fertile. The price of corn, determined in national markets, is $10. Farmers rent land from landowners, and there are no restrictions on entry into the corn market.

Table 6–1 shows how to compute the maximum amount a tenant farmer is willing to pay for land. The less fertile land produces two units of corn per hectare, so total revenue per hectare is $20. The cost of nonland inputs (capital, labor, fertilizer) is $15, so the farmer's profit before paying for land is $5. This is the maximum amount a farmer is willing to pay (WTP) for a hectare of low-fertility land. In contrast, highly fertile land produces twice as much output per hectare for the same cost. The more fertile land generates $20 of extra revenue, so the farmer is willing to pay $20 more to use the more fertile land.

How much will farmers offer to pay (bid) for land? Recall the fifth axiom of urban economics:

Competition drives economic profit to zero

There are no restrictions on entry into corn farming, and we assume that all farmers have access to the same production technology and the same inputs. Therefore, competition among prospective farmers will bid up the price of land until economic profit is zero. Farmers are willing to pay up to $25 per hectare for the high-fertility land, and that's how much they bid for it. If a farmer were to bid less than $25, the landowner could find another farmer willing to pay $25 to use the land. This is the *leftover principle:* Because of competition among farmers for land, the landowner gets the leftovers, equal to total revenue minus total nonland costs. Less fertile land has a lower rent ($5) because there is less money left over after paying the nonland production cost.

BID-RENT CURVES FOR THE MANUFACTURING SECTOR

In an urban environment, the willingness to pay for land depends on its accessibility rather than its fertility. Suppose manufacturing firms in a city assemble bicycles, using land, labor, and imported parts (such as wheels and frames), and then export their output to consumers outside the city. Imported parts and finished bikes are transported by truck on a highway that runs through the city. Let's assume that the price of bikes is determined in the world market and is unaffected by changes in the city.

We can use the leftover principle to determine how much bike producers will bid for land at different locations in the city. We are interested in the bid rent per

TABLE 6–2 Computing the Manufacturing Bid Rent

Distance	Total Revenue	Nonland Production Cost	Freight Cost	WTP for Land	Production Site (hectares)	Bid Rent (per hectare)
0	$250	$130	—	$120	2	$60
1	$250	$130	$20	$100	2	$50
2	$250	$130	$40	$ 80	2	$40
3	$250	$130	$60	$ 60	2	$30

hectare of land, which equals the firm's willingness to pay for a lot large enough for its factory (revenue minus cost) divided by the size of the factory lot (in hectares).

$$\text{Rent per hectare} = \frac{\text{Total revenue} - \text{nonland production cost} - \text{freight cost}}{\text{Lot size (quantity of land)}}$$

Suppose each firm produces five units of output and its output price is $50, so its total revenue is $250. The nonland production cost is $130. As shown in the first row of numbers in Table 6–2, a firm at the highway has no freight cost, so its willingness to pay for a factory lot (the numerator of the rent expression) is $120. If a firm occupies two hectares of land, its bid rent per hectare is $60.

The Negatively Sloped Rent Curve

The firm's bid rent for land decreases as the distance to the highway increases. Suppose the unit freight cost, defined as the cost of transporting one unit of output one mile, is $4. The firm produces five units of output per day, so its daily freight cost is $20 at a site one mile from the highway, $40 two miles from the highway, and so on. As shown in the last column of Table 6–2, the firm bids $50 per hectare for a site one mile from the highway, $40 for a site two miles away, and so on.

In Figure 6–1 (page 130), the manufacturing bid-rent curve is negatively sloped, reflecting rising freight cost as the firm moves away from the highway. The slope is the change in the bid rent from a one-unit increase in distance:

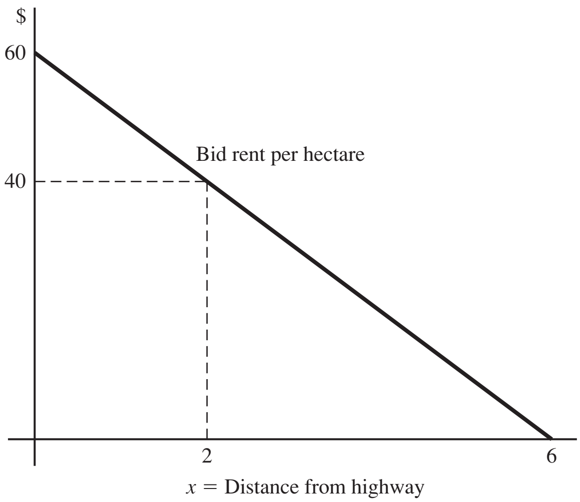
$$\frac{\Delta R}{\Delta x} = \frac{\text{Unit freight cost} \cdot \text{output}}{\text{Lot size}} = \frac{-\$4.5}{2} = -\$10$$

For a one-mile move away from the highway, freight cost increases by the unit freight cost ($4) times the output transported (5), or $20 in total. Dividing this by the size of the production site (two hectares), we get a slope of −$10. A one-mile move away from the highway increases freight cost by $10 per hectare and decreases the bid rent for land by the same amount. Recall the first axiom of urban economics:

Prices adjust to generate locational equilibrium

In this case, variations in the bid rent for land make firms indifferent among all locations: Differences in freight cost are exactly offset by differences in land rent.

FIGURE 6–1 Freight Cost and Manufacturing Bid-Rent Curve



Freight cost increases with distance to the highway, so the bid rent for land decreases to generate zero economic profit at every location.

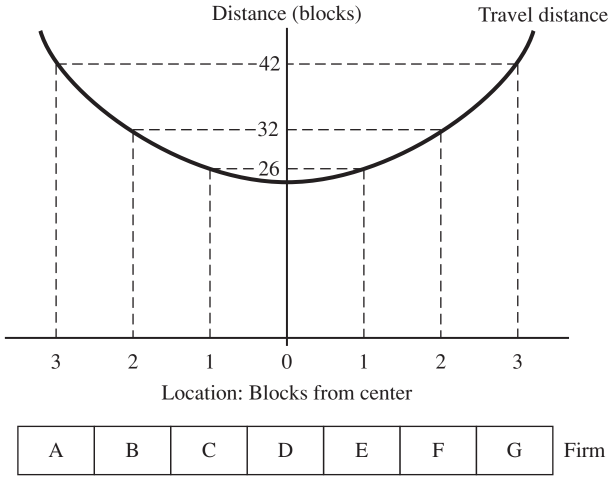
Rent on Industrial Space in Los Angeles

Sivitanidou and Sivitanides (1995) explored the spatial variation in industrial rent in the Los Angeles area. They measured industrial rent as the annual rent per square foot of industrial space. Across the metropolitan area, the rent ranged from $3.12 to $8.12, and the average rent was $5.35. Rent was higher in industrial areas with relatively high freeway density (freeway miles per square mile of land) and higher in areas close to the intersections of major freeways. In addition, rent was higher for industrial sites close to a major airport.

BID-RENT CURVES FOR THE INFORMATION SECTOR

Consider next the city's office sector. Although firms that produce their output in offices provide a wide variety of services, they have a common input and output: information. The firms gather, process, and distribute tacit information, defined as information that cannot be codified in an encyclopedia or operating manual. The transmission of tacit information requires face-to-face contact between people exchanging information—typically high-skilled workers who face a high opportunity cost of travel. Some examples of workers who transmit input and output in this way are accountants, financial consultants, marketing strategists, designers, and bankers. Office firms have an incentive to cluster in an area that provides ready access to information provided by other office firms.

FIGURE 6–2 Travel Distances for Information Exchange



Each office firm interacts with all other office firms in the CBD to exchange information. The total travel distance for information exchange is minimized at the center of the CBD and grows at an increasing rate as the distance to the center increases.

TABLE 6–3 Travel Distance for Firms in a CBD

Firm	Location	Travel Distance: West	Travel Distance: East	Total Travel Distance
D	0	6 = 1 + 2 + 3	6 = 1 + 2 + 3	12 × 2 = 24
E	1	10 = 1 + 2 + 3 + 4	3 = 1 + 2	13 × 2 = 26
F	2	15 = 1 + 2 + 3 + 4 + 5	1	16 × 2 = 32
G	3	21 = 1 + 2 + 3 + 4 + 5 + 6	0	21 × 2 = 42

Travel for Information Exchange

Suppose there are seven firms in a central business district (CBD), spaced one block apart in a straight line. A worker from each firm travels to each of the other firms to exchange information and makes a separate trip to and from each firm. In other words, each trip starts and ends at the firm's location. In Figure 6–2, firms A through G are located one block apart in the CBD, with D at the center. As shown in Table 6–3, firm D travels west to firms C (one block), B (two blocks), and A (three blocks), so its one-way westward travel distance is six blocks. Similarly, the firm travels to the east to firms E, F, and G so its one-way eastward travel distance is six blocks. The firm's one-way travel distance, the sum of westward and eastward travel, is 12 blocks, so its two-way or total travel distance is 24 blocks.

A firm's total travel distance increases as we move away from the center. For firm E, located one block east of the center, the one-way westward distance is longer

(10 blocks for travel to A, B, C, and D) and the one-way eastward distance is shorter (three blocks for travel to F and G), for a one-way distance of 13 blocks and a total travel distance of 26 blocks. Total travel distance is 32 blocks for firm F and 42 blocks for firm G. As shown in Figure 6–2, the total travel distance is minimized at the central location.

Why does the central location minimize total travel distance? The center is the median location, the location that splits travel destinations into two equal halves. One of the fundamental concepts in location theory is the principle of median location:

The median location minimizes total travel distance

When a firm moves away from the median location (the CBD center), its total travel distance increases because the firm moves *farther from at least half* of its destinations and moves *closer to fewer than half* of its destinations. To illustrate, suppose D and E swap places, meaning that D moves one block away from the median location. Firm D is now one block farther from three firms (A, B, C) and one block closer to only two firms (F, G), so its one-way travel distance increases by one block (from 12 to 13) and its total travel distance increases by 2 blocks, from 24 to 26.

As shown in Table 6–3 and Figure 6–2, as we move away from the center location, travel distance increases at an increasing rate. As we move to the east in one-block increments, the total distance increases from 24 to 26 to 32 to 42. This occurs because as a firm moves to the east, it is moving farther from progressively *more* firms to its west and closer to progressively *fewer* firms to its east. For example, if E swaps locations with F, firm E gets closer to one firm (G), but farther from four firms (A, B, C, D). At the extreme, when F swaps places with G, firm F gets farther away from five firms and closer to none.

Office Bid-Rent Curve with a Fixed Lot Size

We can apply the leftover principle to the office sector. Suppose each office firm has a four-story building on 1/4 hectare of land. Each firm produces \$510 worth of output per day and has two types of production costs: the capital cost of the building (\$100) and other costs (for labor, materials, and other inputs) of \$150. For zero economic profit, land rent is computed as follows:

$$\text{Rent per hectare} = \frac{\text{Total revenue} - \text{capital cost} - \text{other production cost} - \text{travel cost}}{\text{Lot size (quantity of land)}}$$

As before, the numerator is the firm's willingness to pay for a lot large enough for the production facility, and the denominator is the lot size. For a firm with a \$10 travel cost, the bid rent per hectare is \$1,000:

$$\text{Rent per hectare} = \frac{\$510 - \$100 - \$150 - \$10}{0.25} = \$1{,}000$$

TABLE 6–4 Office Bid Rent without Factor Substitution

Distance (blocks)	Building Height (floors)	Total Revenue	Capital Cost of Building	Other Nonland Cost	Travel Cost	WTP for Land	Production Site (hectares)	Bid Rent per Hectare
0	4	$510	$100	$150	$ 10	$250	0.25	$1,000
1	4	$510	$100	$150	$ 46	$214	0.25	$ 856
5	4	$510	$100	$150	$210	$ 50	0.25	$ 200

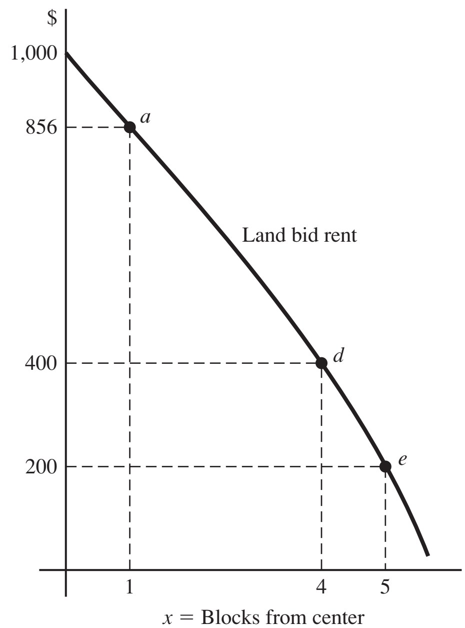
Table 6–4 shows the computed bid rents for different distances from the center. As we move away from the center and travel cost increases, the bid rent for land decreases. If the travel cost at a distance of one block is $46, the bid rent is $856:

$$\text{Rent per hectare} = \frac{\$510 - \$100 - \$150 - \$46}{0.25} = \$856$$

Similarly, if the travel cost at a distance of five blocks is $210, the bid rent is $200.

In Figure 6–3, the bid-rent curve is negatively sloped and concave. The slope is negative because travel cost increases with the distance to the center. The curve

FIGURE 6–3 The Office Bid-Rent Curve without Factor Substitution



The bid-rent curve of office firms is negatively sloped because as we move away from the center, the cost of travel for information exchange increases. The curve is concave because travel cost increases at an increasing rate.

is concave because as we move away from the center, travel cost increases at an increasing rate, so rent decreases at an increasing rate. For example, a move from the center to one block away increases travel cost by $36 and decreases the bid rent per hectare by $144 (point *a*). A move from four blocks to five blocks increases travel cost by $50 and decreases the bid rent per hectare by $200 (point *d* to point *e*). The farther from the center, the larger the increase in travel cost and the larger the reduction in the bid rent for land. Recall the first axiom of urban economics:

Prices adjust to generate locational equilibrium

Differences in the cost of travel for information exchange are fully offset by differences in land rent, so economic profit is zero at all locations.

OFFICE BID-RENT CURVES WITH FACTOR SUBSTITUTION

A key assumption for the bid-rent curve in Figure 6–3 is that each office firm uses a standard office building on the same amount of land. In other words, office buildings at all locations are assumed to be the same height. In fact, office firms near the center occupy tall buildings on small lots.

Building Options: The Office Isoquant

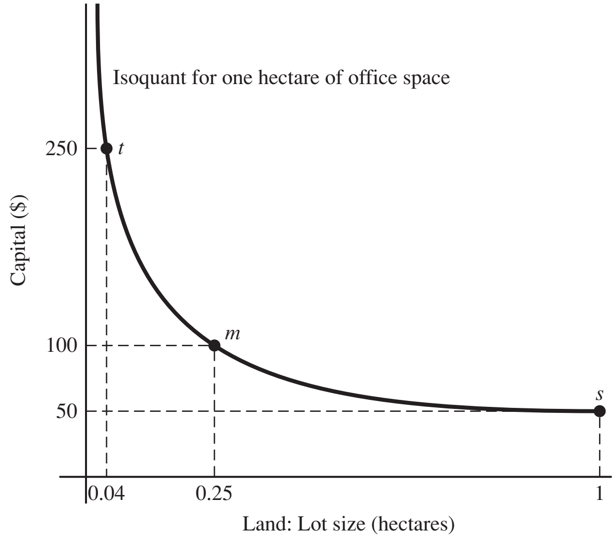
An office firm bases its choice of a building height on the trade-offs between the costs of land and capital. Suppose each office produces the same quantity of output per day and occupies 10,000 square meters of office space (a 100-meter square, which is one hectare). The office space could be in a tall building on a small lot or in a short building on a big lot. The first two rows of Table 6–5 show three options, a 25-story building on 1/25 hectare, a four-story building on 1/4 hectare, and a one-story building on one hectare.

If every office building contains one hectare of office space, do they all have the same amount of capital? A taller building requires more capital because it requires extra reinforcement to support its more concentrated weight, along with extra equipment for vertical transportation (elevators). To see why taller buildings are more expensive, imagine that you borrow a crane and build a 25-story office building by stacking 25 regular mobile homes on top of one another. In addition to the accordion problem (upper floors crushing lower ones), the lack of a vertical transportation system would require workers to rappel from one floor to another. We can

TABLE 6–5 Lot Size, Building Heights, and Capital Cost

	Tall	Medium	Short
Land (hectares)	0.04	0.25	1.0
Building height (floors)	25	4	1
Capital cost ($)	250	100	50

FIGURE 6–4 Isoquant for Office Building



The building isoquant shows the different combinations of land and capital that provide a fixed amount of office space (one hectare = 10,000 square meters). A taller building requires more capital for reinforcement and vertical transportation, so the isoquant is negatively sloped.

avoid these problems by putting more capital in taller buildings. As shown in the third row of numbers in Table 6–5, the tallest building, which is 25 times taller than the shortest building, requires five times as much capital ($250 versus $50).

Figure 6–4 shows the production isoquant for an office building with one hectare of office space. An *isoquant* shows different combinations of inputs (land and capital in this example) that produce a fixed amount of output (*iso* is Greek for equal). The isoquant in Figure 6–4 is simply a graphical representation of the numbers in the first and third rows of Table 6–5. Point *t* shows the input combination for the tall building, while point *m* represents the medium building and point *s* represents the short building.

Factor Substitution: Choosing a Building Height

The isoquant shows the building options for the office firm, and we can use it to explain why buildings are taller near the city center. The firm's objective is to minimize the cost of the building, equal to the sum of land and capital costs. The question is, Which point on the isoquant minimizes the building cost? The answer is that it depends on the prices of land and capital.

We've seen that office firms are willing to pay more for land near the city center. In contrast, the price of capital will be the same at all locations within the city. Table 6–6 (page 136) continues our example of the three types of buildings and shows how the total building cost varies with land rent.

TABLE 6–6 Lot Size, Building Heights, and Building Costs

	Tall	Medium	Short
Land (hectares)	0.04	0.25	1.0
Capital ($)	250	100	50
Building cost with rent = $40			
Land cost	1.6	10	40
Total cost	251.6	110	**90**
Building cost with rent = $200			
Land cost	8	50	200
Total cost	258	**150**	250
Building cost with rent = $1,600			
Land cost	64	400	1,600
Total cost	**314**	500	1,650

- **Low Rent ($40).** Total cost is minimized (at $90) with a short building (point *s* on the isoquant). When land is cheap, it doesn't make sense to build up because the savings from using less land are dominated by the higher capital cost of a taller building. For example, a firm could save $30 in land cost with a medium building ($40 – $10), but would pay $50 more in capital cost ($100 – $50).
- **Medium Rent ($200).** Total cost is minimized (at $150) with a medium building (point *m* on the isoquant). The cost of land is high enough to justify a four-story building, but not a taller one. A taller building would save $42 in land cost but would require $150 more in capital cost.
- **High Rent ($1,600).** Total cost is minimized (at $314) with a tall building (point *t* on the isoquant). When land is expensive, the savings in land costs from using less land dominate the extra capital costs of a tall building.

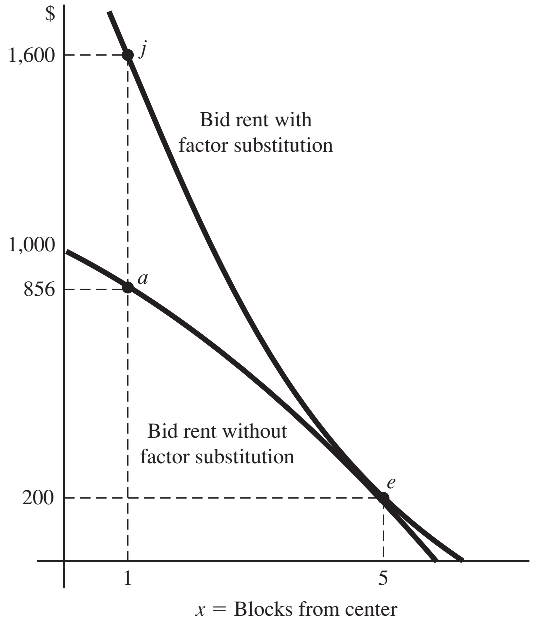
As the price of land increases, a firm responds by substituting capital for land, a process known as input substitution or factor substitution.

How much does an office firm save by engaging in factor substitution? Let's use the medium building as a reference point. An office firm on expensive land ($1,600) builds a tall building at a cost of $314 rather than a medium building at a cost of $500, so the savings from building up (factor substitution) is $186. In the opposite direction, a firm on cheap land builds a short building at a cost of $90 rather than a medium building at a cost of $110, so the savings from factor substitution is $20.

Factor Substitution Generates a Convex Bid-Rent Curve

Figure 6–5 shows the implications of factor substitution for the office bid-rent curve. The concave curve is the bid-rent curve without factor substitution (from Figure 6–3). Suppose that at a site five blocks from the center a four-story building is efficient. As we saw earlier in Table 6–4, the bid rent at this location is $200 (shown by point *e* in Figure 6–5). If an office firm moved to a site one block from the center and continued to use a four-story building, its bid rent would increase to

FIGURE 6–5 The Office Bid-Rent Curve with Factor Substitution



The bid-rent curve for office firms is concave without factor substitution and convex with factor substitution. A move from five blocks from the center to one block increases the bid rent because travel cost decreases (point *e* versus point *a*) and factor substitution saves on building costs (point *a* versus point *j*).

TABLE 6–7 Office Bid Rent with Factor Substitution

Distance (blocks)	Building Height (floors)	Total Revenue	Capital Cost of Building	Other Nonland Cost	Travel Cost	Total Rent Paid	Production Site (hectares)	Bid Rent per Hectare
1	25	$510	$250	$150	$ 46	$64	0.04	$1,600
5	4	$510	$100	$150	$210	$50	0.25	$ 200

$856 (point *a*; computed in the second row of Table 6–4). The increase in bid rent reflects the savings in travel cost at the more central location.

Factor substitution increases the slope of the bid-rent curve. Land is more expensive closer to the center, so it will be rational to occupy a taller building. Suppose at a site one block from the center, the efficient building height is 25 stories. As shown in the first row of Table 6–7, the bid-rent per hectare is $1,600.

$$\text{Rent per hectare} = \frac{\$510 - \$250 - \$150 - \$46}{0.04} = \frac{\$64}{0.04} = \$1{,}600$$

In other words, factor substitution increases the bid rent for land from $856 to $1,600. So a move closer to the center increases the bid rent for land because (1) travel cost decreases and (2) factor substitution cuts the cost of an office building.

The same logic applies to moves away from the center, with a slight twist. Suppose an office firm starts at a site five blocks from the center and then moves farther from the center. In Figure 6–5, if the firm were to use a four-story building at the more distant location, the bid for land would decrease by an amount equal to the increase in travel cost. But given the lower price of land at a more distant location, a shorter building will be efficient, and the cost savings from factor substitution partly offset the effect of higher travel costs. Because of factor substitution, the bid rent decreases by an amount less than the increase in travel cost.

The general effect of factor substitution is to increase the office firm's bid rent for land. An office firm will engage in factor substitution only if it decreases production costs and thus increases its ability to pay land rent. In Figure 6–5, factor substitution transforms a concave bid-rent curve into a convex curve, meaning that as we approach the city center, the price of land rises at an increasing rate. The rapidly increasing price of land in turn encourages more factor substitution, resulting in tall office buildings close to city centers.

We have explained factor substitution with an isoquant and a numerical example. In the appendix to this chapter, we use the full input choice model (with isoquants and isocosts) to provide a more general analysis of factor substitution and its implications for land rent.

Rent on Office Space: Los Angeles and Atlanta

A number of studies have explored the spatial variation in office rent, defined as annual rent per square meter of office space. The principal conclusions of two studies of Los Angeles by Sivitanidou (1995, 1996) are as follows.

1. The office-rent curve is negatively sloped: As the distance to downtown Los Angeles increases, rent decreases. The negative slope reflects the continued attraction power of downtown Los Angeles.
2. Office rent is higher for office sites close to large employment subcenters (clusters of employment). The higher the density of the subcenter, the larger the effect on office rent.
3. Office rent is higher at sites with superior access to freeways and major airports.

The principal conclusions of a study of Atlanta by Bollinger, Khlanfeldt, and Bowes (1998) are as follows.

1. The office-rent curve is relatively flat, with a small negative slope as we move south from the city center, and a small positive slope as we move north from the city center.

2. Between 1990 and 1996, the southern slope became flatter (a smaller negative number), while the northern slope became steeper (a larger positive number).
3. Office rent is higher for sites with superior opportunities for face-to-face meetings among office workers. A location's opportunity for face time was measured by (a) the number of nearby professional workers and (b) the number of nearby workers in industries that provide services to office firms.
4. Office rent is higher for office sites close to highway interchanges.

HOUSING PRICES

Consider next the residential sector of the urban economy. Our goal is to derive the residential bid-rent curve, which shows the bid rent of housing producers for land at different locations in the city. Their bids for land depend on how much consumers are willing to pay for housing, so we'll start by showing how the price of housing varies within the city.

Our model of the housing market focuses attention on commuting as the key location factor for households.

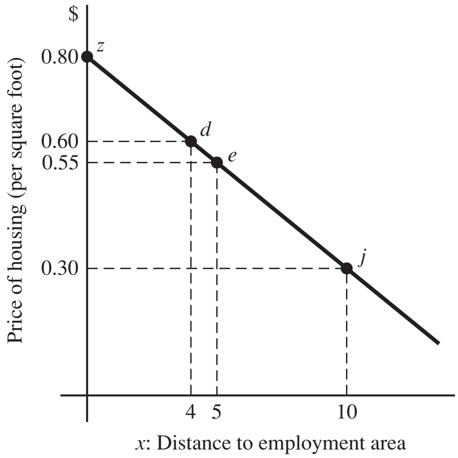
1. The cost of commuting is strictly monetary, a cost of $\$t$ per mile. We ignore the time cost of commuting.
2. One member of each household commutes to a job in an employment area, either the CBD or a manufacturing district.
3. Noncommuting travel is insignificant.
4. Public services and taxes are the same at all locations.
5. Amenities such as air quality, scenic views, and weather are the same at all locations.

These assumptions make the employment area the focal point for city residents. The other things that people care about (e.g., public services, taxes, amenities) are distributed uniformly throughout the city. Later in the chapter, we will relax these assumptions and explore the implications for the price of housing.

Linear Housing-Price Curve: No Consumer Substitution

Suppose for the moment that households do not obey the law of demand. Regardless of the price of housing, each household occupies a standard dwelling, with 1,000 square feet of living space. Suppose the typical household has a fixed amount ($800) to spend on housing and commuting each month. The cost of commuting is $50 per month per mile: A household living one mile from the employment district incurs $50 per month in commuting costs, compared to $100 for a household two miles away, and so on. The price of housing is defined as the price per square foot of housing per month. If a household rents a 1,000 square-foot house (a "standard" house) for $600, the price of housing is $0.60 per square foot ($600/1,000).

FIGURE 6–6 The Housing-Price Curve without Consumer Substitution



The price of housing decreases as the distance to the employment area increases, offsetting commuting costs and ensuring locational equilibrium for households. In the absence of consumer substitution, the housing-price curve is linear.

Figure 6–6 shows the equilibrium housing-price curve. For a standard dwelling next to the employment area ($x = 0$), commuting cost is zero, so the household can spend its entire \$800 budget on housing, paying \$0.80 per square foot for a 1,000 square-foot dwelling (point *z*). In contrast, at a distance of four miles from the employment area, commuting cost is \$200 (equal to \$50 per mile times four miles), so the household has \$600 of its \$800 budget left over to spend on housing and is willing to pay \$0.60 per square foot for a standard dwelling (point *d*). Similarly, at a distance of 10 miles, the household is willing to pay \$0.30, as shown by point *j*.

To see the logic of the negatively sloped housing-price curve, consider what would happen if it were horizontal, with a constant price of \$0.60 at all locations in the city. For a household living 10 miles from the employment area, a move to a location next to the employment area would eliminate \$500 of commuting cost without any change in housing costs. Other households have the same incentive to move closer to the employment area. The demand for housing near the employment area will increase, pulling up housing prices. At the same time, the demand will decrease at more remote locations, pushing down housing prices. In other words, a horizontal housing-price curve will be transformed into a negatively sloped curve.

The equilibrium housing-price curve makes residents indifferent among all locations. Recall the first axiom of urban economics:

Prices adjust to generate locational equilibrium

A move toward or away from the employment area changes commuting cost by the change in distance (Δx) times the commuting cost per mile (t) and changes housing cost by the change in price of housing (ΔP) times housing consumption (h). For locational indifference, the two changes must sum to zero:

$$\Delta P \cdot h + \Delta x \cdot t = 0$$

We can rewrite this expression to show that the change in housing cost equals the negative of the change in commuting cost:

$$\Delta P \cdot h = -\Delta x \cdot t$$

In Figure 6–6, if a household moves from $x = 10$ to $x = 5$ and the price increases by \$0.25, a \$250 increase in housing cost is exactly offset by a \$250 decrease in commuting cost.

$$\$0.25 \cdot 1{,}000 = -(-5) \cdot \$50 = \$250$$

We can use the trade-off expression to get an equation for the slope of the housing-price curve. Dividing each side of the expression by Δx and h,

$$\frac{\Delta P}{\Delta x} = -\frac{t}{h}$$

In our example, $t = \$50$ and $h = 1{,}000$ square feet, so the slope of the housing-price curve is $-\$0.05$:

$$\frac{\Delta P}{\Delta x} = -\frac{t}{h} = -\frac{\$50}{1{,}000} = -\$0.05$$

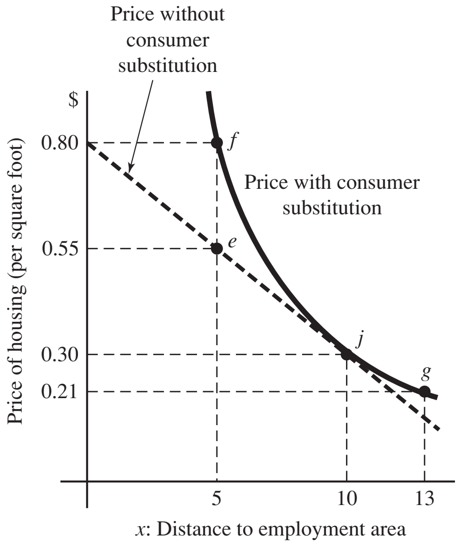
In Figure 6–6, points e and d show the direct computation of the slope: A one-mile move toward the employment center increases the price per square foot of housing by \$0.05.

Consumer Substitution Generates a Convex Housing-Price Curve

The linear housing-price curve in Figure 6–6 reflects the assumption of perfectly inelastic demand for housing. Everyone lives in a 1,000-square-foot house, regardless of the price of housing. In fact, real households obey the law of demand, responding to a higher price by consuming fewer square feet of housing. What are the implications for the housing-price curve?

We saw earlier that a move closer to the employment area generates offsetting changes in commuting costs and housing costs. For a five-mile move inward, commuting cost falls by \$250 (equal to \$50 times five miles) and the price of housing increases by \$0.25. Therefore, a household could use the \$250 savings in commuting costs to exactly cover the higher cost of a 1,000-square-foot dwelling. But if a 1,000-square-foot dwelling is affordable, is that the best use of the consumer's money?

FIGURE 6–7 Consumer Substitution and the Price of Housing



Consumer substitution generates a convex rather than a linear housing-price curve. As distance (x) decreases and the price rises, housing consumption (square feet of space) decreases, increasing the slope of the curve (in absolute value).

In the appendix to this chapter, we use the consumer choice model to show that the consumer will in fact consume less housing at the higher price. The economic intuition for this result is as follows. When the price of housing increases, the opportunity cost of housing increases: The amount of other goods sacrificed for each square foot of housing increases. For example, if the price per square foot increases from \$0.30 at 10 miles to \$0.55 at five miles, the opportunity cost of 100 square feet of housing increases from \$30 worth of meals to \$55 worth. Given the higher opportunity cost, the consumer will rent a house with fewer square feet of space and get more restaurant meals.

As shown in Figure 6–7, consumer substitution increases the slope of the housing-price curve. We can modify the expression for the slope to incorporate consumer substitution by simply replacing h with $h(x)$:

$$\frac{\Delta P}{\Delta x} = -\frac{t}{h(x)}$$

As we approach the employment area (as x decreases), the price of housing increases, so housing consumption decreases. Therefore, the denominator of the slope equation decreases, and the slope increases (in absolute value). The housing-price curve is steeper closer to the employment area, meaning that the curve is convex, not linear. For example, if housing consumption at $x = 5$ is 500 square feet, the slope is

$$\frac{\Delta P}{\Delta x} = -\frac{t}{h(x)} = -\frac{50}{\$500} = -\$0.10$$

Apartment Rents in Portland, Oregon

A recent study of the apartment market in Portland, Oregon, illustrates the notion that housing prices are higher for locations that are accessible to employment centers. A hedonic model of housing is based on the notion that the price or value of a housing unit—house or apartment—is determined by various attributes of the housing unit, for example, its location, size, or age. A hedonic study measures the implicit price of each attribute. The study by Wilson and Frew (2007) shows that an additional bedroom increases monthly rent by about \$85, and access to a swimming pool increases the monthly rent by \$2.

How does the monthly rent on apartments vary within the metropolitan area? In 2002, the monthly rent was about \$700 for apartments just two miles from the city center, compared to only \$300 for apartments 20 miles from the center. Over the first 10 miles, the monthly rent decreases from \$700 to \$400. Between 10 and 14 miles, the monthly rent increases slightly, and then beyond 14 miles, rent decreases rapidly, falling to \$300 for apartments 20 miles from the center. Like other metropolitan areas, Portland area has circumferential highways (beltways) in suburban areas, and apartments located close to beltways have relatively good access to employment areas along the beltways. As a result, there is a local peak in the apartment-rent curve at 14 miles.

THE RESIDENTIAL BID-RENT CURVE

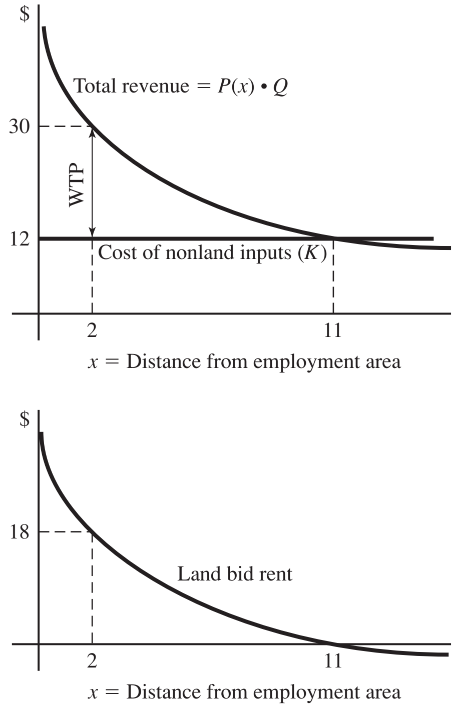
We can use the housing-price curve to derive the residential bid-rent curve, which shows how much housing producers bid for land at different locations. As in the case of manufacturing and office firms, the leftover principle applies: The bid rent generates zero economic profit at each location.

Fixed Factor Proportions

Consider first the situation in which housing is produced with fixed factor proportions. Suppose each housing firm produces Q square feet of housing, using one hectare of land and \$$K$ worth of capital. Once the firm erects a building, it can be used as a single dwelling (with Q square feet of space), or divided into q units, each of which has (Q/q) square feet of living space. For example, a building with 10,000 square feet could be divided into 10 units, each with 1,000 square feet of living space.

In Figure 6–8, we derive the residential land bid rent at different distances from the employment area. The firm's total revenue is the fixed quantity per hectare (Q)

FIGURE 6–8 Residential Land Bid-Rent Curve



The land bid rent of housing producers equals total revenue minus nonland cost (C). The bid-rent curve is negatively sloped and convex, reflecting the negatively sloped and convex housing-price curve. At $x = 11$, total revenue = total nonland cost, so the bid rent is zero.

times the price of housing $P(x)$, which decreases as the distance to the employment area increases. Because the housing-price curve is negatively sloped and convex, so is the firm's total-revenue curve. The gap between the total-revenue curve and the cost curve shows the firm's willingness to pay for a one-hectare lot for a housing complex. Since each firm occupies one hectare of land, the bid-rent per hectare equals the willingness to pay.

Factor Substitution

The bid-rent curve shown in Figure 6–8 is based on the assumption that housing is produced with fixed factor proportions. Housing firms produce the same amount of housing on each hectare at all locations, regardless of the price of land. How would things change if housing firms engaged in factor substitution?

TABLE 6–8 Population Density in Suburbs versus Central City

	Housing (square feet)	Land per Square Foot of Housing	Land per Household (square feet)
Suburb	2,000	2	4,000
City center	1,000	0.10	100

As we saw earlier in the chapter, an increase in the price of land causes firms to substitute capital for land, economizing on land by building taller buildings. As we approach the employment area and the price of land increases, housing firms will build taller buildings on smaller lots. The cost savings from factor substitution are incorporated into the bid-rent curve, causing the bid rent for land to increase more rapidly as we approach the city center. In other words, factor substitution makes the convex residential bid-rent curve even more convex.

Residential Density

How does population density vary within the city? Density increases as we approach employment areas for two reasons:

- **Consumer substitution.** The price of housing increases, and households respond by consuming fewer square feet. In Table 6–8, a suburban resident occupies a 2,000-square-foot house, while a central-city resident occupies a 1,000-square-foot apartment.
- **Factor substitution.** Housing firms respond to higher land prices by using less land per unit of housing. In Table 6–8, a suburban resident lives on a lot that is twice the living area of the house, while a central-city resident lives in a 10-story apartment building, with 0.10 square feet of land for each square foot of housing.

Putting these two factors together, the city-center resident uses 100 square feet of land, while the suburbanite uses 4,000 square feet. Therefore, in this example, population density is 40 times higher in the central city.

RELAXING THE ASSUMPTIONS: TIME COSTS, PUBLIC SERVICES, TAXES, AMENITIES

As explained earlier in the chapter, the basic model of residential land use has a number of convenient but unrealistic assumptions.The first assumption is that commuting has no time cost. In fact, commuting time comes at the expense of work or leisure, so there is an opportunity cost. Studies of travel behavior suggest that the typical person values commuting time at between one-third and one-half the wage rate. The higher the opportunity cost of commuting, the steeper the housing-price curve and the residential bid-rent curve.

The model also assumes that there is one worker per household. For the more realistic case of two-earner households, we must keep track of two commuters and their costs. If the two earners have the same workplace, the savings in commuting costs from moving toward the employment area would double, increasing the slope of the housing-price curve. If the two earners work at different locations, things are not so tidy. The housing-price curve could be steeper, flatter, or even positively sloped.

The basic model also assumes that noncommuting travel—for shopping, entertainment, and other activities—is insignificant. If the destinations for noncommuting travel are distributed uniformly throughout the urban area, this assumption is harmless. A change in residence in one direction would decrease travel costs to some destinations but increase travel costs to others, and the net change in total travel costs would be relatively small. In contrast, if noncommuting trips are concentrated, households will be oriented to this destination, along with the employment area. Everything else being equal, a move away from the noncommuting destination would decrease the price of housing.

The fourth assumption is that public services and taxes are the same at all locations. Suppose instead that a city has two school districts with the same taxes, but one district has better schools. Competition among households will bid up the price of housing in the superior district. Instead of paying directly for better schools with higher taxes, people pay indirectly with higher housing prices. The same logic applies to variation in taxes. If two communities have the same level of public services but different taxes, competition among households will bid up the price of housing in the low-tax community.

The model also assumes that environmental quality is the same at all locations in the city. Suppose instead that a polluting factory moves into the center of a previously clean city, and the smoke and smell from the factory are heaviest in the central area. The factory will decrease the relative attractiveness of dwellings near the city center, decreasing the price of housing. For the opposite case of positive amenities, housing prices will be higher for sites that provide scenic views or access to parks.

LAND-USE PATTERNS

We can use the bid-rent curves of different land users to determine a city's equilibrium land-use patterns. In the market equilibrium, land is allocated to the highest bidder. In our simple model of the urban economy, three sectors compete for land: manufacturers, office firms, and residents.

The first step in determining land-use patterns is to specify the features of the urban transportation system. We will assume that manufacturers export their output from the urban area, using trucks that travel on intercity highways. An intercity highway goes through the center of the metropolitan area, and a circumferential highway (a beltway) is connected to the intercity highway. Firms in the office sector exchange information in a central business district. Residents work in offices and firms and travel by automobile from their homes to their workplaces.

Bid-Rent Curves for Business

Figure 6–9 shows the bid-rent surfaces of the office and manufacturing sectors. Panel A shows bid rent for the office sector, which peaks at the center of the metropolitan area. Because of the face-to-face contact required in the exchange of information in the office sector, the bid-rent curve is relatively steep. As an office firm moves away from the center, the cost of interacting with other office firms increases rapidly, so the bid rent decreases rapidly. As shown in Panel B of Figure 6–9, the bid rent of manufacturers reaches its highest level for sites along the highway (shown by the straight ridge) and the beltway (shown by the circular ridge four miles from the city center). As we move away from the highway or beltway, intracity freight costs increase, decreasing the bid rent for manufacturing land.

Panel C of Figure 6–9 puts the two sets of business bid-rent curves together and shows the maximum bid rents for business land users. At the city center, the office bid rent exceeds the manufacturing bid, indicating that the office sector has more to gain from being in the center. This is sensible because the transportation cost of office firms involves the travel of people rather than the shipping of goods. Because the office bid rent falls rapidly as distance to the center increases, manufacturing firms outbid the office firms for more distant locations. Along the highway, manufacturing takes over once we reach about half a mile from the center (where the highway ridge meets the office bid-rent cone). The beltway is far enough from the city center that office firms don't provide any competition for manufacturing firms.

FIGURE 6–9 Panel A: Bid Rent of the Office Sector

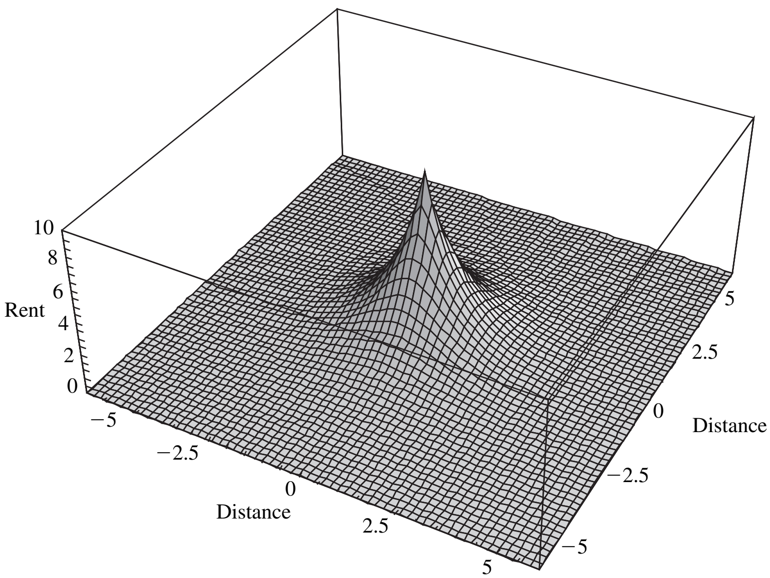


FIGURE 6–9 Panel B: Bid Rent of the Manufacturing Sector

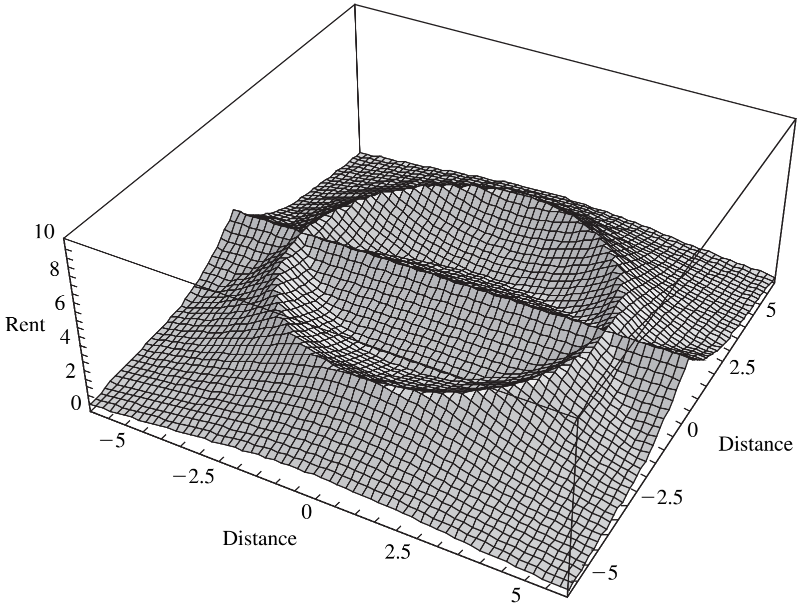


FIGURE 6–9 Panel C: Maximum Bid Rent of Employers

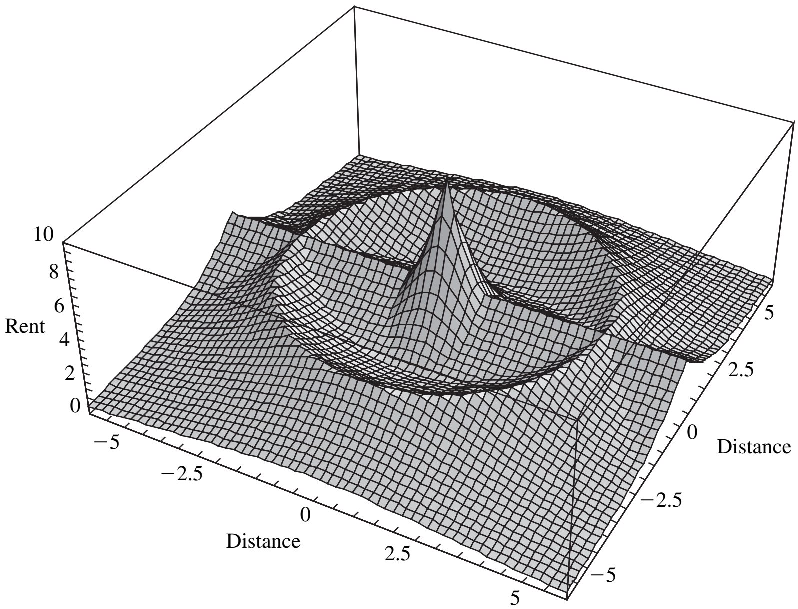
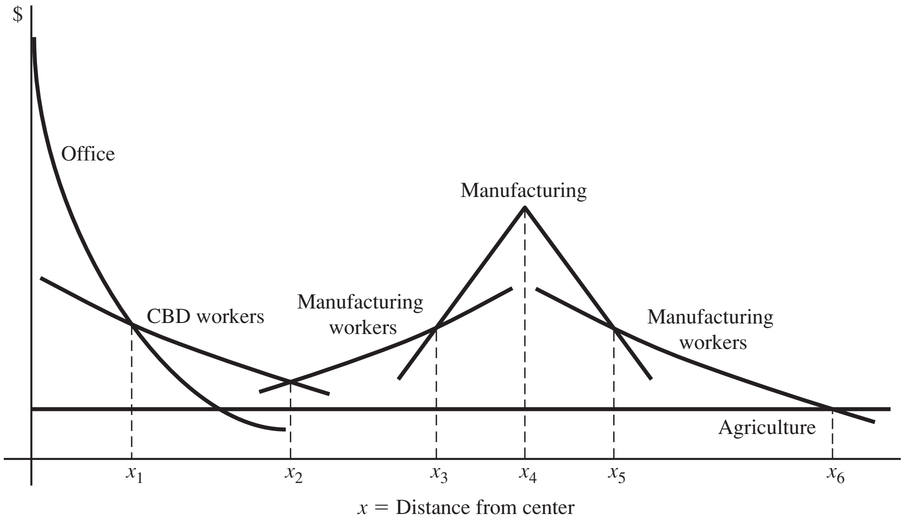


FIGURE 6–10 Bid Rents and Land Use Patterns



The equilibrium land-use pattern is determined by the bid-rent curves of firms and residents. The CBD is the area over which office firms outbid other users (from $x = 0$ to x_1). The area between x_1 and x_2 is occupied by residents who work in the CBD. Manufacturing workers live in the areas between x_2 and x_3 and x_5 and x_6. Manufacturers occupy the area between x_3 and x_5.

Territories of Different Sectors

In our simplified model of the urban economy, we assume that each household is oriented toward its workplace. There are three employment districts: the circular CBD, the area close to the highway, and a ring surrounding the beltway. If we superimposed all the residential bid-rent curves, one for each employment area, we could show the different residential and business areas.

To provide a clearer picture of the different territories, we use Figure 6–10 to take a two-dimensional perspective, with distance from the center on the horizontal axis. This figure omits the highway manufacturing districts, but it includes the beltway districts. The office sector has a negatively sloped bid-rent curve that peaks at the city center. The first residential bid-rent curve is for office workers who travel inward to the CBD. The beltway is located x_4 miles from the center, and the bid rent of the manufacturing sector peaks at the beltway location. There are two bid-rent curves for manufacturing workers, one for those who travel outward and one for those who travel inward.

Figure 6–10 shows competition between workers and firms for urban land. Land is occupied by the highest bidder, and the intersection of the bid-rent curves of office firms and office workers shows the boundary between the business and residential area, x_1. Similarly, the boundary between office workers and manufacturing workers occurs where the bid-rent curves of the two worker types intersect, at x_2.

Manufacturing firms outbid their workers for locations between x_3 and x_5, defining the manufacturing district. The outer residential district, defined as the area over which manufacturing workers outbid their employers and a nonurban land use (agriculture), is between x_5 and x_6.

HENRY GEORGE AND THE SINGLE TAX ON LAND

We've seen that the price of agricultural land is determined by its fertility, and the price of land in cities is determined by its accessibility. The rapid development of cities in the late 19th century caused substantial increases in land prices. In 1880, Henry George proposed a 100 percent tax on land rent. The proposed tax was dubbed the "single tax" because it would have generated enough revenue to support all levels of government at the time. The spirit of the single tax is best expressed by George himself. The following is from an interview with David Dudley Field (in the *North American Review* in 1885):

FIELD: Then suppose A to be the proprietor of a thousand acres on the Hudson, chiefly farming land, but at the same time having on it houses, barns, cattle, horses, carriages, furniture; how is he to be dealt with under your theory?

GEORGE: He would be taxed on the value of his land, and not on the value of his improvements and stock. . . . The effect of our present system, which taxes a man for values created by his labor and capital, is to put a fine upon industry, and repress improvement. The more houses, the more crops, the more buildings in the country, the better for us all, and we are doing ourselves an injury by imposing taxes upon the production of such things.

FIELD: Then you would tax the farmer whose farm is worth $1,000 as heavily as you would tax the adjoining proprietor, who, with the same quantity of land, has added improvements worth $100,000; is that your idea?

GEORGE: It is. The improvements made by the capitalist would do no harm to the farmer, and would benefit the whole community, and I would do nothing to discourage them.

FIELD: A large landlord in New York owns a hundred houses, each worth, say, $25,000 (scattered in different parts of the city); at what rate of valuation would you tax him?

GEORGE: On his houses, nothing. I would tax him on the value of the lots.

FIELD: As vacant lots?

GEORGE: As if each particular lot were vacant, surrounding improvements remaining the same.

FIELD: Well, what do you contemplate as the ending of such a scheme?

GEORGE: The taking of the full annual value of land for the benefit of the whole people. I hold that land belongs equally to all, that land values arise from the presence of all, and should be shared among all.

George proposed the single tax for both equity and efficiency reasons. On the equity issue, George argued that land rent is determined by nature and society, not by the efforts of landowners. The rent on a piece of urban land is determined by the site's accessibility to other activities. In George's time, public investments in transportation (railroads, cable cars) increased accessibility, increasing the demand for urban land and urban land rent. George argued that windfall gains from urban growth should be taxed away. On the efficiency issue, George argued that the land tax would eliminate the need for taxes on improvements, increasing investment in houses and buildings. In contrast, because the supply of land is fixed, the land tax would simply redistribute income without affecting the quantity of land. George argued that replacement of the improvement tax with the land tax would increase the total wealth of society.

The single tax has been criticized for three reasons. First, the single tax would decrease the net return to the landowner (net land rent) to zero, making the market value of land zero. In other words, the government would essentially confiscate the land. This strikes many people as inequitable. Second, if the net return on land were zero, landowners would abandon their land, leaving government bureaucrats to decide who uses the land. Unlike the private owner, who receives more income if the land is used efficiently, the bureaucrat has nothing to gain from the efficient use of land. Therefore, the government land market is less likely to allocate land to its highest and best use. The third criticism is that it is difficult to measure land rent (and the appropriate tax). Most land has structures or other improvements, and it is difficult to separate the value generated by the raw land from the value generated by the improvements.

There are two alternatives to a single tax. Under a partial land tax, the tax rate is less than 100 percent. A partial land tax would leave landowners with a positive net return, so the land market would continue to be run by those who have a private interest in allocating land to its highest bidder. A second alternative is the two-rate tax, or the split property tax. Under the conventional property tax, land and improvements are taxed at the same rate. A 3 percent property tax is actually a 3 percent tax on land and a 3 percent tax on improvements. Under a split tax, the tax rate on land may be 9 percent, while the tax rate on improvements may be 1 percent. The split tax is widely used in Australia and New Zealand. It is also used in some cities in Pennsylvania. By imposing a lower tax on improvements, the split tax would increase investment in housing, buildings, and other improvements.

SUMMARY

In this chapter, we've shown that the price of urban land is determined by its accessibility. Here are the main points of the chapter:

1. The leftover principle tells us that the bid rent for land equals the excess of total revenue over nonland cost.
2. Manufacturing firms are oriented toward highways that link the city to markets outside the city. Intracity freight cost increases with distance to the highway, so the bid-rent curve is negatively sloped.

3. **3.** Office firms exchange information, and the median location has the minimum travel cost and the maximum bid rent.
4. **4.** Tall buildings result from factor substitution in response to high land prices. The savings in production costs from factor substitution increase the bid rent for land.
5. **5.** Residents are oriented toward employment areas, and commuting costs generate negatively sloped and convex housing-price curves.
6. **6.** Land is allocated to the highest bidder, so we can use the bid-rent curves of different land users to predict land-use patterns.
7. **7.** Henry George proposed a single tax on land and the elimination of the tax on improvements.

APPLYING THE CONCEPTS

For exercises that have blanks (________), fill each blank with a single word or number. For exercises with ellipses (…), complete the statement with as many words as necessary. For exercises with words in square brackets ([increase, decrease]), circle one of the words.

1. Rent for GreenGene

Mr. Greengene grows corn on the land he rents from Lauren, initially at a price of $500 per hectare. Suppose Mr. Greengene develops a new method for growing corn that decreases the growing cost by $300 per hectare. Lauren rejoices, citing the leftover principle as she counts on collecting $800 in rent instead of $500.

a. Greengene's rent will increase by $300 per hectare if ....

b. Greengene's rent will be unchanged if ....

2. Gandhi and the Leftover Principle

In 1917, Mahatma Gandhi settled a dispute between Indian farmers and British landowners. Under a share-cropping arrangement, each indigo farmer gave 15 percent of the harvest to the landowner. When landowners heard about the development of synthetic indigo, they quickly sold the land to the farmers, who at the time didn't know about synthetic indigo and the upcoming collapse of indigo prices. When the price of indigo dropped, the farmers who had purchased land demanded their money back. Gandhi negotiated a partial refund of the payments. Imagine that you are Gandhi's research assistant and must compute the appropriate refund for the typical new landowner.

- The initial price of indigo is $10. The annual output is 100 units per hectare, and the annual nonland cost.
- To purchase land, farmers borrow money at an interest rate of 10 percent per year.
- The alternative crop is rice, which has a price of $8, an annual output of 100 units, and a nonland cost (including the opportunity cost of the farmer) equal to the nonland cost of indigo.

a. Under the original share-cropping arrangement, the effective annual rent on land is _______ per hectare, computed as … This implies that the nonland cost per hectare is _______, computed as ….

b. Before farmers find out about synthetic indigo (while they are assuming that the price will be $10), the anticipated annual profit from indigo is _______ per hectare, computed as … A farmer's willingness to pay for a hectare of land is _______, computed as ….

c. Suppose that the development of synthetic indigo decreases the price of indigo to $5. The new annual profit from a hectare of land is _______, computed as ….

d. A farmer's willingness to pay is _______ per hectare. The appropriate refund, to be paid by the old landowners to the new landowners, is _______, computed as ….

3. Compute the Manufacturing Bid Rent

Consider a manufacturing firm that occupies two hectares of land. The firm produces 10 tons of output per day and sells its output at a price of $80 per ton. The firm does not engage in factor substitution as the price of land changes. Intraurban transportation is on trucks, with a unit cost of $12 per ton per mile. The firm's nonland cost is $200 per day. The firm exports its output via circumferential highway (beltway).

a. Draw the firm's bid-rent curve for land for different distances from the beltway, from a distance of zero to five miles.

b. The bid rent at the beltway is _______ per hectare and the slope of the bid-rent curve is _______ per mile, computed as ….

4. Matter Transmitter for Manufacturing

Consider a manufacturing industry that exports its output by ship. Each firm has total revenue per month of $1,400 and a monthly nonland production cost of $400. Each firm initially transports its output from its factory at location x to the port ($x = 0$) on trucks. A firm's freight cost is $100 per block from the port. Suppose a second transport option is developed: For a monthly rental cost of $300, a firm can use a matter transmitter to transport its output from its factory to the port, up to a distance of seven blocks. The marginal cost of the matter transmitter is zero.

a. Draw two bid-rent curves for manufacturers, one for firms that use the truck and a second for firms that use the transmitter, for zero to 10 blocks from the port.

b. Firms continue to use the truck over the following interval(s): _______ to _______ ; _______ to _______.

c. Firms use the transmitter over the interval _______ to _______.

5. Benefits from Street Improvements

Suppose a city widens the streets in an industrial area, thus improving the access of trucks to an interstate highway.

a. In the long run, the bulk of the benefits of the street improvements go to [workers, trucking firms, landowners, manufacturing firms] because ….

b. Illustrate with a graph.

6. **Opportunity Cost and Office Bid-Rent Curve**
   Using Table 6–4 and Figure 6–3 as a starting point, suppose the opportunity cost of travel by office workers increases by 25 percent.
   a. At a distance of five blocks, travel cost = ______ ; the willingness to pay for land = ______; the bid rent per hectare = ______.
   b. Draw the old and the new bid-rent curves for the office sector.
   c. The new bid-rent curve is [flatter, steeper] and reaches a value of zero at a distance of $x =$ ______.
7. **Segway for Information Workers**
   The Segway Human Transporter is a self-balancing personal transportation vehicle that is clean (battery powered) and small (its footprint is 19 by 25 inches) so it can be used on sidewalks and inside buildings. Suppose the segway is introduced into a CBD, doubling the speed of travel for information exchange.
   a. Show the effects of the segway on the bid-rent curve for the office sector in two circumstances:
   - Fixed building heights: Office firms do not engage in factor substitution. The gap between the old bid-rent curve and the new bid-rent curve [increases, decreases] as distance increases.
   - Variable building heights.

   b. In the long run, the bulk of the benefits of the segway go to [office firms, office workers, landowners].
8. **Housing Price and Land Bid-Rent Numbers**
   Consider a monocentric city where the cost of commuting is $40 per mile per month. A household located eight miles from the city center occupies a dwelling with 1,000 square feet at a monthly rent of $600. Nonland cost per dwelling is $250, and there are 10 houses per hectare.
   a. The price of housing at a distance of eight miles is ______ per square foot, computed as . . . .
   b. The bid rent for land at a distance of eight miles is ______ per hectare, computed as . . . .
   c. Assume that the demand for housing is perfectly inelastic. The price of housing at a distance of five miles is ______ per square foot, computed as . . . .
   d. Assume that housing firms do not engage in factor substitution. The bid rent for land at a distance of five miles is ______ per hectare, computed as . . . .
   e. Suppose consumers engage in consumer substitution and firms engage in factor substitution. The bid rent for land at a distance of five miles would be [greater, less] than the number computed in part (d) because . . . .
9. **Violating the Law of Demand**
   Consider a region with two cities: Obeyburg (B) and Vioville (V). The cities differ in the individual demand curves for housing. Consumers in Obeyburg obey the law of demand, with negatively sloped individual demand curves. Consumers in Vioville violate the law of demand, with positively sloped

individual demand curves. Draw the housing-price curves for the two cities (labeled P_B for Obeyburg and P_V for Vioville) under the assumption that $P_B = P_V$ at a distance of five miles from the city center. Hint: How does conventional consumer substitution affect the shape of the housing-price curve, and how would contrary consumer substitution affect the shape?

10. Crime and Housing Prices

Consider a city where everyone commutes to the city center, and commuting cost per mile per month is \$40. Each household occupies a 1,000-square-foot dwelling and has \$7,000 worth of possessions in its dwelling. The probability that any particular household will be burglarized and lose all its posessions (no insurance) is 0.10 at the city center and decreases by 0.01 per mile (to 0.09 at one mile, 0.08 at two miles, and so on). The price of housing is \$1.00 per square foot at the city center.

a. Starting from the center, a one-mile move outward changes the expected value of the loss from crime from ________ to ________, a change of ________ per square foot.

b. The slope of the housing-price curve is ________, computed as . . .

c. Draw the housing-price curve for locations up to five miles from the city center. The price changes from ________ at the city center to ________ five miles away.

REFERENCES AND ADDITIONAL READING

1. Bollinger, Christopher, Keith Ihlanfeldt, and David Rowes. “Spatial Variation in Office Rents within the Atlanta Region.” *Urban Studies* 35 (1998), pp. 1097–1118.
2. Fujita, Mashisa, and Jacques-Francois Thisse. *Economics of Agglomeration*. Cambridge: Cambridge University Press, 2002.
3. George, Henry. *Progress and Poverty.* New York: Schalkenbach Foundation, 1954.
4. O’Hara, D. J. “Location of Firms within a Square Central Business District.” *Journal of Political Economy* 85 (1977), pp. 1189–1207.
5. Sivitanidou, Rena. “Urban Spatial Variations in Office-Commercial Rents: The Role of Spatial Amenities and Commercial Zoning.” *Journal of Urban Economics* 38 (1995), pp. 23–49.
6. Sivitanidou, Rena. “Do Office-Commercial Firms Value Access to Service Employment Centers?” *Journal of Urban Economics* 40 (1996), pp. 125–49.
7. Sivitanidou, R., and P. Sivitanides. “Industrial Rent Differentials: The Case of Greater Los Angeles.” *Environment and Planning* A 27 (1995), pp. 1133–46.
8. Wilson, Beth, and James Frew. “Apartment Rents and Locations in Portland, Oregon: 1992–2002.” *Journal of Real Estate Research* 29 (2007), pp. 201–217.

Appendix: Consumer and Factor Substitution

In this appendix, we provide more rigorous analysis of two results in Chapter 6. First, consumers obey the law of demand, consuming less housing as the price of housing increases. This result explains the convex shape of the housing-price curve. Second, firms engage in factor substitution, substituting capital for land as the relative price of land increases. This result explains why the bid-rent curve of the office sector is convex rather than concave. For a review of the consumer choice model and the input choice model, see the Appendix at the end of the book, "The Tools of Microeconomics."

CONSUMER CHOICE AND THE LAW OF DEMAND

The consumer choice model is a model of constrained maximization. It shows how consumers make choices to maximize their utility, subject to the constraints imposed by their income and the prices of consumer goods.

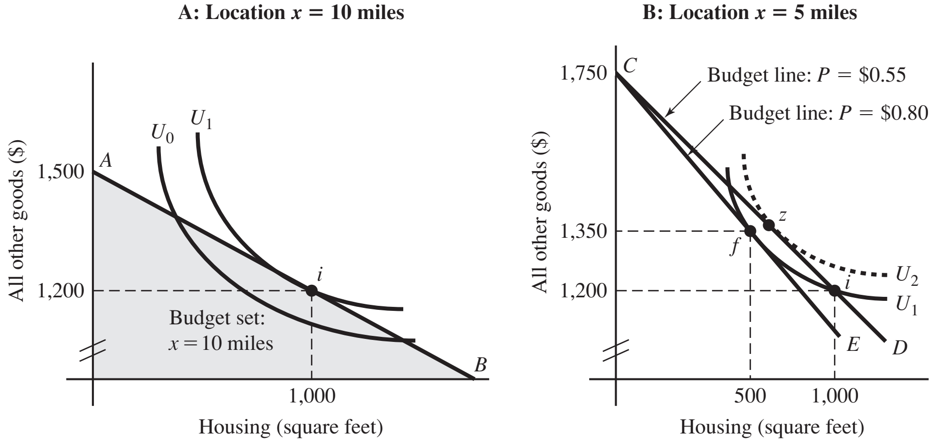
A consumer's budget set shows all the affordable combinations of two goods. In Panel A of Figure 6A–1, the shaded area is the budget set for a household located 10 miles from the employment area. The two goods are housing (measured in square feet) and all other goods (measured in dollars). Suppose the household has an income of $2,000 and its commuting cost is $500 per month ($50 per month per mile times 10 miles). The price of housing at 10 miles is $0.30 per square foot. The budget line *AB* shows the combinations of housing and other goods that exhaust the household's budget. Consider two points on the budget line.

- **Point *A*.** If housing consumption is zero, the household can spend $1,500 on other goods: $1,500 = $2,000 income −$500 spent on commuting.
- **Point *i*.** If housing consumption is 1,000 square feet, the household spends $300 on housing and $500 on commuting, leaving $1,200 for other goods.

The slope of the budget line shows the market trade-off between housing and other goods. The slope is the quantity of other goods that must be sacrificed for each additional square foot of housing. If the price of housing is $0.30, each square foot of housing reduces the amount of other goods by $0.30. In other words, the slope is simply the price of housing.

Consumer preferences are represented by indifference curves. Each indifference curve shows the different combinations of two products that generate the same level of utility. The marginal rate of substitution (MRS) is the amount of other goods the consumer is willing to sacrifice to get one more square foot of housing, and it is shown by the slope of the indifference curve. For example, if the slope is −$0.50, the household is willing to sacrifice $0.50 units of other goods for one square foot of housing: MRS = 0.50.

FIGURE 6A–1 Economic Choice Model



At location x = 10 miles, utility is maximized where the budget line is tangent to an indifference curve (point i), where MRS equals the price ratio of $0.30.

At location x = 5, commuting cost is $250 lower and the price of housing is higher. Although point i is affordable, it does not maximize utility. Location equilibrium is restored at point f, with a price of $0.80 and the same utility as at x = 10.

Maximizing Utility: MRS = Price Ratio

To maximize utility, a consumer finds the highest indifference curve within its budget. In Panel A of Figure 6A–1, the indifference curve U_1 is the highest indifference curve within the consumer budget set, so utility is maximized at point i, with 1,000 square feet of housing and $1,200 of other goods. At this point, the indifference curve is tangent to the budget line, meaning that the slope of the budget line (the price ratio = 0.30) equals the slope of the indifference curve (the marginal rate of substitution). In fact, this is the rule for utility maximization:

Marginal rate of substitution = Price ratio

If the price ratio (the market trade-off) equals the marginal rate of substitution (the consumer's own trade-off), the consumer can't do any better.

Consumer Substitution

Panel B of Figure 6A–1 shows budget lines for a location five miles from the employment area. When a household moves inward from 10 miles to five miles, its commuting cost decreases by $250. If the price of housing at 5 miles is $0.55 per square foot, CD is the budget line.

- **Point *C*.** If housing consumption is zero, the household can spend all the money it saves on commuting costs (\$250) on other goods. The maximum for other goods increases from \$1,500 (point *A* in Panel A) to \$1,750.
- **Point *i*.** If housing consumption remains at 1,000 square feet and the price of housing is \$0.55, the change in commuting cost (\$250) is exactly offset by higher housing cost (\$250), so the original combination shown by point *i* is still affordable.

Although point *i* is still affordable, the household will not choose it. Given the higher price of housing (\$0.55, up from \$0.30), the indifference curve U_1 is not tangent to the budget line at point *i*, so utility is no longer maximized at point *i*. Point *i* violates the utility-maximizing rule MRS = price ratio: The price ratio is 0.55 and the MRS (the consumer's own trade-off) is, of course, still 0.30. As a result, the household can do better.

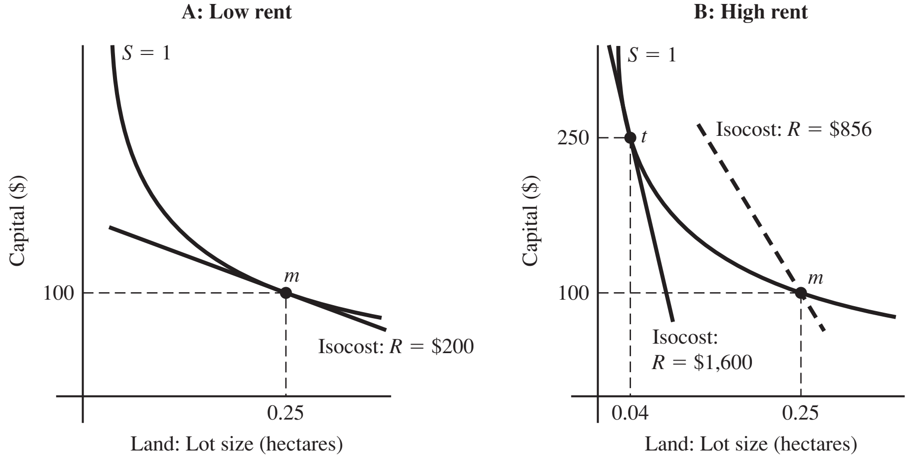
Will housing consumption increase or decrease? At a price of \$0.55, the household now sacrifices \$0.55 of other goods for every square foot of housing, but it is willing to sacrifice only \$0.30 of other goods (given the MRS). The household has too much housing and will reduce its consumption. For each square foot reduced, the household gets \$0.55 of other goods—more than the amount the household needs to be indifferent about the change (MRS = 0.30 at point *i*). If the price of housing were to remain at \$0.55, the household would move up the budget line *CD* to a higher indifference curve (U_2) and a higher utility level, consuming less housing and more of other goods (shown by point *z*).

Locational Equilibrium

Unfortunately for the household, the price of housing at location $x = 5$ will not be \$0.55 but will be higher. Imagine for the moment that a household at $x = 5$ reaches utility level $U_2 > U_1$. Everyone will want to live at $x = 5$, and the resulting increase in demand for housing there will bid up its price until households are indifferent between the two locations ($x = 5$ and $x = 10$). For locational equilibrium, the utility level at $x = 5$ must be the same as the utility level at $x = 10$ (equal to U_1). To equalize utility, the price of housing at $x = 5$ rises above \$0.55, causing the budget line to tilt inward, with a steeper slope. The price will continue to increase until we get a tangency between the new budget line and the indifference curve associated with the original utility level U_1. In Panel B of Figure 6A–1, this happens at point *f*: When the price of housing is \$0.80, utility is maximized at level U_1.

This analysis shows how households living in different locations can achieve the same level of utility. Compared to the household living 10 miles from the employment area, the household living at five miles consumes less housing (500 square feet compared to 1,000) but consumes more of other goods (\$1,350, compared to \$1,200). Both households reach the same utility level because points *i* (chosen by the 10-mile household) and *f* (chosen by the five-mile household) lie on the same indifference curve. Housing consumption is lower with the higher housing price, consistent with the law of demand.

FIGURE 6A–2 Factor Substitution and the Price of Land



A: When land rent = \$200, cost is minimized at point *m*, where the isocost line is tangent to the isoquant. B: With offsetting changes in land rent and travel cost, the firm could still choose point *m*, but because the land rent is higher, factor substitution will decrease cost. The resulting increase in the bid rent for land increases the slope of the isocost, and cost is minimized at point *t*.

INPUT CHOICE AND FACTOR SUBSTITUTION

Consider next the input choices of office firms. As we saw in the chapter, the isoquant shows the different combinations of land and capital that generate a fixed amount of office space (one hectare). The objective of the office firm is to find the input combination on its isoquant that has the lowest building cost, equal to the sum of capital and land costs.

Figure 6A–2 shows how a firm picks the least costly input combination. The linear curve is an isocost line, the analog of the consumer's budget line. An isocost shows the combinations of two inputs that exhaust a fixed input budget. The slope of the isocost line is the market trade-off between the two inputs, the price of land (on the horizontal axis) divided by the price of capital (on the vertical axis). The price of capital is assumed to be \$1, so the slope is simply the price of land. A higher (more northeasterly) isocost represents a bigger budget (more spent on both inputs). To minimize cost, the firm gets on the lowest (most southwesterly) isocost that touches the isoquant.

Panel A of Figure 6A–2, shows the firm's choice when the price of land is \$200 per hectare. Cost is minimized at point *m*, with 0.25 hectares of land and 100 units of capital. Point *m* minimizes cost because it is on the lowest feasible isocost. The two curves are tangent at the cost-minimizing point, consistent with the cost-minimizing rule for input choices:

Marginal rate of technical substitution = Input price ratio

The marginal rate of technical substitution (MRTS) is the analog of the marginal rate of substitution. The MRTS of capital for land is the change in capital per unit change in land, keeping the quantity produced constant. When the MRTS equals the input price ratio, the production trade-off between two inputs equals the market trade-off, so the firm can't produce its target output at any lower cost. At point *m*, the MRTS = 200, equal to the input price ratio (land price = $200 and capital price = $1).

Suppose the firm moves toward the city center from a location five blocks from the center to a location one block from the center. If the firm does not engage in factor substitution, the decrease in travel cost will be exactly offset by higher rent, so point *m* will still be possible with the original budget: The firm takes its savings in travel cost and puts it into land. According to the leftover principle, the rent is $856:

$$\text{Rent per hectare} = \frac{\$510 - \$100 - \$150 - \$46}{0.25} = \frac{\$214}{0.25} = \$856$$

Although point *m* is affordable, the firm will not choose it. The higher price of land generates a steeper isocost: In Panel B of Figure 6A–2, the solid isocost is steeper than the dashed original isocost. As a result, point *m* is not the cost-minimizing point. The MRTS (still 200 at point *m*) is less than the price ratio (now 856), so the firm will substitute capital for the more expensive land, moving upward along the isoquant to a lower isocost line (lower cost).

Factor substitution cuts the firm's cost and increases its bid rent for land. The leftover principle tells us that the cost savings from factor substitution bid up the price of land, tilting the isocost line and increasing its slope. In Panel B of Figure 6A–2, cost is minimized at point *t*, with a 25-story building and 250 units of capital. The MRTS equals the price ratio of 1,600. As in the case of consumer substitution, factor substitution is caused by an increase in land rent, and in turn increases the bid rent for land.

CHAPTER 7

Land-Use Patterns

Otis's apparatus (the elevator) recovers the uncounted planes that have been floating in the thin air of speculation.

—Rem Koolhaas

In modern metropolitan areas, jobs are divided between central business districts, suburban subcenters, and "everywhere else." It turns out that most jobs are elsewhere—widely dispersed throughout the metropolitan area—and most people work and live far from the center. In this chapter, we describe the spatial distributions of employment and population within cities, then look back about 100 years to a different urban reality. In the heyday of the monocentric city, between two-thirds and three-fourths of jobs were near the center. We'll explore the market forces behind the transformation of cities and discuss the causes and consequences of urban sprawl.

THE SPATIAL DISTRIBUTION OF EMPLOYMENT

How are jobs distributed across the typical metropolitan area? One approach to answering this question is to divide a metropolitan area into two parts, a central area and the rest of the metropolitan area, and show how jobs are divided between the two areas. A second approach is to take a closer look at the spatial distribution of jobs throughout the metropolitan area, using a smaller geographical unit such as a census tract—a small, relatively permanent statistical subdivision of a county, with between 2,500 and 8,000 residents. We'll start with the division of jobs between the central area and the rest of the metropolitan area and then switch to census-tract data.

Jobs Inside and Outside the Central Area

Recall that the central city of a metropolitan area is defined as the large central municipality. In other words, the boundary of a central city is a political—not economic—boundary. The typical metropolitan area has many other municipalities;

TABLE 7–1 Employment Inside and Outside Central Cities, 1980–2000

	1980	1990	2000
In Central Cities	35.21	46.47	49.03
In Other Municipalities (numbers in millions)	31.58	43.75	53.75

Source: U.S. Census, Journey to Work.

these other municipalities comprise the "suburban" area—the rest of the metropolitan area. Table 7–1 shows the number of metropolitan jobs inside and outside central cities. In 1980, central cities had about 11 percent more jobs. Between 1980 and 2000, central cities grew slower than the other municipalities and, by 2000, central cities had 10 percent fewer jobs. This trend is actually a continuation of a long trend of employment decentralization. In 1948, central cities had roughly twice as many jobs as the other metropolitan municipalities.

We can also distinguish between a fixed central area and the rest of the metropolitan area. For the largest 100 metropolitan areas, roughly 22 percent of jobs are within three miles of the center, and 65 percent of jobs are within 10 miles of the center (Glaeser, Kahn, and Chu, 2001). The median location (where half of jobs are closer and half are farther away) is seven miles. Looking across the four regions of the country (Northeast, Midwest, South, West), the 10-mile shares are remarkably close, with all four lying in the range of 64 percent to 67 percent. The three-mile shares are similar for all but the Northeast region, whose share of 29 percent lies well outside the 19 to 21 percent range of the other regions.

As shown in Table 7–2 there is substantial variation in the 3-mile and 10-mile employment shares across U.S. metropolitan areas. The 3-mile share ranges from 8 percent for Los Angeles to 40 percent for Boston. The 10-mile share ranges from 28 percent for Los Angeles to 79 percent for Indianapolis. In terms of the actual number of jobs within three miles of the center, the metropolitan areas with largest concentrations of central employment are not listed in the table: New York has over 1.4 million jobs within three miles of the center, and Chicago has 530,000.

TABLE 7–2 Employment within Three Miles and 10 Miles of the City Center: Selected Cities

	Indianapolis	Portland	Boston	Minneapolis	Atlanta	Los Angeles
Total Employment	635,818	762,677	1,152,387	1,294,873	1,604,716	4,680,802
Jobs within three miles of center	179,893	235,057	459,936	267,798	221,986	382,465
Percent within three miles of center	28	31	40	21	14	8
Percent within 10 miles of center	79	76	76	64	43	28

MAP 7–1 The Spatial Distribution of Employment: Portland

Each jigsaw-piece census tract is extruded to a height equal to its employment density, defined as the number of workers per hectare. The ribbons show freeways, extruded to a height of 25. Employment density reaches its maximum of 539 per hectare in the central business district.

A Closer Look at the Spatial Distribution of Jobs: Portland and Boston

Map 7–1 shows the distribution of employment within the Portland metropolitan area, with a bird's-eye view from the southwest. Each census tract is shown as a jigsaw puzzle piece and is extruded (pushed up) to a height equal to its employment density, defined as the number of workers per hectare. The ribbons in the map depict the freeways that run through the metropolitan area, extruded to a value of 25. Employment density reaches its highest level in the central area, with a top density of 539 workers per hectare (137,890 workers per square mile). Outside the central area, employment density is highest near the freeways, reaching 25 (the height of the freeway ribbons) at several places.

How are jobs in the Portland area divided between areas of different job densities? As shown in the first column of Table 7–3, almost one-fifth of the jobs are in census tracts with high density—at least 50 workers per hectare. Urban economists define an employment subcenter as an area with a job density of at least 25 workers per hectare and total employment of at least 10,000 jobs. In Map 7–1, several areas of the city have job density exceeding 25 worker per hectare, including areas in the

TABLE 7–3 Distribution of Jobs across Density Classes: Selected Cities

	Portland	Boston	Indianapolis	Minneapolis	Atlanta	Los Angeles
Total Employment	762,677	1,152,387	635,818	1,294,873	1,604,716	4,680,802
Percent of Jobs						
High Density (D > 50)	18	37	18	17	10	17
Medium Density (25 < D < 50)	8	10	3	12	8	22
Low Density (12.5 < D < 25)	16	22	14	24	12	28
Very Low Density (D < 12.5)	57	31	65	47	70	32

southwest (Beaverton, OR) and the north (Vancouver, WA). As shown in the first column of Table 7–3, roughly 8 percent of jobs in the metropolitan area have a job density between 25 and 50 workers per hectare. The rest of the jobs are in areas with low or very low density: 16 percent are in census tracts with between 12.5 and 25 workers per hectare, and 57 percent are in census tracts with fewer than 12.5 workers per hectare.

Map 7–2 shows the density of employment within the Boston metropolitan area. The upper panel shows a full view. Employment density reaches its maximum at 1,953 workers per hectare (500,001 workers per square mile) near the center. The lower panel provides a close-up view and cuts off the map at an employment density of 976: The actual density in the center is roughly twice as high as indicated by the map. As in the other density maps, the freeways are shown as ribbons extruded to a value of 25 and help identify areas where job density exceeds the threshold for an employment subcenter.

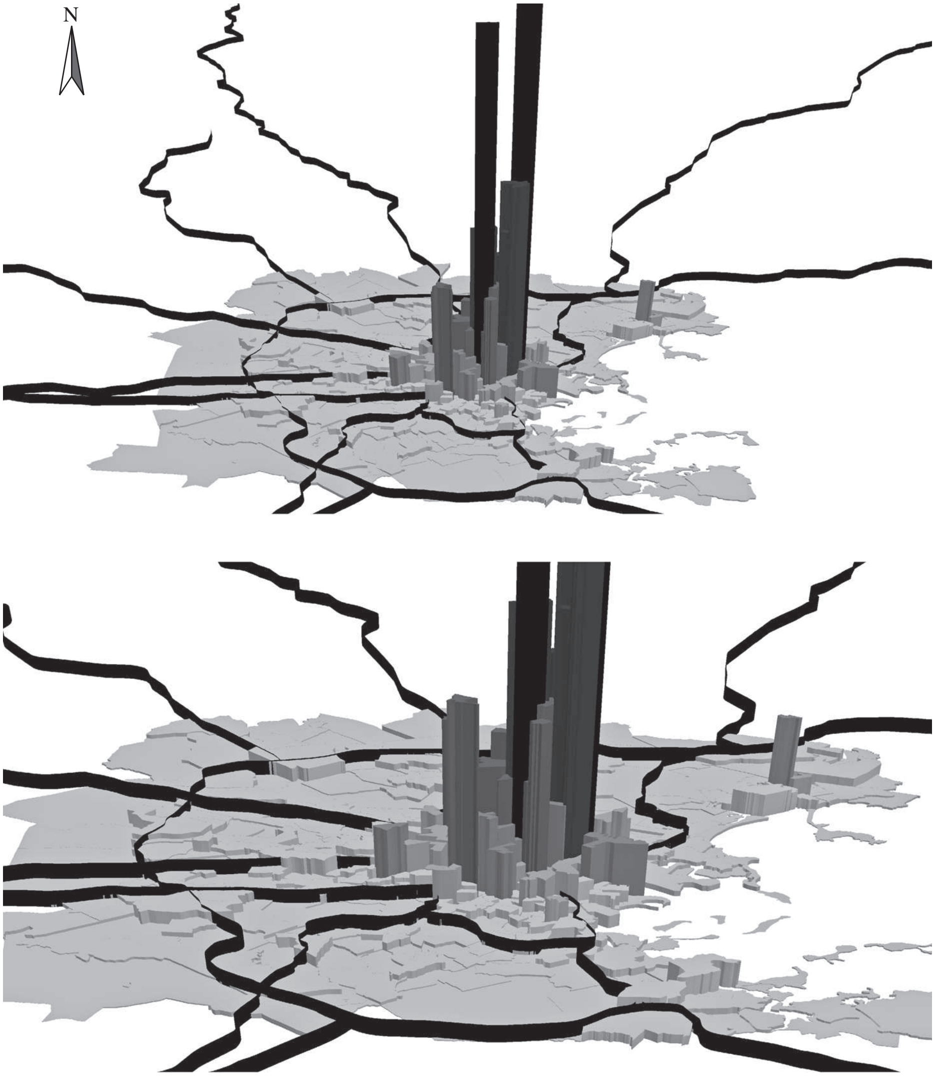
Table 7–3 shows the distribution of jobs in Boston and other metropolitan areas across different density classes. Compared to Portland, Boston has roughly twice the share of jobs in the high-density class and a much lower share in the lowest density classes. Los Angeles has roughly the same share of jobs in the high-density class as Portland, Indianapolis, and Minneapolis, but a much higher share in the medium class, so the share in the lowest two classes is relatively small. Atlanta stands out with its small share in the high-density class and its large share in the lowest class. In Indianapolis and Atlanta, roughly four out of five jobs are in areas with density less than 25 workers per hectare.

Employment Subcenters: Los Angeles and Chicago

As we saw in Chapter 3, agglomeration economies encourage firms to locate close to one another, and one manifestation of clustering is a subcenter. For example, manufacturing firms in a cluster can purchase maintenance and repair services from a common supplier, and office firms in a cluster can purchase glossy brochures from the same printing firm. Firms in office clusters also share restaurants and hotels. One rule of thumb is that a cluster of 2.5 million square feet of office space can support a 250-room hotel.

Giuliano and Small (1991) explore the spatial distribution of employment in the Los Angeles metropolitan area. The CBD had total employment of 496,000 (11 percent of total employment), with a density of 90 workers per hectare. They define

MAP 7–2 The Spatial Distribution of Employment: Boston



Each census tract is extruded to a height equal to its employment density, defined as the number of workers per hectare. The ribbons show freeways, extruded to a height of 25. Employment density reaches its maximum of 1,953 per hectare in the central business district. The top panel is a full view, while the bottom panel is a close-up view, cut off at a vertical value of 976 workers per hectare.

a subcenter as an area where employment density is at least 25 workers per hectare and total employment is at least 10,000 workers. In 1990, there were 28 subcenters, with an average employment density of 45 workers per hectare. Together the subcenters contained 23 percent of the metropolitan employment, leaving about two-thirds of employment as dispersed—outside the center and the subcenters. Giuliano and Small divide the 28 subcenters in Los Angeles into five types, according to the products produced:

- **Mixed-industrial subcenters** started out as low-density manufacturing areas near a transport node (airport, port, or marina) and grew as they attracted other activities.
- **Mixed-service subcenters,** like traditional downtowns, provide a wide range of services, and many functioned as independent centers before they were absorbed into the metropolitan economy.
- **Specialized-manufacturing subcenters** include old manufacturing areas as well as newer areas near airports that produce aerospace equipment.
- **Service-oriented subcenters** employ workers in service activities such as medical care, entertainment, and education.
- **Specialized entertainment subcenters** employ workers in television and film.

McMillen and McDonald (1998) explore the spatial distribution of employment in the Chicago metropolitan area. Using the same subcenter definition as Guiliano and Small, they identify 20 subcenters, including nine old industrial areas, three old satellite cities, two subcenters that mix new industry and retailing, and three that mix services and retailing.

The Spatial Distribution of Office Employment and Office Space

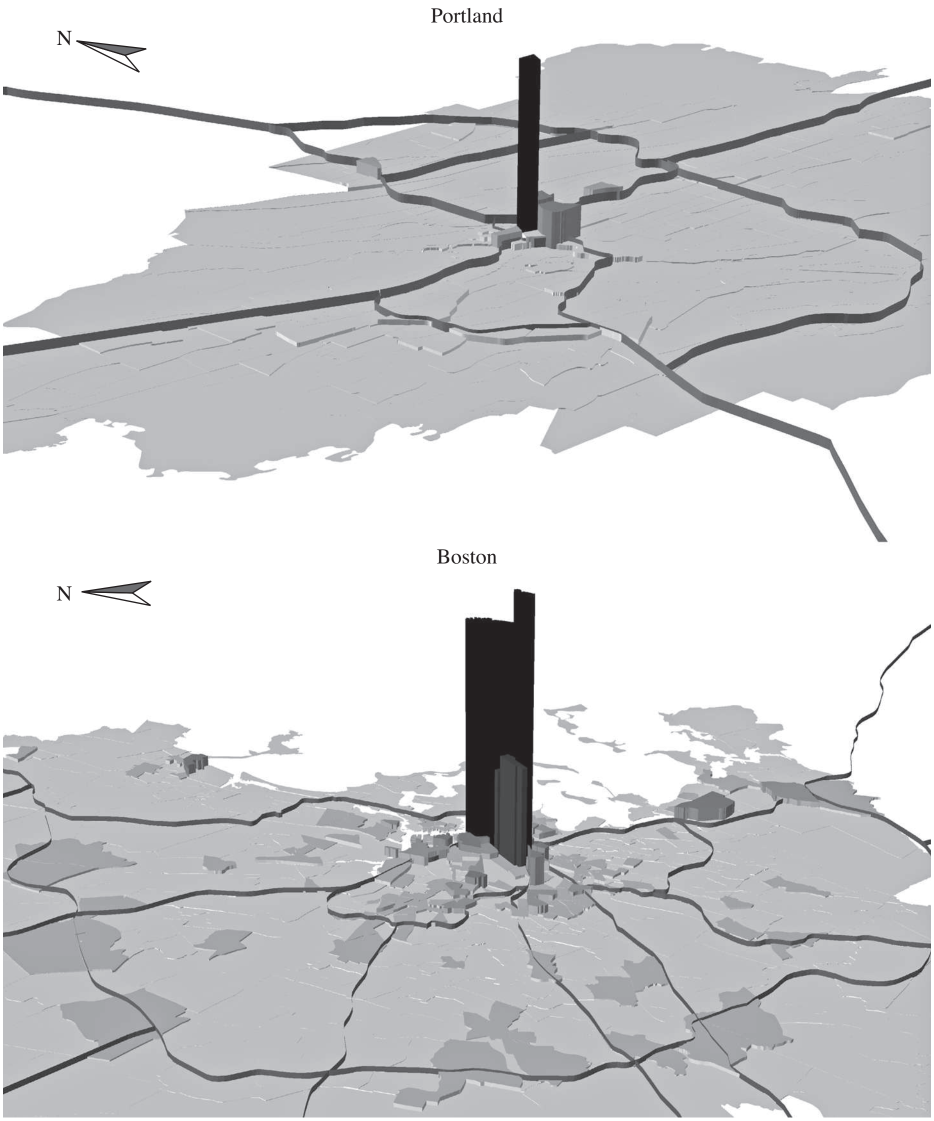
As explained in the previous chapter, activities that require face-to-face contact between workers from different firms benefit from clustering near the city center. Map 7–3 shows the spatial distribution of employment in FIRE (finance, insurance, real estate) in Portland and Boston. Census tracts are extruded to the value of employment density (workers per hectare) and the freeway ribbons are extruded to a value of 12.5 (half the value for freeway extrusion in the earlier maps). Both metropolitan areas have large concentrations of office employment in the central business district, and smaller clusters along the freeways.

Consider next the spatial distribution of office space. An office subcenter is defined as an area with at least 5 million square feet of office space in a relatively compact area. A recent study of 13 large U.S. metropolitan areas revealed a total of 81 office subcenters (Lang, 2003). For the 13 cities as a whole, 38 percent of office space is located in the CBD; 26 percent is in subcenters; and 36 percent is in dispersed locations. Table 7–4 (page 168) shows the distribution of office space across CBDs, subcenters, and elsewhere (dispersed). For three metropolitan areas—New York, Chicago, and Boston—over half of office space is in the CBD. But in each of the other metropolitan areas, the dispersed share exceeds the CBD share.

Edge Cities

Garreau (1991) introduced the notion of edge cities, relatively new concentrations of office and retail space. More recent work focuses on total office space as the

MAP 7–3 Distribution of Office Employment: Portland and Boston



Each census tract is extruded to a height equal to its office-employment density, defined as the number of workers in FIRE (finance, insurance, real estate) per hectare. The ribbons show freeways, extruded to a height of 12.5.

defining characteristic of an edge city, using the threshold of 5 million square feet (Lang, 2003). Edge cities are distinguished from other subcenters by their more recent development. The last two columns of Table 7–4 show, for selected metropolitan areas, the number of edge cities and their shares of metropolitan office space.

TABLE 7–4 Distribution of Office Space, Selected Metropolitan Areas

	Percent of Office Space in CBD	Percent of Office Space in Subcenters	Percent of Office Space Dispersed	Number of Edge Cities	Percent of Office Space in Edge Cities
Atlanta	24	35	41	2	25
Boston	57	23	32	4	19
Chicago	54	20	27	6	20
Dallas	21	45	35	6	40
Denver	30	34	36	4	29
Detroit	21	40	39	2	40
Houston	23	38	39	6	38
Los Angeles	30	33	37	6	25
Miami	13	21	66	2	17
New York	57	13	30	6	6
Philadelphia	34	12	54	2	9
San Francisco	34	23	43	4	14
Washington, DC	29	40	32	8	27

Source: Computations based on Robert E. Lang, *Edgeless Cities* (Washington DC: Brookings, 2003).

The Role of Subcenters in the Metropolitan Economy

Based on recent studies, we can draw several conclusions about the nature and role of subcenters in the metropolitan economy (Anas, Arnott, and Small, 1998; Sivitanidou, 1996, McMillen, 1996; Schwartz, 1992).

1. Subcenters are numerous in both new and old large metropolitan areas.
2. In most metropolitan areas, most jobs are dispersed rather than concentrated in CBDs and subcenters.
3. Many subcenters are highly specialized, indicating the presence of large localization economies.
4. Subcenters have not eliminated the importance of the main center. In 7 of the 13 cities studied by Lang, the ratio of CBD office space to the space in the largest subcenter is at least 4. The metropolitan areas with the largest CBD subcenter ratios are New York (32) and Chicago (12). Detroit is the only metropolitan area with a ratio less than 1.
5. In the typical metropolitan area, employment density (jobs per hectare) decreases as distance from the center increases, even though the center contains a relatively small share of total employment.
6. Firms in subcenters interact with the center, and the value of access to firms in the center is reflected in higher land prices near the center.
7. Firms in different subcenters interact, indicating that subcenters have different functions and are complementary.

What is the economic relationship between a central business district and the surrounding subcenters and dispersed firms? The central business district provides better opportunities for the face time required for the production of services such as advertising, accounting, legal counsel, and investment banking. Although advances in telecommunications have reduced the need for some types of interaction, face time is still required to exchange complex and tacit information as well as to establish trust.

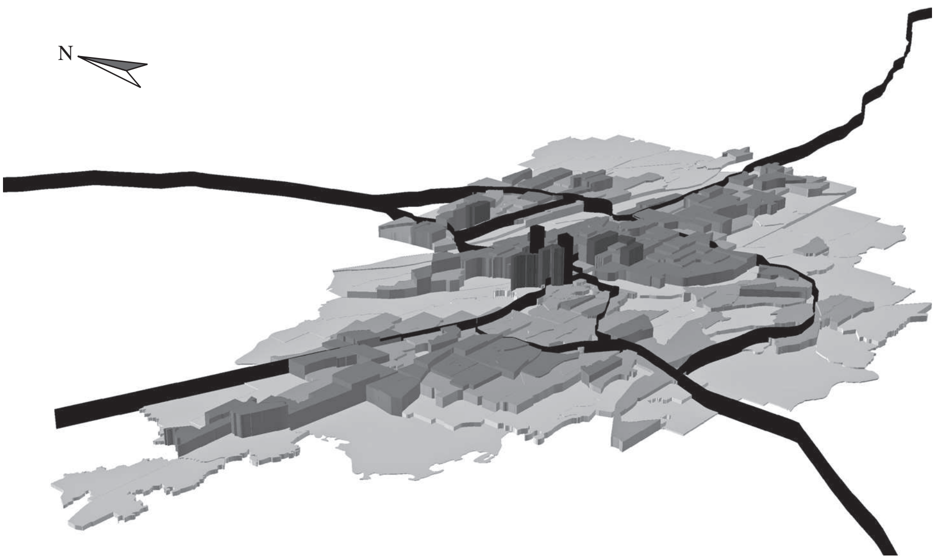
THE SPATIAL DISTRIBUTION OF POPULATION

Consider next the spatial distribution of population within metropolitan areas. For U.S. metropolitan areas as a whole, 36 percent of people live in central cities and the remaining 64 percent live in other municipalities. For the largest 100 metropolitan areas, 20 percent of people live within three miles of the center, and 65 percent live within 10 miles of the center (Glaeser, Kahn, and Chu, 2001). The median residential location is eight miles from the city center, one mile beyond the median location for employment. In other words, the urban population is a bit more decentralized than urban employment.

Population Density in Portland and Boston

Maps 7–4 and 7–5 show the density of population within two metropolitan areas, Portland and Boston. Each census tract is shown as a jigsaw piece, extruded to a height equal to its population density, defined as the number of residents per hectare. As usual, the freeways that run through the metropolitan areas are shown as

MAP 7–4 Population Density: Portland



The ribbons show freeways, extruded to a height of 25. Each census tract is extruded to a height equal to population density (people per hectare). Population density reaches its maximum of 91 per hectare in the central area.

MAP 7–5 Population Density: Boston

The ribbons show freeways, extruded to a height of 25. Each census tract is extruded to a height equal to population density (people per hectare). Population density reaches its maximum of 420 people per hectare in the central area.

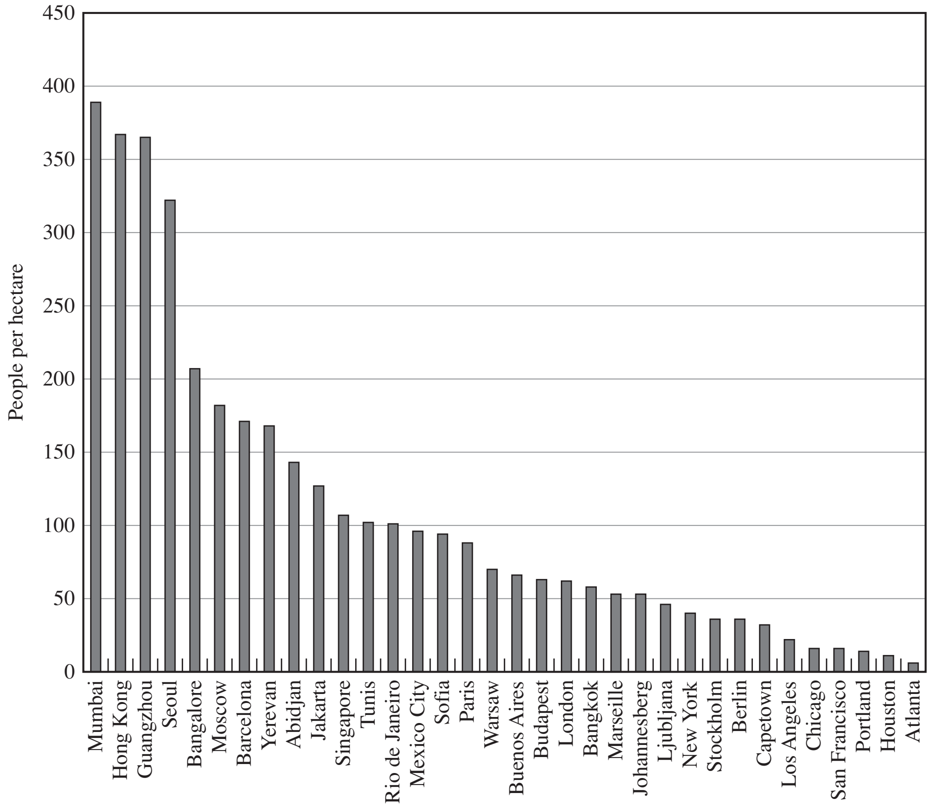
ribbons, extruded to a value of 25. In Portland, the population density reaches its highest level in the central area, with a top density of 91 per hectare (23,376 per square mile). In Boston, the population density reaches its maximum at 420 per hectare (107,760 per square mile) near the center.

As we saw in the previous chapter, the price of land generally decreases as we move away from the city center, reflecting the superior accessibility of central locations. As a result, people living close to the center economize on land, leading to higher population density. Map 7–5 shows that in Boston, population density decreases as the distance to the center increases. In Paris, the population density in the central area is roughly six times the density at a distance of 20 kilometers. In New York, population density near the center is roughly four times the density at a distance of 20 kilometers. The density gradient is defined as percentage change in density per additional mile from the city center. For example, the density gradient is 0.13 in Boston: Population density decreases by 13 percent for each additional mile from the city center. For most large metropolitan areas in the United States, the density gradient is in the range 0.05 to 0.15.

Density in World Cities

Figure 7–1 shows the urban density in selected cities around the world. In this figure, urban density is the total population of a metropolitan area divided by the amount of land in urban use, including residential areas, industrial districts, commercial areas, roads, schools, and city parks. This is called built-up density, as opposed to residential density. Asian cities are at the top of the density list, and U.S. cities are at the bottom. New York is the densest U.S. metropolitan area, yet its density is about half the density of Paris, one-fourth the density of Barcelona, and one-tenth the density of Mumbai (formerly known as Bombay). Los Angeles, the second densest U.S. metropolitan area, is roughly half as dense as New York. All the European cities shown rank above Los Angeles in density, and most rank above New York.

FIGURE 7–1 Population Density in World Cities



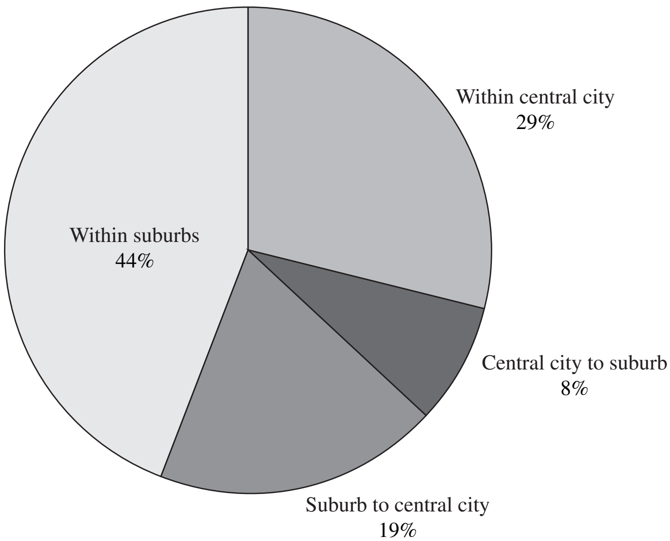
Based on: Alain, and Stephen Malpezzi. "The Spatial Distribution of Population in 48 World Cities: Implications for Economies in Transition." Working Paper, Center for Urban Land Economics Research, University of Wisconsin, 2003.

Commuting Patterns

We've seen that the majority of jobs are outside central cites, and so is the majority of the population. It should be no surprise that the most frequent commuting trip is between two municipalities outside the central city. As shown in Figure 7–2, about 44 percent of commuter trips occur outside the central city, while 29 percent are within central cities. For commuting between central cities and other municipalities, 19 percent of commuters live outside the central city and commute to the central city, and 8 percent do the reverse commute.

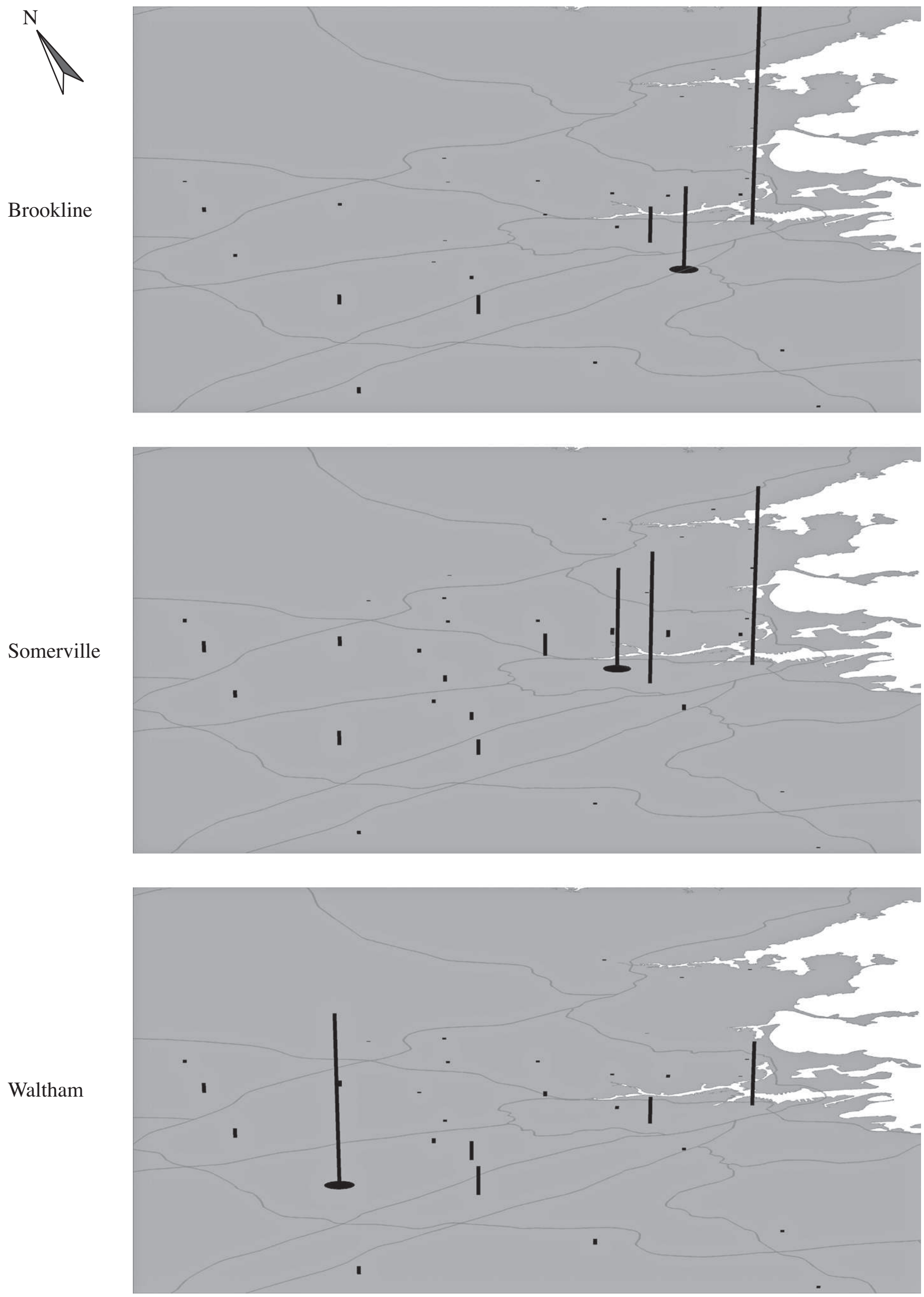
Map 7–6 shows some examples of commuting in the Boston metropolitan area. Each panel shows the number of commuters from a particular municipality (marked with a disk) to other municipalities in the metropolitan area. The top panel shows commuting from Brookline: The tallest bar shows the number of workers who commute to Boston (50 percent of Brookline commuters), and the second tallest bar shows the number of workers who commute within Brookline (19 percent). The shorter bars show the numbers of commuters to other municipalities. The middle panel shows commuting from Somerville. Roughly 30 percent of Somerville commuters travel to jobs in the city of Boston, while 22 percent commute to a nearby municipality and 17 percent work within Somerville. As shown in the lower panel, roughly 39 percent of Waltham commuters travel to jobs within the municipality, and 16 percent commute to jobs in the city of Boston.

FIGURE 7–2 Metropolitan Commuting Patterns, 2000



Source: U.S. Census, Journey to Work.

MAP 7–6 Commuting Patterns in the Boston Metropolitan Area



THE RISE OF THE MONOCENTRIC CITY

Cities looked very different just 100 years ago. At the start of the 20th century, jobs were concentrated near the city center. Manufacturing firms located close to railroad terminals and ports to economize on the cost of transporting inputs and outputs within the city. Office firms clustered in the CBD to facilitate the rapid exchange of information. Workers either lived near the central city and commuted by foot or rode streetcars from suburbs to the city center.

Before exploring the reasons for the demise of the monocentric city, we will explore why it arose in the first place. Recall the fourth axiom of urban economics:

Production is subject to economies of scale

As we saw in Chapter 2, the Industrial Revolution of the 19th century generated innovations in production and energy that increased the scale of production. Firms used indivisible inputs and specialized labor to produce on a large scale, and they located in cities to exploit agglomeration economies. The Industrial Revolution also generated innovations in intercity transportation that allowed the wider exploitation of comparative advantage, leading to increased trade and larger trading cities.

Innovations in Intracity Transportation

The Industrial Revolution also generated a series of innovations in intracity transportation that decreased commuting costs. Before the 1820s, most urban travel was by foot, although a few wealthy people traveled by private horse-drawn carriage. Beginning in the late 1820s, innovations in transportation included the following:

- **Omnibus (1827).** The omnibus, a horse-drawn wagon introduced in New York, was the first public transit mode. The name means "for all" (French via Latin), indicating the public nature of the mode, and is the predecessor of "bus." When the horse wagons were put on rails, the travel speed increased to six miles per hour.
- **Cable cars (1873).** Steam-powered cable cars were introduced in San Francisco and spread to other cities.
- **Electric trolley (1886).** The trolley was powered by an on-board electric motor connected to overhead power lines with dangling wires that apparently reminded harried city dwellers of fishing (trolling). Chicago built its elevated trolley line (with a dedicated throughway) in 1895.
- **Subways (1895).** The world's first practical subway started operation in London in 1890 with electric traction replacing smoky steam power. Boston built the first U.S. subway, a 1.5-mile line that used streetcars. It was followed by systems in New York (1904) and Philadelphia (1907).

These innovations decreased commuting costs and increased the feasible radius of cities. One rule of thumb is that the radius of a city is the distance that can be traveled in an hour. In the "walking city" of the early 19th century, the maximum radius was about two miles. The series of innovations in intracity transport increased travel speeds and increased the feasible radius of cities.

The design of the public transit systems of the 19th century facilitated the large concentrations of employment near city centers. These were hub-and-spoke systems, with spokes radiating out from a central hub. They were designed to transport workers and shoppers from suburban areas along the spokes to the city center.

The Technology of Building Construction

Another limit to city size comes from the costs of building high-density housing to accommodate workers. In the early 1800s, wood buildings were made of posts and beams, with 16-inch timbers, and the practical height limit was three floors. The construction of a three-story building required highly skilled labor to fasten the posts and beams, so urban buildings were relatively expensive. Masonry buildings could be a bit taller but were inflexible because every wall was a load-bearing wall.

The balloon-frame building, introduced in 1832, used smaller pieces of lumber, fastened by nails using less skilled labor. The first balloon-frame building was a warehouse in Chicago. A critical element in the spread of the balloon-frame building was the introduction of inexpensive manufactured nails. Before they were introduced in the 1830s, handcrafted nails were expensive enough to be listed as valued possessions in wills (Bartlett, 1998). The combination of the balloon frame and manufactured nails decreased the cost of urban buildings significantly, contributing to the growth of monocentric cities.

Office buildings were transformed by the switch from masonry to steel frames. In 1848, a five-story building in New York used cast-iron columns instead of masonry walls. The switch to steel followed, providing framing material that was stronger and more elastic and workable than cast iron. The world's first skyscraper, an 11-story building housing the Home Insurance Company, was built in 1885 with a steel skeleton frame.

One limit on building heights is the cost of vertical transportation. The burden of walking up stairs imposed a practical height limit on buildings. In 1854, Elisha Otis demonstrated the safe use of a steam-powered elevator. The key innovation was a safety latch that prevented the elevator car from plummeting down when the rope connecting the car to the pulley system broke. By 1857, the Otis elevator was being used in a five-story building. When the dedicated steam engine was replaced by electricity, the cost of running elevators decreased and their range increased. In the world's first skyscraper, a bank of elevators carried people up and down at a speed of 500 feet per minute.

The elevator changed the pricing of space on different floors of an office building. Recall the first axiom of urban economics:

Prices adjust to achieve locational equilibrium

In a tall office building, locational equilibrium requires firms to be indifferent between different floors. Before the elevator, upper floors were rented at a discount to offset the cost of climbing four or five flights of stairs. The elevator reduced the cost of vertical travel, and the upper floors became more desirable. Upper floors, with their better views, rented at a premium rather than a discount (Bartlett, 1998).

The Primitive Technology of Freight

Consider finally the technology of freight transportation in the 19th century. As we saw earlier in the book, most intercity freight traveled by railroad or water (on rivers or oceans). For transportation within the city, manufacturers used horse-drawn wagons to transport their freight from factories to the city's port or railroad terminal. This, of course, was the most primitive and costly part of the transportation system, and it tied manufacturers to the central export node—a railroad terminal or port.

THE DEMISE OF THE MONOCENTRIC CITY

What caused the demise of the monocentric city, with its large concentration of employment in the central area? We'll discuss the decentralization of manufacturing, office activity, and population.

Decentralization of Manufacturing: Trucks and Highways

The share of metropolitan manufacturing employment in central cities decreased from about two-thirds in 1948 to less than half in 2000. Mills (1972) provides evidence that the decentralization of manufacturing started long before 1948. What caused the suburbanization of manufacturing employment?

Moses and Williamson (1972) explain the role of the intracity truck in the suburbanization of manufacturing. The truck was developed in 1910, providing an alternative to the horse-drawn wagon used for the trip from factory to port or rail terminal. The truck was twice as fast as a horse-drawn wagon and half as costly, with a unit freight cost of only $0.15 per ton per mile. Between 1910 and 1920, the number of trucks in Chicago increased from 800 to 23,000.

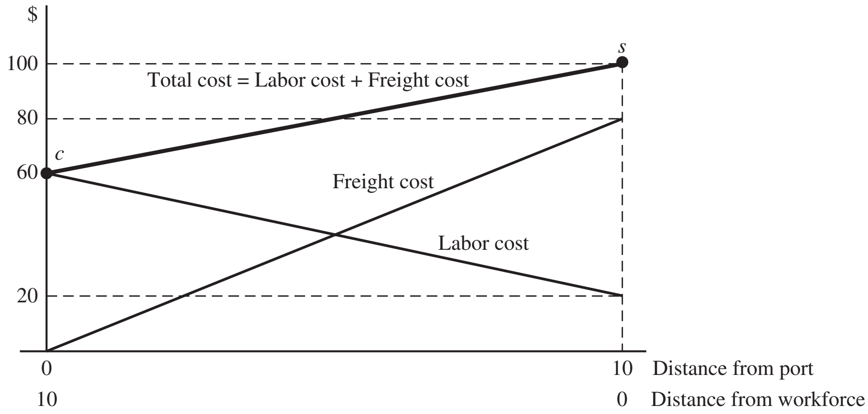
Consider the trade-offs faced by a manufacturing firm that moves away from a central port to a suburban location:

- **Higher freight cost.** The cost of transporting output to the port increases.
- **Lower wages.** As the factory moves closer to its workforce, commuting time decreases, decreasing wages.

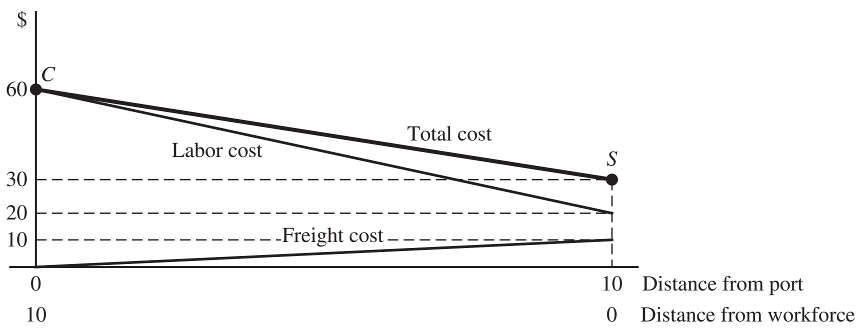
In the era of the horse-drawn wagon and the streetcar, the cost of moving freight was high relative to the cost of moving workers, so as a firm moved away from the city center, freight costs increased more rapidly than wages dropped. It was cheaper to move the workers from the suburbs to the central factory than to move output from a suburban factory to the export node.

Figure 7–3 illustrates the influence of intracity freight technology on the location choices of manufacturers. Consider a firm that transports its output from its factory to a port at the city center, using a horse-drawn wagon. As shown in the upper

FIGURE 7–3 The Truck and the Suburbanization of Manufacturing



Total transport cost (labor cost + freight cost) is minimized at the city center (the port) because the cost of moving output (on horse-drawn wagons) is high relative to the cost of moving workers (on streetcars).



Total transport cost (labor cost + freight cost) is minimized at suburb (where workers live) because the cost of moving output (on trucks) is low relative to the cost of moving workers (on streetcars).

panel of Figure 7–3, the firm's daily freight cost of \$8 per mile, increases from zero for a factory located next to the port to \$80 for a factory in a suburb 10 miles from the port. The firm's workers live in the suburb and commute by streetcar to the factory. The wage demanded by workers increases with commuting distance, and the firm's daily labor cost increases from \$20 in a factory in the suburb to \$60 in a factory in the center. The firm's total cost, equal to freight cost plus labor cost, is \$60 at the city center (point *c*), compared to \$100 in the suburb (point *s*). Starting from a central location, a one-mile move toward the suburb decreases labor cost by \$4 but increases freight cost by \$8, so total cost increases by \$4. When the firm moves the entire 10 miles from center to suburb, the firm's total cost increases from \$60 to \$100. Total cost is minimized at the center because the cost of moving output on horse-drawn wagons is large relative to the cost of moving workers on streetcars.

The lower panel of Figure 7–3 shows the effects of replacing the horse-drawn wagon with the intracity truck. In our example, the firm's freight cost decreases from \$8 per mile to \$1 per mile, decreasing the slope of the freight-cost curve. In contrast, the labor-cost curve is unchanged. Starting from a central location (point *C*), a one-mile move toward the suburb decreases labor cost by \$4 but increases freight cost by only \$1, so total cost decreases by \$3. Moving the entire 10 miles from center to suburb causes the total cost to decrease from \$60 to \$30. Total cost is minimized at the suburban location (point *S*) because the cost of moving output on trucks is low relative to the cost of moving workers on streetcars. The truck allowed manufacturers to benefit from lower wages in the suburbs without a large penalty in freight cost, so many firms moved to the suburbs.

Two decades after the truck was first introduced, manufacturers started using it for intercity transport. Improvements in the truck made long-distance travel feasible, and the expansion of the intercity highway system facilitated intercity truck traffic. Eventually, the truck became competitive with the train and the ship for intercity freight. As manufacturers switched from trains and ships to trucks, they were freed from their dependence on the railheads and ports in city centers, and they moved to sites accessible to the intercity highways. In modern cities, manufacturers locate close to highways and beltways to get easy access to the interstate system.

Map 7–7 shows the spatial distribution of manufacturing in two metropolitan areas, Chicago and Los Angeles. Census tracts are extruded to a value equal to manufacturing employment per hectare. Freeways are shown as ribbons, extruded to a value of four workers per hectare. In both metropolitan areas, the bulk of manufacturing employment is in suburban areas.

Other Factors: Automobile, Single-Story Plants, and Airports

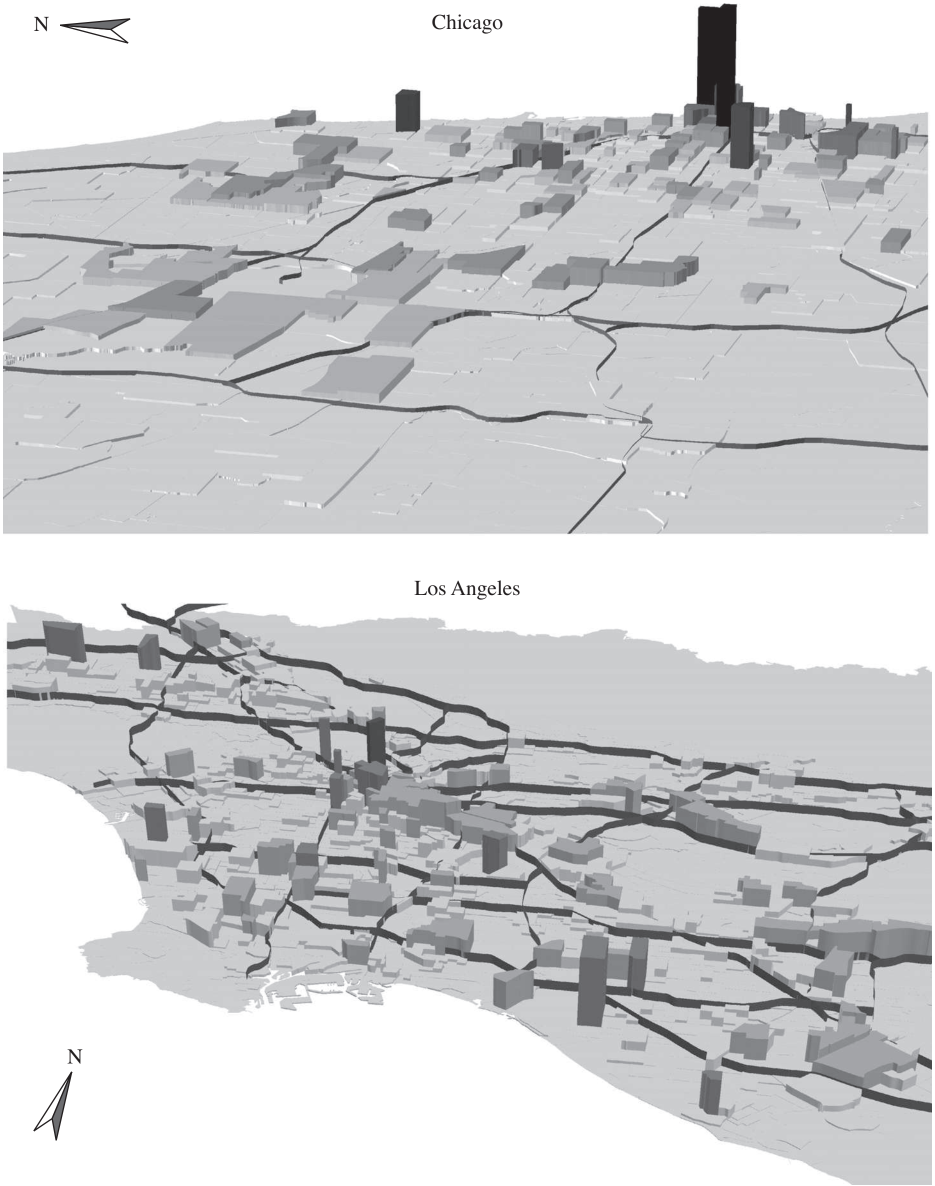
The automobile contributed to the suburbanization of manufacturers. In a streetcar city, a firm that drew workers from locations throughout the metropolitan area located near the hub of the streetcar system to be accessible to the metropolitan workforce. In a modern auto-based city, production sites along highways and beltways are accessible to the metropolitan workforce, so firms have more location options, including suburban locations.

Two other factors contributed to the suburbanization of manufacturing. First, the switch from the traditional multistory plants of the 19th century to modern single-story plants increased the relative attractiveness of suburbs, where land prices are lower. Second, an increase in the importance of air freight caused firms to locate near suburban airports. For some firms, the suburban airport has replaced the port as the point of orientation.

Decentralization of Office Employment

Before the early 1970s, most office firms located in the CBD because the central location facilitated face time with other office firms. There was some suburban office activity, but most of it involved back-office operations—paper processing

MAP 7–7 Distribution of Manufacturing Employment: Chicago and Los Angeles



Each census tract is extruded to a height equal to the density of manufacturing employment defined as the number of workers per hectare. The ribbons show freeways, extruded to a height of four workers per hectare.

rather than information exchange. For most office activities, the advantages of a CBD location (timely contact with other firms) outweighed the disadvantages (high wages and rents).

In the last 30 years, advances in communications technology have allowed more office activities to be performed outside CBDs. The electronic transmission of information allows workers to exchange a wider variety of information without face time. Firms can decouple their operations, with information processing in the suburbs and activities requiring face time in the CBD. For example, a firm's accountants can locate in the suburbs and transmit reports electronically to executives in the CBD, who then use the reports in their interactions with other firms.

Decentralization of Population

We can use the population density gradient, defined as the rate at which population density decreases with distance, to document the suburbanization of population. A smaller gradient indicates that density decreases less rapidly with distance and population is less centralized. As Mills (1972) shows, the U.S. density gradients have been decreasing for the last 120 years. The average gradient for four cities (Baltimore, Milwaukee, Philadelphia, Rochester) was 1.22 in 1880, indicating that 88 percent of the population lived within three miles of the center. By 1948, the gradient had dropped to 0.50 (a three-mile population share of 44 percent), and by 1963, the gradient had dropped to 0.31 (a three-mile share of 24 percent).

The decentralization of metropolitan population is a worldwide phenomenon (Anas, Arnott, and Small, 1998). Between 1801 and 1961, London's density gradient decreased from 1.26 to 0.34, meaning that the percentage of its population living within three miles of the city center dropped from 88 percent to 28 percent. In Paris, the gradient decreased from 2.35 in 1817 to 0.34 in 1946. In cities throughout the world, population has been shifting outward away from the city center.

What factors contributed to the decentralization of population over the last several decades? One factor is rising income. The demand for housing increases with income, and because housing prices are generally lower in suburban areas, rising income increases the relative attractiveness of suburban locations. Of course, an increase in income also increases the opportunity cost of commuting, increasing the relative attractiveness of locations close to workplaces. So it is not clear, in theory, whether higher income leads to more distant residential locations. There is evidence that income growth encourages suburbanization (Anas, Arnott, and Small, 1998).

Another factor in suburbanization of population is lower commuting costs. As we saw earlier in the chapter, technological innovations over the last 180 years, from the omnibus of 1827 to the fast and comfortable automobiles of today, have decreased the monetary and time costs of commuting. A decrease in commuting costs decreases the relative cost of living far from the city center, contributing to suburbanization. In addition, the suburbanization of jobs and people reinforces one another: Jobs follow workers to the suburbs, and workers follow jobs to the suburbs.

Several other factors contribute to the suburbanization of population:

1. **Old housing.** The deterioration of central-city housing encourages households to move to the suburbs, where most of the new housing is built.
2. **Central-city fiscal problems.** Many central cities have relatively high taxes, encouraging households to move to low-tax suburbs. The causality goes both ways: Fiscal problems cause suburbanization, and suburbanization contributes to central-city fiscal problems.
3. **Crime.** Most central cities have relatively high crime rates, encouraging households to move to the suburbs. Later in the book, we'll explore the reasons for higher crime rates in central cities.
4. **Education.** Suburban schools are often considered superior to central-city schools, encouraging households to relocate to the suburbs. Later in the book we'll explore the reasons for differences between central-city and suburban schools.

Empirical studies of the suburbanization process provide support for these other factors in suburbanization. Bradbury, Downs, and Small (1982) show relatively rapid suburbanization in metropolitan areas in which the central city had (1) a relatively old housing stock, (2) relatively high taxes, (3) a relatively large black population, and (4) a large number of suburban governments. Frey (1979) found that suburbanization was relatively rapid in metropolitan areas with high taxes, high crime rates, and low spending on schools. Cullen and Levitt (1999) estimate that for each additional central-city crime, one additional person relocates from the central city to a suburb.

URBAN SPRAWL

There is a spirited debate among economists and policy makers about "urban sprawl." As a city's population increases, the city can grow up by building taller buildings, or it can grow out by occupying more land. The people concerned about urban sprawl suggest that there is too little "up" and too much "out." Between 1950 and 1990, the amount of urbanized land in the United States increased by 245 percent while the urban population increased by only 92 percent. The "footprint" of the typical metropolitan area increased more rapidly than its population, so urban density decreased.

Sprawl Facts

One measure of urban sprawl is the density of economic activity. The lower the density, the larger the land area required to accommodate a given population, and the greater the spread or sprawl of the metropolitan area. We saw earlier in the chapter that U.S. cities are much less dense than cities in the rest of the world, including cities in Europe with similar education and income levels. Metropolitan areas in Germany are four times as dense as U.S. metropolitan areas, and Frankfurt is three times as dense as New York City.

Table 7–5 compares Barcelona to Atlanta. Barcelona is 28 times as dense as Atlanta, and the land used per person is 58 square meters, compared to 1,712 in Atlanta. A trip between the two most distant points in Atlanta is 86 miles, compared to only 23 miles in Barcelona. Reflecting its greater density, Barcelona has more trips by public transit and walking.

Population density varies considerably among U.S. metropolitan areas (Fulton, Pendall, Nguyen, Harrison, 2001). Among the 20 most densely populated areas, the range is 40 people per hectare in New York to 14 per hectare in Santa Barbara. The median value for the top 20 metropolitan areas is 18 people per hectare. In contrast with popular perceptions, 12 of the top 14 metropolitan areas (and 13 of the top 20) are in the West, including eight cities in California. The high density in western cities reflects relatively high land prices. In fact, the two poster-cities for sprawl, Los Angeles (number 2, with 21 people per hectare) and Phoenix (number 11, with 18 per hectare), are more dense than Chicago (number 15, with 15 per hectare) and Boston (number 19, with 14 per hectare).

Over the last few decades, the density of U.S. cities has decreased significantly. One way to convey the change is by computing the elasticity of urbanized land (the percentage change in land in urban use) with respect to the population (the percentage change in the urban population). Table 7–6 shows the elasticities for different regions over the period 1982–1997. In the United States, urbanized land grew 2.76 times faster than the urban population. The largest elasticities—and the largest

TABLE 7–5 Population Density in Atlanta and Barcelona

	Atlanta	Barcelona
Population in 1990 (million)	2.5	2.8
Average density (people per hectare)	6	171
Land per person (square meters)	1,712	58
Maximum distance between two locations (kilometers)	138	37
Percent of trips walking	Less than 1	20
Percent of trips on public transit	4.5	30

Source: Computations based on Alain Bertaud. "The Spatial Organization of Cities: Deliberate Outcome or Unforeseen Consequence?" Working Paper, Institute of Urban and Regional Development, University of California, Berkeley, 2004.

TABLE 7–6 Changes in Urban Land and Population, 1982–1997

Region	Percentage Increase in Urban Land	Percentage Increase in Urban Population	Elasticity of Urban Land with Respect to Urban Population
United States	47	17	2.76
West	49	32	1.53
South	60	22	2.73
Northeast	39	7	5.57
Midwest	32	7	4.57

Source: Computations based on William Fulton, Rolf Pendall, Mai Nguyen, and Alicia Harrison. "Who Sprawls Most? How Growth Patterns Differ across the U.S." *The Brookings Institution Survey Series*, July 2001, pp. 1–23.

decreases in density—occurred in the Northeast and the Midwest. In Chicago between 1970 and 1990, the urbanized land area increased by 46 percent while population grew by only 4 percent. In Cleveland over the same period, urbanized land area increased by 33 percent while the population actually decreased by 8 percent.

The Causes of Sprawl

What causes urban sprawl—low density cities? Living at a low density means consuming a large quantity of land. Land is a normal good, so the higher the income, the larger the consumption of land and the lower the population density. A second factor is a low cost of travel, which allows workers and shoppers to live relatively long distances from jobs, shops, and destinations for social interaction. Distant land is cheaper, so lot sizes are larger and population density is lower. Putting these two factors together, high income makes people demand large lots, and a low travel cost allows them to move to the suburbs where land is relatively cheap. So we get low-density development at distant locations, also known as urban sprawl.

Is there a cultural dimension to urban density and sprawl? Bertaud and Malpezzi (2003) suggest that cultural differences explain some of the dramatic differences in urban density across world cities. Asia has much higher urban density than other continents, much higher than could be explained by other factors such as income. Similarly, the variation in density across other continents could reflect differences in preferences for living space. In U.S. metropolitan areas, the presence of immigrants tends to increase density, suggesting that culture is relevant (Fulton, Pendall, Nguyen, Harrison, 2001).

A number of government policies in the United States encourage low densities in large metropolitan areas.

- **Congestion externalities.** As we discuss later in the book, people who use streets and highways during the peak travel period slow other drivers down, imposing an external cost. This underpricing of urban transportation encourages people to commute relatively long distances from locations far from the city center where the low price of land encourages large lots.
- **Mortgage subsidy.** Interest on housing mortgages is a deductible expense for federal and state income taxes, providing a subsidy for housing that increases housing consumption. Land and housing are complementary goods, so the mortgage subsidy increases lot sizes, decreasing density.
- **Underpricing of fringe infrastructure.** In some metropolitan areas, the infrastructure cost of new development at the urban fringe is not fully borne by developers and their customers. Many states use development fees (impact fees) to impose the cost of fringe development on developers and their customers.
- **Zoning.** Many suburban municipalities use zoning to establish minimum lot sizes. One motivation is to exclude low-income households, whose tax contribution may fall short of the costs they impose on municipal government.

Glaeser and Kahn (2004) argue that sprawl is caused mainly by the automobile and the truck. These two travel modes eliminated the orientation of firms and workers toward the indivisible transportation infrastructure near the city center. The authors show that sprawl is ubiquitous across metropolitan areas with all levels of income, poverty, and government fragmentation, suggesting that something else—the internal combustion engine—is the driving force behind sprawl. The authors suggest that the subsidies for highways and housing are too small to have much of an effect.

European Policies

Why is urban population density higher in European cities? Nivola (1998) discusses various public policies that promote higher urban density in European cities. One factor is a higher cost of personal transportation. Because of high taxes, the price of gasoline in Italy is roughly four times the price in the United States. Another factor is the policy of heavier taxes on consumption rather than income. Sales taxes on cars sold in Europe are much higher than in the United States—nine times higher in the Netherlands and 37 times higher in Denmark.

A number of policies in Europe promote the small neighborhood shops that facilitate high-density urban living. Electricity is more costly in Europe, so it would be very expensive to operate the huge refrigerators and freezers that allow Americans to make infrequent trips to suburban megastores. As a result, most Europeans rely to a greater extent on more frequent trips to neighborhood stores. In addition, many European countries restrict the pricing and location of large retailers, protecting small shops from competition. The result is more neighborhood shops—and higher prices for consumers.

Several other policies in Europe promote higher density living. Large agricultural subsidies allow small farmers on urban fringes to outbid city dwellers for land. In 1995, the subsidy per hectare was $791 in the European Union, compared to $79 in the United States. In Europe, investment in transportation infrastructure favors mass transit rather than highways. Britain and France allocate between 40 percent and 60 percent of their transport investment to mass-transit networks, compared to 17 percent in the United States.

The Consequences of Sprawl

A recent study measures some of the consequences of low-density living in U.S. cities (Kahn, 2000). Compared to the typical central-city household, a suburban household requires 58 percent more land (1,167 square meters versus 739). A suburban household actually consumes about the same amount of energy: Although suburban dwellings are larger, they are newer and more energy-efficient. A suburban household drives about 30 percent more than a central-city household. In general, low density means more travel: The elasticity of vehicle miles traveled with respect to urban density is -0.36, meaning that a 10 percent decrease in density increases vehicle miles by 3.6 percent.

What about air pollution? In the last several decades, urban density decreased and urban travel increased, but urban air quality actually improved. Between 1980 and 1995, the annual number of days exceeding the ozone standard in Los Angeles dropped by 27. Although vehicle miles traveled increased, improvements in emissions technology cut emissions per mile. Since most of the "dirty" pre-1975 cars are now off the road, the opportunity to drive more and still get better air quality is evaporating.

What about greenhouse gases, the gases responsible for rising levels of carbon dioxide in the atmosphere? The volume of greenhouse gases generated from a car is determined by the amount of fuel burned. Every gallon of gasoline emits about 20 pounds of greenhouse gases. Over the period 1983–1990, vehicle miles traveled increased by about 4 percent per year, and the average fuel efficiency didn't change much, so the volume of greenhouse gases increased. Since then, vehicle miles traveled have continued to increase, while average fuel efficiency has decreased—a result of the popularity of SUVs, vans, and pickup trucks—so greenhouse emissions have continued to rise.

As we saw earlier in the chapter, the amount of urban land has increased in the last few decades. Between 1980 and 1990, the total urban land increased from 18.9 million hectares to 22.4 million hectares. Counties that experienced urban growth lost agricultural land, but the effect was relatively small. For the nation's counties as a whole, the elasticity of farmland acreage with respect to population is −0.02: A 10 percent increase in population causes a 0.2 percent decrease in farmland. The elasticity is much larger (−0.20) for a subset of states (Illinois, Indiana, Michigan, North Carolina, and Pennsylvania).

The loss of farmland at the city fringe indicates that the land is more valuable in urban use. As we saw earlier in the chapter, various public policies increase the residential value of fringe land, and the solution is to correct the distortionary policies. There is no evidence that urban sprawl has created a shortage of either agricultural land or agricultural products. If it had, the prices of agricultural products would increase, pulling up the price that farmers are willing to pay for land, allowing them to outbid developers for land on the urban fringe.

Bertaud (2004) discusses the challenges associated with providing mass transit in low-density areas. Mass transit is feasible only if density around bus stops or transit stations is high enough to attract a sufficient number riders. For most people, the maximum walking time to a transit stop is about 10 minutes, so a transit stop can serve households within an 800-meter radius. To support a bus system with an intermediate service level (two buses per hour and 1/2 mile between stops), the population density in the service area must be at least 31 people per hectare. There are two U.S. metropolitan areas with at least 31 people per hectare—New York (40) and Honolulu (31). Of course, density is higher closer to centers and subcenters, and these areas are likely to have high enough density to support mass transit. For example, the density of New York City is 80 people per hectare (compared to 40 for the metropolitan area). Later in the book, we'll explore various issues concerning the provision and pricing of mass transit.

A comparison of Barcelona to Atlanta reveals the transit challenge for U.S. cities (Bertaud, 2004). As shown in Table 7–5, Barcelona is 28 times as dense as Atlanta. In Barcelona, 60 percent of the population lives within 600 meters of a transit station, compared to only 4 percent living within 800 meters of a transit station in Atlanta. To duplicate the accessibility and ridership of the Barcelona system, Atlanta would have to build an additional 3,400 kilometers of metro tracks and 2,800 more stations. In contrast, the Barcelona system has just 99 kilometers of tracks and 136 stations.

Policy Responses to Sprawl?

There are many factors behind urban sprawl. It partly reflects consumer choice—a rational choice of a large lot at the expense of other consumer products. A number of public policies contribute to urban sprawl, and the appropriate response is to eliminate these distortions. Would land-use patterns change by a little or a lot? If the relatively low density in U.S. cities results in large part from high income, low transport costs, and strong preferences for space, eliminating the policy distortions won't change density very much. But if the distortions—from congestion externalities, mortgage subsidies, underpricing of fringe infrastructure, and large-lot zoning—are significant, we would expect larger changes in density.

An alternative approach is to adopt antisprawl policies such as urban growth boundaries and development taxes. We'll discuss the trade-offs associated with these policies later in the book. If an antisprawl policy succeeds in increasing density, what are the benefits and costs?

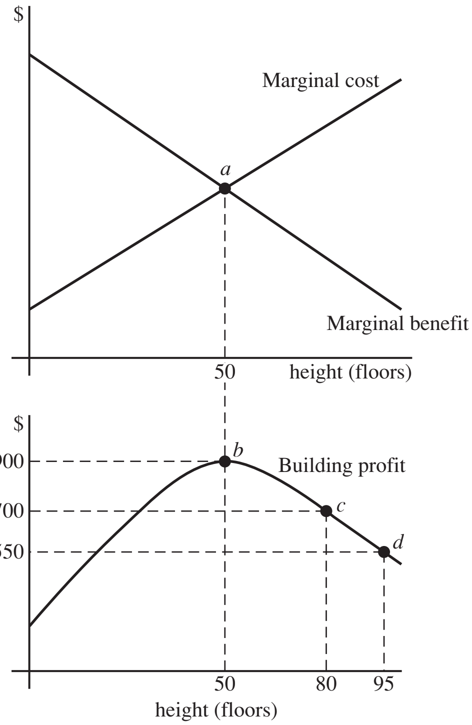
ARE SKYSCRAPERS TOO TALL?

As we saw in the previous chapter, a high price of land generates tall buildings as firms substitute capital for relatively expensive land. Does the high price of land fully explain the massive skyscrapers in modern cities? A recent study suggests that skyscrapers result from competition between firms for the tallest building in a city (Helsley and Strange, 2008). Competition to be the tallest building reduces efficiency and profits.

Consider a single firm that will construct an office building on a given plot of land. The firm will choose a building height, measured in building floors. The firm can use the marginal principle to decide how high to go, choosing the height that makes the marginal benefit equal to the marginal cost.

- The marginal benefit of height equals the rent to be collected from office firms on an additional floor. In the upper panel of Figure 7–4, the marginal-benefit curve is negatively sloped because a taller building devotes more space for vertical transportation (elevators), leaving less rentable space. As building height increases, total rent increases, but at a decreasing rate.
- The marginal cost of height is the additional construction cost from building one more floor. In Figure 7–4, the marginal-cost curve is positively sloped because a taller building requires extra reinforcement to support its more concentrated weight. As building height increases, the total construction cost increases at an increasing rate.

FIGURE 7–4 Are Skyscrapers Too Tall?



Building profit is maximized at the height where marginal benefit equals marginal cost. For a height of 80 floors, building profit is $200 lower than the $900 maximum. If the prize for the tallest is $200, the first firm will win the contest with a building of at least 80 floors.

The firm's profit is maximized at point *a*, where the marginal benefit of height equals the marginal cost. In this example, the profit-maximizing height is 50 floors. The lower panel of Figure 7–4 shows the firm's total profit for different heights. For a particular height, the profit is shown by the area between the marginal-benefit curve and the marginal-cost curve. The total profit reaches its maximum at 50 floors. For a taller building, the marginal benefit of a floor (additional rent) falls short of the marginal cost (additional construction cost), so total profit is lower.

Consider next the implications of competition for being the tallest building in the city. Suppose a second firm has access to the same construction technology, building costs, and rental income potential. Suppose each firm places a value of v on having the tallest building. This could be the value of free corporate advertising from being labeled the tallest building. Alternatively, v could be the value of signaling to potential investors that the firm has enough resources to lavish on a tall building.

Suppose the two firms play a sequential game: firm 1 builds first, followed by firm 2. Before firm 1 decides how tall to build, it must anticipate firm 2's response. If firm 1 is naïve and assumes that firm 2 will choose 50 floors (the height at which the marginal benefit equals the marginal cost) firm 1 could choose 51 floors, expecting to be the tallest and getting the prize of v. The problem is that firm 2 wants to win the prize too, and if firm 1 builds a 51-floor building, firm 2 would choose 52 floors and win the contest. Knowing this, firm 1 might be tempted to build a 53-floor building instead. How high will they go?

Firm 1 must determine how high it must go to prevent its building from being topped by firm 2. If the prize for the tallest is $v = \$200$, firm 2 is willing to sacrifice up to $200 worth of building profit to be the tallest and get the prize. Suppose firm 1 chooses a 79-floor building, meaning that firm 2 would be forced to build at least 80 floors to win the prize. As shown in the lower panel of Figure 7–4, the profit from a 80-floor building is $700, or $200 less than the profit for a 50-floor building. In this case, firm 2 will be indifferent between (*a*) conceding the contest and choosing a 50-floor building, with $900 profit, and (*b*) winning the contest with an 80-floor building, getting $700 in building profit and the $200 prize. If firm 1 chooses an 80-floor building, firm 2 will definitely concede because it would sacrifice more than $200 in building profit to get a $200 prize. To summarize, firm 1 will win the contest if its building has at least 80 floors.

The competition for being the tallest generates a winner with an 80-floor building and a loser with a 50-story building. There is a large gap between the heights of the two tallest buildings, a result consistent with observation in real cities. In the largest 20 U.S. cities, the average gap between the tallest and second tallest is about 27 percent. A second implication is that competition is wasteful because it reduces the total profit of the two firms. When firm 1 chooses an 80-floor building and firm 2 chooses 50 floors, total profit between the two firms is $1,800, equal to the $1,600 building profit ($700 for firm 1 plus $900 for firm 2) plus the $200 prize. One of the firms could still get the $200 prize if the tallest building were 51 floors instead of 80 floors. In that case, total profit would be just under $2000, equal to a building profit just below $1,800 (just under $900 for firm 1 and $900 for firm 2) plus the $200 prize. In other words, competition for the prize dissipates building profit.

SUMMARY

In modern cities, most jobs are dispersed, and most people live and work far from the city center. Over the last 200 years, innovations in transportation have caused the rise of the monocentric city and then its demise. Here are the main points of the chapter.

1. The median job location is seven miles from the city center, and the median residential location is eight miles.
2. Cities in the United States are much less dense than cities in the rest of the world.

3. The key factors in the rise of the large monocentric city were innovations in intraurban transportation that decreased the cost of commuting and innovations in construction that decreased the cost of tall buildings.
4. The key factors in the decentralization of jobs and people were the development of the truck, the automobile, and the highway system; increases in income; and the switch to one-story production facilities.
5. Between 1950 and 1990, the amount of urban land increased more than twice as fast as the urban population.
6. Competition to have the tallest building leads to inefficiently tall buildings.

APPLYING THE CONCEPTS

For exercises that have blanks (_______), fill each blank with a single word or number. For exercises with ellipses (. . .), complete the statement with as many words as necessary. For exercises with words in square brackets ([increase, decrease]), circle one of the words.

1. **Elevator and the Vertical Price of Office Space**
   Draw a curve showing the price of office space (per square foot) within a building, with the horizontal axis measuring the (vertical) distance from the street (in meters) for the following cases.
   a. People walk up and down stairs, and there are no view amenities.
   b. People ride elevators up and down, and there are no view amenities.
   c. People ride elevators up and down, and the view is better on higher floors.
2. **Commuting Methods and Housing Prices**
   The 19th century brought many changes in commuting, from walking to omnibuses, electric trolleys (streetcars), and subways.
   a. For each commuting method, draw a housing-price curve for a monocentric city, labeling each curve with the date on which the method was introduced.
   b. Over time, the feasible radius of the city [increased, decreased] because. . . .
3. **Manufacturing: Labor versus Freight Cost**
   Consider a manufacturing firm that exports its output from a port in the city center and employs workers who live in a suburb six miles from the center. The daily freight cost is $9 per mile. The firm's daily labor cost increases from $20 at the suburb to $50 at the center.
   a. Using Figure 7–3 as a model, show the firm's labor cost, freight cost, and total cost for locations from the center to the suburb. Total cost is minimized at the [city center, suburb].
   b. Suppose the intracity truck is introduced, decreasing the daily freight cost. The firm will be indifferent between the city center and the suburb if the daily freight cost is _______ per mile. Illustrate with a graph. In this case, total cost in the suburb is _______, computed as. . . .
   c. Suppose the daily freight cost per mile is three-fifths of the value you computed in part (b). Total cost is minimized at the [city center, suburb] with a cost of _______, computed as. . . . Illustrate with a graph.

d. For part (*c*), the firm chooses [city center, suburb] because the cost of moving its output one mile is _______, while the cost of moving the workforce one mile is _______.

4. Dink Commuting

Consider the Dinks (double income, no kids). Mr. Dink commutes to a job in the city center ($x = 0$), while Ms. Dink commutes to a suburban subcenter four miles east of the city center. The Dinks consume the same quantity of housing at all locations. Travel speed is the same in both directions.

a. Draw the household's housing-price curve up to a distance of seven miles.

b. Draw the housing-price curve under the assumption that the speed of inward commuting (toward the city center in the morning and away from the center in the evening) is half the travel speed of outward commuting.

c. Draw the housing-price curve under the assumption that travel speed is the same in both directions, but Ms. Dink has a higher opportunity cost of travel time.

5. Gas Tax and Suburb versus Central City

Consider a software firm with 10 workers who live in a suburb due east of the city center and commute by automobile at a cost of $20 per worker per day. The firm exchanges products and information with other software firms in the center, and the savings in exchange costs from locating in the center as opposed to the suburb is $300 per day.

a. The firm will be willing to pay more for a site in the [city center, suburb] because. . . .

b. Suppose a gas tax doubles the daily commuting cost. The firm will be willing to pay more for a site in the [city center, suburb] because. . . .

c. Suppose that cost of commuting by public transit is $28 per worker per day. The gas tax remains. The firm will be willing to pay more for a site in the [city center, suburb] because. . . .

6. Improving Mass Transit and Suburbanization

Consider a modern city in which office employment has been steadily shifting from the city center to suburban locations along beltways. Suppose that the city improves its mass-transit system, decreasing the monetary and time costs of radial travel. Assume the improvement in the transit system does not affect the spatial distribution of residents. The improvement in the transit system will [speed up, slow down] the movement of office firms to the suburbs because. . . .

7. Was the Monocentric City a Fluke?

Consider the following statement: "The large traditional monocentric city of the 19th century was a fluke, a result of a particular sequence of technological innovations in intraurban transportation. If the sequence had been slightly different, the large monocentric city would have never developed. Instead, we would have gone from the small cities of the 18th century directly to the large multicentric, suburbanized city that we see today."

a. For each of the following intraurban transport technologies, specify its starting date (the year it was introduced): horse-drawn wagon (chariot) _______; omnibus _______; cable car _______; trolley (streetcar) _______; elevator _______; intracity truck _______.

b. If we change the date for the introduction of the _______ to _______, the large monocentric city would never have developed because. . . .

8. Urban Villages: Two Assumptions

The traditional monocentric city was a segregated city in the sense that all employment was in the central core and most residents lived outside the core. The modern city is less segregated in the sense that employment is dispersed. According to Mr. Wizard, "If my assumptions are correct, land use in the typical American city will soon be completely integrated, with each manufacturer and each office firm surrounded by its workforce. Every worker will travel less than a mile to work." Assume that Mr. Wizard's reasoning is correct. Also assume that each household has a single worker.

a. The most important assumption is that. . . .

b. Another, less important assumption is that. . . .

9. By Land and by Air

Consider a manufacturing firm that occupies one hectare of land in rectangular city. The firm produces 10 tons of output per day and transports six tons on trucks via an interstate highway four miles east of the city center ($x = 4$) and transports the remaining four tons on airplanes that leave from an airport seven miles east of the city center ($x = 7$). Intraurban transportation is via trucks, with a unit cost of $20 per ton per mile. The firm does not engage in factor substitution as the price of land changes. To simplify, assume that labor costs are the same at all locations. The firm's bid-rent for land at the city center is $500.

a. Draw the firm's bid-rent curve for land from the city center to 10 miles east.

b. The slope of the bid-rent curve for the segment from $x = 0$ to $x = 4$ is _______, computed as. . . .

c. The slope of the bid-rent curve for the segment from $x = 4$ to $x = 7$ is _______, computed as. . . .

d. The slope of the bid-rent curve for the segment from $x = 7$ to $x = 10$ is _______, computed as. . . .

10. Bigger Prize for Tallest Building

Using Figure 7–4 as a starting point, suppose the prize for having the tallest building increases from $200 to $350.

a. The outcome of the competition for being the tallest will generate one building with _______ floors and a second building with _______ floors because. . . .

b. With the larger prize, the building profit of the winner (excluding the prize) is _______ and the building profit of the loser is _______.

c. The increase in the prize [increases, decreases, does not change] the total payoff, equal to the sum of building profits and the prize.

REFERENCES AND ADDITIONAL READING

1. Anas, Alex, Richard Arnott, and Kenneth A. Small. "Urban Spatial Structure." *Journal of Economic Literature* 34 (1998), pp. 1426–64.
2. Bartlett, Randall. *The Crisis of American Cities*. Armonk, NY: Sharp, 1998.

3. Bertaud, Alain. "The Spatial Organization of Cities: Deliberate Outcome or Unforeseen Consequence?" Working Paper, Institute of Urban and Regional Development, University of California, Berkeley, 2004.
4. Bertaud, Alain. "Clearing the Air in Atlanta: Transit and Smart Growth or Conventional Economics?" *Journal of Urban Economics* 54 (2003), pp. 379–400.
5. Bertaud, Alain, and Stephen Malpezzi. "The Spatial Distribution of Population in 48 World Cities: Implications for Cities in Transition." Working Paper, The Center for Urban Land Research, University of Wisconsin, 2003.
6. Bradbury, Katharine L., Anthony Downs, and Kenneth A. Small. *Urban Decline and the Future of American Cities*. Washington DC: Brookings Institution, 1982.
7. Brueckner, Jan K., Jacques-Francois Thisse, and Yves Zenou. "Why Is Central Paris Rich and Downtown Detroit Poor? An Amenity-Based Theory." *European Economic Review* 43 (1999), pp. 91–107.
8. Cullen, J. B., and S. D. Levitt. "Crime, Urban Flight, and the Consequences for Cities." *Review of Economics and Statistics* 81 (1999), pp. 159–69.
9. Frey, W. H. "Central City White Flight: Racial and Non-Racial Causes." *American Sociological Review* 44 (1979), pp. 425–88.
10. Fulton, William, Rolf Pendall, Mai Nguyen, and Alicia Harrison. "Who Sprawls Most? How Growth Patterns Differ across the U.S." *The Brookings Institution Survey Series*. July 2001, pp. 1–23.
11. Garreau, Joel. *Edge City: Life on the New Frontier*. New York: Doubleday, 1991.
12. Giuliano, Genevieve, and Kenneth Small. "Subcenters in the Los Angeles Region." *Regional Science and Urban Economics* 21 (1991).
13. Glaeser, Edward, and Matthew Kahn. "Decentralized Employment and the Transformation of the American City." NBER Working Paper, March 2001.
14. Glaeser, Edward, and Matthew Kahn. "Sprawl and Urban Growth." Chapter 56 in *Handbook of Regional and Urban Economics 4: Cities and Geography*, eds. Vernon Henderson and Jacques-Francois Thisse. Amsterdam: Elsevier, 2004.
15. Glaeser, Edward, Matthew Kahn, and Chenghuan Chu. "Job Sprawl: Employment Location in U.S. Metropolitan Areas." *Brookings Institution Survey Series,* May 2001, pp. 1–8.
16. Helsley, Robert, and William Strange. "A Game Theoretic Analysis of Skyscrapers," *Journal of Urban Economics*. (2008).
17. Hohenberg, Paul M., and Lyann H. Lees. The Making of Urban Europe 1000–1950. Cambridge, MA: Harvard University Press, 1986.
18. Kahn, Matthew. "The Environmental Impact of Suburbanization." *Journal of Policy Analysis and Management* 19 (2000).
19. Lang, Robert E. "Office Sprawl: The Evolving Geography of Business." *The Brookings Institution Survey Series*, October 2000.
20. Lang, Robert E. *Edgeless Cities*. Washington DC: Brookings, 2003.
21. McMillen, Daniel P. "One Hundred Fifty Years of Land Values in Chicago: A Nonparametric Approach." *Journal of Urban Economics* 40 (1996), pp. 100–24.

22. McMillen, Daniel P., and John F. McDonald. "Suburban Subcenters and Employment Density in Metropolitan Chicago." *Journal of Urban Economics* 43 (1998), pp. 157–80.
23. Mills, Edwin S. *Studies in the Structure of the Urban Economy*. Baltimore: Johns Hopkins, 1972.
24. Moses, Leon, and Harold Williamson. "The Location of Economic Activity in Cities." In *Readings in Urban Economics*, eds. Matthew Edel and Jerome Rothenberg. New York: Macmillan, 1972.
25. Nivola, Pietro. "Fat City: Understanding American Urban Form from a Transatlantic Perspective." *Brookings Review*, Fall 1998, pp. 17–20.
26. Schwartz, Alex. "Corporate Service Linkages in Large Metropolitan Areas: A Study of New York, Los Angeles, and Chicago." *Urban Affairs Quarterly* 28 (1992), pp. 276–96.
27. Sivitanidou, Rena. "Do Office-Commercial Firms Value Access to Service Employment Centers? A Hedonic Value Analysis within Polycentric Los Angeles." *Journal of Urban Economics* 40 (1996), pp. 125–49.
28. Wheaton, William. ''Income and Urban Residence: An Analysis of Consumer Demand for Location." *American Economic Review* 67 (1977), pp. 620–31.

Appendix: The Monocentric Model and Applications

This appendix describes and applies the model of a monocentric city, the prevailing urban form until the early 20th century. After describing the model and its implications, we apply the model in two ways. First, we explore the reasons for the concentration of low-income households in central cities. Second, we develop a general-equilibrium model of the urban economy that captures the interactions between the urban labor market and the land market.

THE MONOCENTRIC MODEL

We can summarize the transportation technology of the monocentric city as follows:

- **Central export node.** Manufacturing firms export their output from the city through a central export node—a railroad terminal or a port.
- **Horse-drawn wagons.** Manufacturing firms transport their output on horse-drawn wagons from their factories to the central node.
- **Hub-and-spoke streetcar system.** Workers travel by streetcar from residential areas to their jobs in the central business district (CBD).
- **Central information exchange.** The employees of different office industries meet in the city center to exchange information.

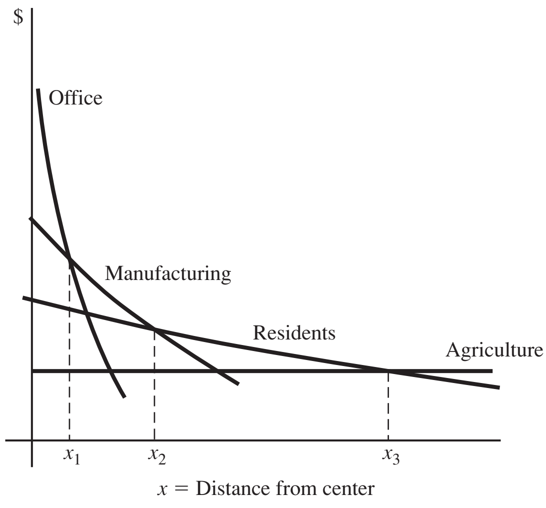
Under these assumptions, the city center is the focal point of the metropolitan area: Manufacturers are oriented toward the export node, office firms are oriented toward each other, and households are oriented toward manufacturing and office jobs.

Figure 7A–1 shows the allocation of land in the monocentric city. The office district is the area over which the bid rent of office firms exceed the bid rents of manufacturing firms, generating an office area with a radius of x_1 miles. Moving outward, manufacturing firms outbid other land users for land between x_1 and x_2, so the manufacturing district is a ring of width $x_2 - x_1$. Residents have the maximum bid rent for the area between x_2 and x_3, so the residential district is a ring of width $x_3 - x_2$.

In the monocentric city, both manufacturers and office firms are oriented toward the central business district. Why is the central area of the city occupied by the office industry? The geometric answer is that the activity with the steeper bid-rent curve occupies more central land. As we saw in Chapter 6, the slope of the bid-rent curve is determined by transport costs. The office industry has higher transport costs because it uses people—with a high opportunity cost of travel time—to transmit output. In contrast, manufacturing firms use horses and wagoneers.

Does the land market allocate land efficiently? In the terms used by land developers, Is land allocated to its "highest and best use"? The office industry, with its higher transport costs, occupies the land closest to the city center. This allocation is efficient because the office industry has more to gain from proximity to the city

FIGURE 7A–1 Bid Rents and Monocentric Land Use



The equilibrium land-use pattern is determined by the bid-rent curves of firms and residents. The office area is the area over which office firms outbid other users (from $x = 0$ to x_1). The area between x_1 and x_2 is the manufacturing district. Residents live in the area between x_2 and x_3.

center. If an office firm were to swap places with a manufacturing firm, the travel cost of the office firm would increase by a relatively large amount, while the freight cost of the manufacturing firm would decrease by a relatively small amount. As a result, total transport costs would increase. The market allocation, which gives central land to the office industry, economizes on transport costs.

The second feature of the monocentric model is that employment is concentrated in the central area, not distributed throughout the metropolitan area. Why do all the manufacturers and office firms locate in the central area? As we saw in this chapter, the cost of commuting on streetcars was low relative to the cost of moving output on horse-drawn wagons, so it was cheaper to transport workers from the suburbs to central factories rather than transporting output from a suburban production site to the central export node. The same logic applies to the office sector: A firm that moved to the suburbs would experience a large increase in the travel cost for information exchange and a relatively small reduction in wages.

INCOME AND LOCATION

In U.S. cities, the wealthy tend to locate in the suburbs, and the poor tend to locate near the city center. In other words, average household income increases as one moves away from the city center. Because the most expensive land is near the city center, this location pattern is puzzling: Why should the poor occupy the most expensive land?

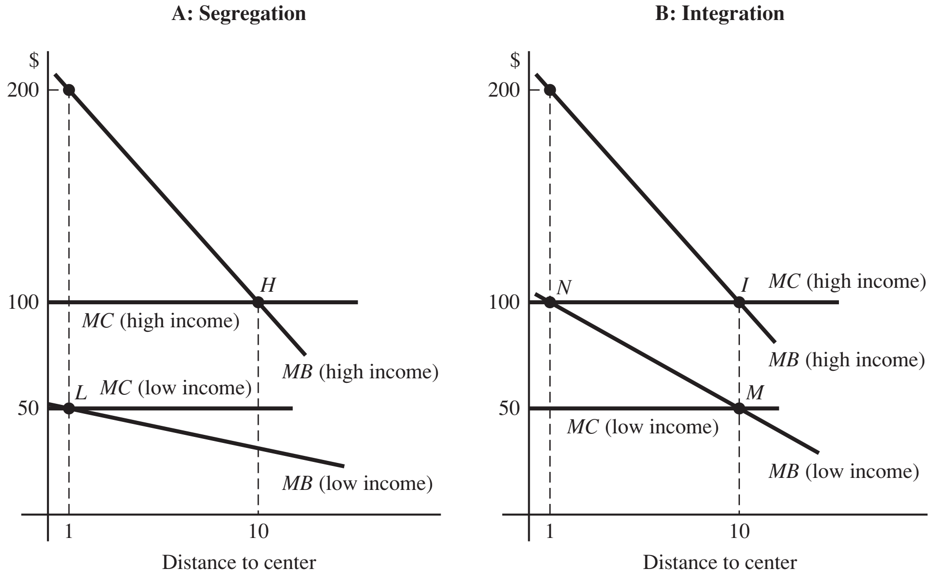
Trade-off between Commuting and Housing Costs

The traditional theory of income segregation suggests that central locations provide the best trade-off between commuting and housing costs for the poor, while suburban locations provide the best trade-off for the wealthy. Consider a household that is thinking about a move outward from a central location. An outward move of one mile generates costs and benefits.

- **Lower housing costs.** The price of housing will decrease, and the marginal benefit of moving outward equals the change in price times housing consumption. For example, consider a household that consumes 2,000 square feet. If the price of housing drops by \$0.10 per square foot, the marginal benefit of moving outward is \$200 (equal to \$0.10 times 2,000).
- **Higher commuting costs.** The marginal cost of a move outward is the marginal cost of commuting.

Figure 7A–2 (page 196) shows how two types of households, low-income and high-income, pick a residential location. The marginal-cost curves are horizontal, based on the assumption that the commuting cost per mile is constant with respect to distance. The marginal cost for the high-income household is more because it has a higher opportunity cost of commuting time. In Figure 7A–2, the marginal cost of the high-income household is \$100, compared to \$50 for the low-income household.

FIGURE 7A–2 Income and Location in Monocentric City



If the income elasticity of housing demand is larger than the income elasticity of commuting costs, the gap between the benefit curves will be large relative to the gap between the cost curves, and the low-income household will locate closer to the center.

If the income elasticity of housing demand equals the income elasticity of commuting costs, both types of households choose the same location.

The marginal-benefit curves are negatively sloped. This reflects the convexity of the housing-price curve. As a household moves away from the center, the reduction in the price of housing for the first mile exceeds the reduction for the second mile, and so on. For example, suppose the price of housing drops by \$0.10 for the first mile and then \$0.095 for the second mile. For a household that consumes 2,000 square feet, the marginal benefit for the first mile is \$200, and the marginal benefit for the second mile is \$190. The marginal-benefit curve for the high-income household is higher because it consumes more housing. In this example, the high-income household consumes four times as much housing as the low-income household (2,000 versus 500 square feet). At point *L*, the marginal benefit for the first mile is \$50 for the low-income household (a price reduction of \$0.10 times 500 square feet).

In Panel A of Figure 7A–2, the low-income household lives in the central city and the high-income household lives in a suburb. For the low-income household, the marginal benefit equals the marginal cost at one mile from the center (point *L*). Moving farther outward generates an additional cost of \$50 and an extra benefit less than \$50, so the more central site is better. The relatively large housing consumption of the high-income household means that the marginal benefit of moving outward exceeds the cost for up to 10 miles (point *H*). Low-income households locate close to the center because their relatively low housing consumption means they have little to gain by moving outward.

The income segregation result shown in Panel A of Figure 7A–2 is based on specific assumptions about the relationships between income, commuting costs, and housing consumption. Specifically, it assumes that housing consumption increases more rapidly with income. The high-income household has four times the housing consumption but just twice the commuting cost. In other words, the income elasticity of demand for housing is larger than the income elasticity of commuting cost. Geometrically, the gap between the benefit curves is larger than the gap between the cost curves, so the high-income household lives farther from the center.

Other Explanations

Wheaton (1977) provides empirical evidence that questions the validity of this conclusion. His results suggest that the income elasticity of demand for housing is roughly equal to the income elasticity of commuting cost. Therefore, the gap between the benefit curves will be roughly the same as the gap between the cost curves. In Panel B of Figure 7A–2, the high-income household has twice the housing consumption (2,000 versus 1,000) and twice the commuting cost, so the gap between the benefit curves matches the gap between the cost curves. As a result, the two households will choose the same residential location 10 miles from the city center.

This suggests that the observed pattern of income segregation cannot be explained by the trade-off between commuting cost and housing cost, so we must look for other explanations:

1. **New suburban housing.** If high-income households prefer new housing to old housing, they are pulled to the suburbs, where new housing is built.
2. **Fleeing central-city problems.** High-income households are relatively sensitive to crime and other problems, so they are willing to pay more for suburban housing.
3. **Suburban zoning.** As explained later in the book, suburban governments use zoning to exclude low-income households.

Other countries have different patterns of location by income. According to Hohenberg and Lees (1986), European cities have large concentrations of high-income households near the center. A good example of the European pattern is the Paris metropolitan area, where the average income in the central city exceeds the average income in the surrounding suburbs.

Brueckner, Thisse, and Zenou (1999) contrast Paris and Detroit. Paris has a rich mixture of cultural amenities (museums, restaurants, parks, street life) that make central Paris attractive relative to the suburbs. The demand for these cultural amenities increases rapidly with income, so the forces pulling the rich toward the central city (access to jobs and cultural amenities) dominate the forces pulling them toward the suburbs (lower housing prices). In contrast, Detroit has few cultural opportunities in the city center, so there is little to counteract the pull of the low housing prices in the suburbs.

A GENERAL EQUILIBRIUM MODEL OF A MONOCENTRIC CITY

We can use the model of the monocentric city to explore the interactions between the urban labor market and land market. The model considers a city that is small (one of many in a nation) and open (people move freely between cities). The utility level of residents is determined at the national level and is unaffected by changes in the city. In other words, the utility level of city residents is fixed, but the population of the city varies.

Interactions between the Land and Labor Markets

To simplify matters, we adopt two assumptions. First, we assume for the moment that there is no consumer substitution or factor substitution. Therefore, population density is the same at all residential locations and employment density (workers per square mile) is the same at all business locations. Later we'll see how variation in density affects the analysis. The second assumption is that the city is not circular, but rectangular with a one-mile width.

Panel A of Figure 7A–3 shows the urban land market. The business bid-rent curve intersects the residential curve at point *b*, generating a two-mile CBD. The total demand for labor in the city equals the land area of the CBD (two square miles) times the employment density (workers per square mile, equal to 60,000), or 120,000 workers. The residential bid-rent curve intersects the agricultural curve at point *r*, generating a six-mile residential area (from mile two to mile eight). The total supply of labor in the city equals the land area of the residential area (six square miles) times residential density (workers per square mile, equal to 20,000), or 120,000 workers.

Panel B of Figure 7A–3 shows the urban labor market, with a negatively sloped demand curve and a positively sloped supply curve.

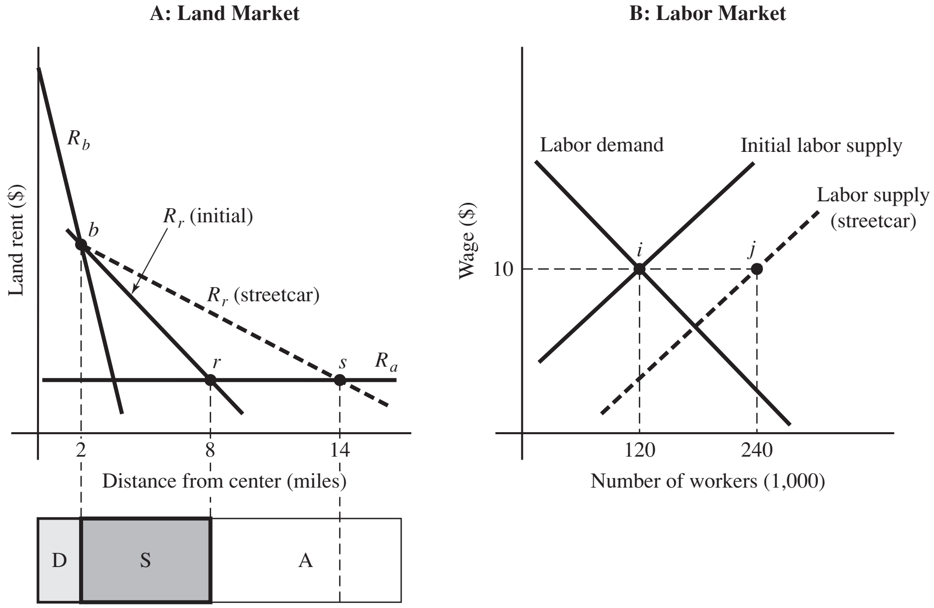
- **Negatively sloped demand curve.** An increase in the wage increases production costs and decreases the bid rent for businesses land, so the business territory shrinks, decreasing the quantity of labor demanded.
- **Positively sloped supply curve.** An increase in the wage increases the bid rent for residential land, increasing the territory of the labor-supply sector.

In the initial equilibrium shown by point *i*, the wage is $10 and the quantity of labor demanded (from the business district) equals the quantity supplied (from the residential district), with 120,000 workers.

The General Equilibrium Effects of the Streetcar

Consider the effects of introducing a streetcar into the monocentric city. In Panel A of Figure 7A–3, the streetcar decreases the unit commuting cost so it tilts the residential bid-rent curve outward. The residential area expands into the previously agricultural area (at point *s*, the residential area has doubled), increasing labor supply.

FIGURE 7A–3 Interactions between Land and Labor Markets and the Streetcar



A: In the initial equilibrium, the business bid-rent curve intersects the residential bid rent at point *b*, generating a two-mile business district. The residential bid-rent curve intersects the agriculture bid-rent curve at point *r*, generating a six-mile residential district.

The streetcar decreases commuting cost and tilts the residential bid-rent curve outward, and the new curve intersects the agriculture bid-rent curve at 14 miles, increasing the size of the residential area.

B: In the initial equilibrium, shown by point *i*, the wage is $10 and there are 120,000 workers. The streetcar increases labor supply, shifting the supply curve to the right and causing excess supply of labor.

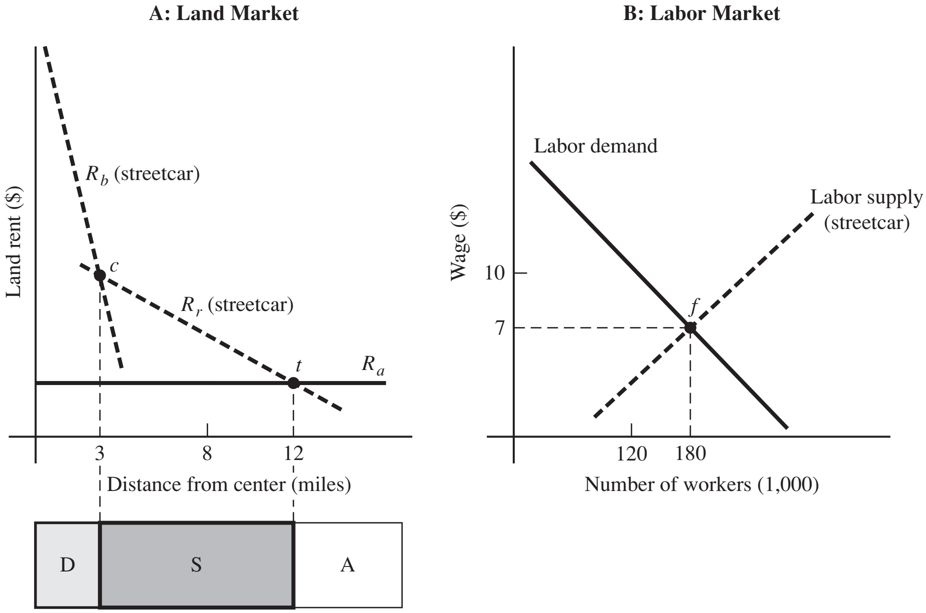
In Panel B, this is shown as a rightward shift of the labor supply curve, which causes an excess supply of labor at the original wage.

The excess supply of labor will cause the city's wage to fall. The wage drop causes two changes in the urban land market:

- The residential bid-rent curve shifts downward because residents' income decreases.
- The business bid-rent curve shifts upward because the decrease in wages decreases production costs.

These changes in the land market eliminate the excess supply of labor. The downward shift of the residential curve decreases the territory of the residential sector, decreasing the quantity of labor supplied. The upward shift of the business curve increases the territory of the business sector, increasing the quantity of labor demanded.

FIGURE 7A–4 Equilibrium after the Streetcar



A: In the streetcar equilibrium, the business bid-rent curve intersects the residential bid rent at point *c*, so the business district is three miles (up from two miles). The residential bid-rent curve intersects the agriculture bid-rent curve at point *t*, so the residential district is nine miles (up from six miles).
B: In the streetcar equilibrium shown by point *f*, the wage is $7 (down from $10) and there are 180,000 workers (up from 120,000).

Figure 7A–4 shows the new equilibrium. In Panel A, point *c* shows that the business district grows to three miles. Point *t* shows that the residential district is now nine miles (from mile three to mile 12). Compared to the initial equilibrium, land rent is higher in both the business and residential areas. In the residential area, land rent is generally higher. Throughout the residential area, the decrease in the wage pushes land rent down while the decrease in commuting cost pulls land rent up. For locations close to the new border at three miles, the negative effect of lower wages may dominate the positive effects of lower commuting costs, so land rent may decrease. More distant locations experience the same reduction of wages but larger savings in commuting costs, so land rent increases. In the business district, the streetcar increases land rent because it increases labor supply and reduces the wage and firm's production costs. In Panel B, the new equilibrium is shown by point *f*: The wage is $7 (down from $10) and there are 180,000 workers in the city (up from 120,000).

So far, we have assumed that both employment density and residential density are fixed. If we relax this assumption, the streetcar will increase density, reinforcing the changes shown earlier. The general effect of the streetcar is to increase land rent, and increases in land rent increase density: Households will economize on land by living on smaller lots; firms will economize on land by producing in taller buildings on smaller production sites.

APPLYING THE CONCEPTS

1. **Paris versus Detroit**
   Suppose the benefit and cost curves of the high-income households shown in Figure 7A–2 are relevant for Detroit. In contrast, high-income households in Paris take 50 percent more trips to the city center (for work and cultural opportunities). While Detroiters commute in air-conditioned land yachts, tuna boats, and Suburban Parcels, Parisians pack into mass transit. The armpit factor (visualize Parisians in packed buses hanging onto overhead straps) makes the unit cost of travel one-third higher in Paris. Use a graph like Figure 7A–2 to predict where high-income Parisians live.
2. **Global Warming and Rising Sea Level**
   Consider Aquaville, a rectangular monocentric city depicted in Figure 7A–3. Suppose that global warming raises the sea level, flooding one-fifth of the CBD. Depict graphically the effects of the flooding on (1) the city's wage, (2) land rent within the CBD, (3) land rent in the residential area, (4) the size of the CBD, and (5) the size of the residential area. Assume that flooding eliminates one-fifth of CBD land at each distance from the central export node.

CHAPTER 8

Neighborhood Choice

Love thy neighbor as yourself, but choose your neighborhood.
—Louise Beal

When a household chooses a house or apartment, it is choosing much more than a dwelling. It is also choosing a set of local public goods (schools, parks, and public safety) and a set of taxes to finance the public goods. The household is also choosing a set of neighbors who provide opportunities for social interactions and send their kids to the same schools. In this chapter, we explore the economics of neighborhood choice. In contrast with our earlier analysis of commuting-based residential choice, the analysis in this chapter considers a variety of neighborhood characteristics.

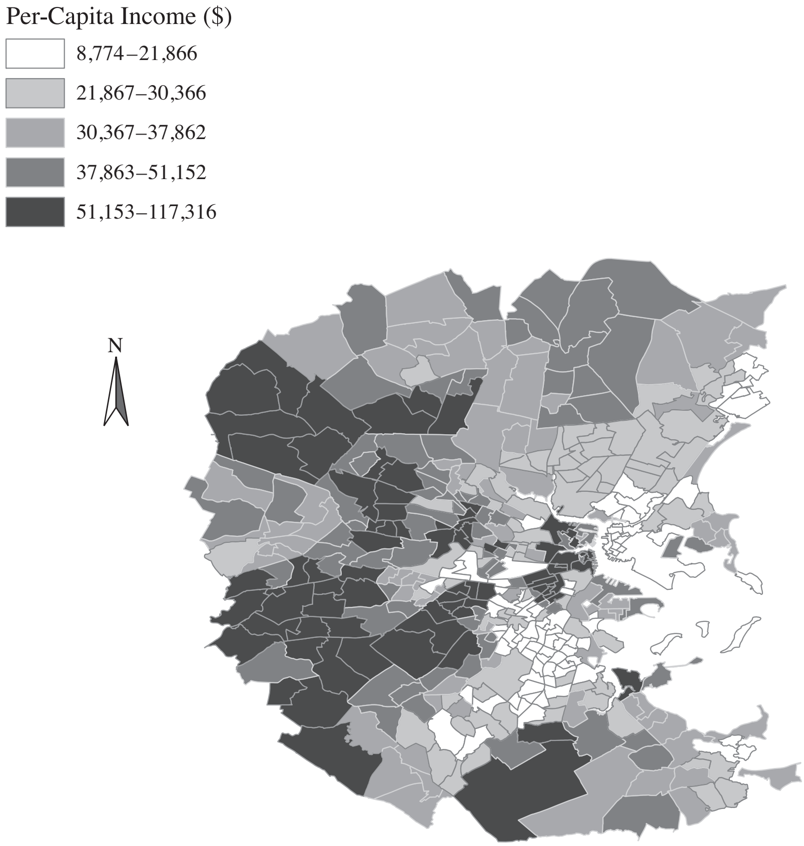
DIVERSITY VERSUS SEGREGATION

Our discussion of neighborhoods considers the issue of diversity. At one extreme is a city of diverse neighborhoods, each with an equal representation of households of different races and income levels. At the other extreme is a segregated city with a neighborhood for each type of household, rich and poor, black and white.

In cities in the United States, there is substantial segregation with respect to income and educational attainment. Map 8–1 shows per-capita income by census tract in Boston. At the tract level, per-capita income ranges from \$8,774 to \$117,316, with a median of \$33,598. The shading is organized by the quintiles of the distribution of per-capita income. The first quintile includes the poorest fifth of census tracts (per-capita income from \$8,774 to \$21,866) and the fifth quintile includes the richest fifth of census tracts (per-capita income from \$51,153 to \$117,316). The higher the per-capita income of a census tract, the darker the shading. For each income quintile, census tracts are not randomly distributed across the metropolitan area but are instead relatively clustered.

There is also substantial segregation within U.S. cities with respect to educational attainment. Map 8–2 shows educational attainment by census tract in Denver, measured as the number of people with four-year college degrees per 1,000 adults

MAP 8–1 Income Segregation: Boston

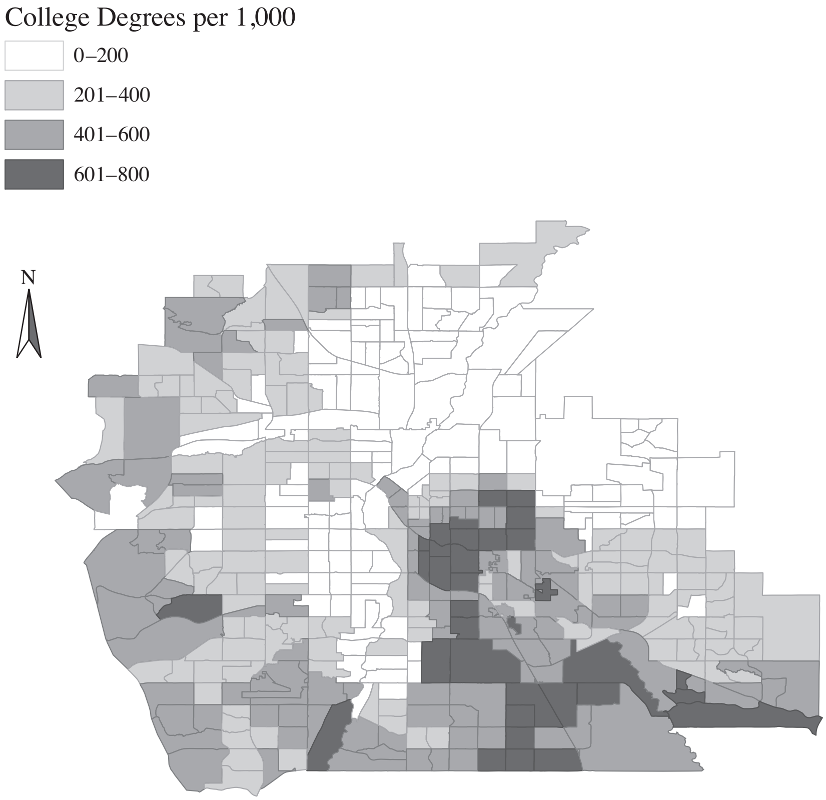


(age 25+). The median college share is 303 per 1,000, with a range of zero to 801. For each attainment class, census tracts are not randomly distributed across the metropolitan area, but instead are clustered.

SORTING FOR LOCAL PUBLIC GOODS

In the typical metropolitan area, there are dozens of municipalities, each with a different mix of local public goods and taxes. In addition, there are many school districts, each with a different education program. The wide variety of municipalities and school districts allows citizens to “vote with their feet,” choosing the jurisdiction with the best combination of public services and taxes. In this part of the chapter, we’ll explore the role of local public goods and local taxes in neighborhood choice. Later in the book, we’ll take a more detailed look at local public goods and taxes.

MAP 8–2 Segregation with Respect to Educational Attainment: Denver

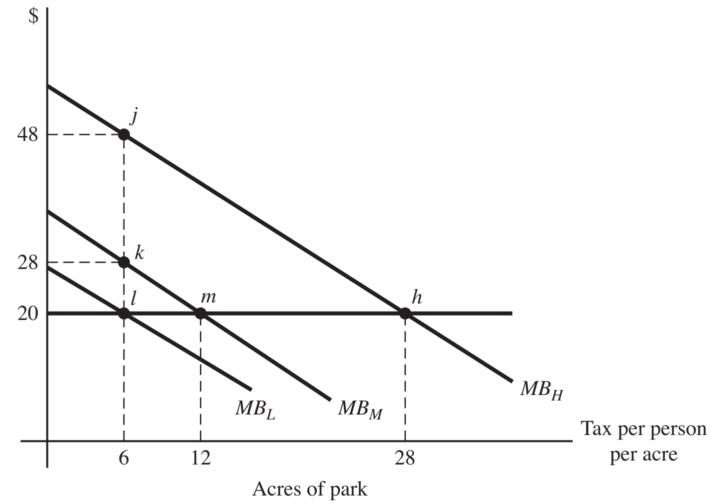


Diversity in Demand for Local Public Goods

Consider a three-person city that provides a local public good, a park. The park is a public good because it is nonrival in consumption: The fact that one person in the city benefits from the park doesn't reduce the benefit for another person. If the benefits of the park are confined to people in the city, the park is a *local* public good. Citizens in the city vary in their demand for park acreage and will collectively decide how big the park will be.

Figure 8–1 shows three individual demand curves for parks, one for each citizen. As shown in Section 2.1 of "Tools of Microeconomics," the appendix at the end of the book, a demand curve shows how much a person is willing to pay for one more unit of a product, so it is a marginal-benefit curve. Lois has a relatively low demand for parks and a low marginal-benefit curve (MB_L). For example, point *l* indicates that she is willing to pay $20 for the sixth acre. In contrast, Marian has a medium demand for park acres, and her marginal-benefit curve (MB_M) shows that she is willing to pay $28 for the sixth acre (point *k*). Hiram, with a high demand (MB_H), is willing to pay $48 for the sixth acre (point *j*).

FIGURE 8–1 Diversity of Demand for Local Public Good



The demand curves of the individual citizens show their marginal benefits of park acres. With a tax of $20 per person per acre, Lois prefers six acres, Marian prefers 12 acres, and Hiram prefers 28 acres.

The three citizens must collectively pick a park size. Suppose the cost per acre of parks is $60, and the city pays for the park with a tax of $20 per citizen per acre. In Figure 8–1, the horizontal line at $20 shows the marginal cost of park acreage for each citizen. For each additional acre, each citizen pays $20, one-third of the $60 cost. For each citizen, the preferred park acreage is the quantity at which his or her marginal benefit equals the $20 marginal cost. Lois prefers six acres (point *l*), while Marian prefers 12 acres (point *m*), and Hiram prefers 28 acres (point *h*). In other words, diversity in demand for park acreage means that citizens disagree about how large the park should be.

Problems with Majority Rule and Formation of Municipalities

Under majority rule, the city will choose Marian's preferred park of 12 acres. Suppose the city holds a series of elections between pairs of parks of different sizes. As shown in Table 8–1, in an election between six acres (Lois's favorite) and 12 acres (Marian's favorite), Hiram joins Marian to approve Marian's choice because

TABLE 8–1 The Median Voter Always Wins

Election	Votes for Median (12 acres)	Votes for Nonmedian
6 acres vs. 12 acres	Marian and Hiram	Lois
28 acres vs. 12 acres	Marian and Lois	Hiram

12 acres is closer to his preferred level. In an election between Marian's favorite and a larger park (28 acres), Lois joins Marian to approve Marian's choice. Marian's preferred size wins both elections because she is the median voter, defined as the voter who splits the rest of the voting public into two equal halves. Marian wins because she can always get one voter to join her to defeat any alternative to her preferred option.

Majority rule leaves two of three citizens unhappy with local government. The park is too big for Lois, who would prefer six acres, and too small for Hiram, who would prefer 28 acres. The necessity of choosing one park size for citizens with different preferences means that two of three citizens will be dissatisfied and will look for alternative arrangements.

One alternative is to form a new municipality with citizens with similar park preferences. Consider a metropolitan area that starts with three heterogeneous municipalities: Each municipality initially has one citizen of each type (Lois, Marian, Hiram). Under majority rule, each municipality will have two dissatisfied citizens looking for alternatives. For example, the three Lois-type citizens could leave their old municipalities and form Loisville, a municipality that will provide a six-acre park. Similarly, Marianville (with three Marians) would provide a 12-acre park, and Hiramville would provide a 28-acre park. By voting with their feet and sorting themselves into homogeneous communities, each citizen gets her or his preferred park size.

Variation in Consumption of the Taxed Good

Up to this point, we have assumed that the local public good is financed with a common head tax. In Loisville, the head tax would be $120 per head ($20 per acre times six acres), just high enough for each citizen to pay one-third of the cost of providing $360 worth of parks ($60 per acre times six acres). Suppose the government switches to a variable head tax: The heavier your head, the higher your tax. Assume that there are three head sizes in Loisville: Pin has a two-pound head, while Avner has a 10-pounder, and Gordo has a 24-pounder. As shown in the first row of Table 8–2, a tax of $10 per pound would generate tax bills of $20 for Pin ($10 times two pounds), $100 for Avner, and $240 for Gordo, just enough to cover the $360 cost of the city park.

TABLE 8–2 Municipality Formation for Tax Purposes

		Tax Bill		
Outcome	Tax Rate per Pound	Pin (small head)	Avner (average head)	Gordo (big head)
Mixed municipality	$10	$ 20	$100	$240
Exclusive small head	$60	$120	—	—
Exclusive average head	$12	—	$120	—
Exclusive big head	$ 5	—	—	$120

Although every citizen in Loisville has the same preferences for parks and benefits equally from the park, they have different tax bills. Gordo pays 12 times as much as Pin and has an incentive to form a new municipality with other big-head people. As shown in the last row of Table 8–2, if Gordo forms a big-head municipality with two other people, they could raise \$120 per capita with a tax rate of only \$5 per pound (\$5 times 24 = \$120). In other words, forming the big-head municipality cuts the head tax in half. Similarly, the average-head people have an incentive to form a municipality and exclude the pinheads. As shown in the second row of Table 8–2, a pinhead municipality needs a tax of \$60 per pound to generate the \$120 per person required to support the preferred park acreage.

The introduction of taxes that vary across individuals increases the equilibrium number of municipalities. In this example, there are nine municipalities, equal to the number of consumer types (three) times the number of head types (three). In equilibrium there will be three low-demand municipalities, each with a different head size. Similarly, there will be three medium-demand municipalities and three high-demand municipalities.

Real municipalities don't tax heads but instead use property taxes to finance local public goods. The basic logic of sorting with respect to head sizes applies to taxation based on housing consumption. Instead of big-head and pinhead municipalities, there will be big-house and small-house municipalities. People who own relatively expensive houses have an incentive to form municipalities with other big-house citizens in order to avoid paying more than their share of taxes. So if there are three types of preferences for local public goods (low, medium, and high) and three house sizes (small, medium, and big), there will be nine municipalities.

The sorting of households with respect to the demand for local public goods and the demand for taxed goods contributes to income segregation. If the demands for local public goods and the taxed goods depend on income, sorting will lead to municipalities with different income levels. Later in the book, we will develop a formal model of this sorting process and explore some of its consequences. Specifically, we will show how sorting with respect to public goods and tax bases generates a fragmented system of local government, with dozens of municipalities in each metropolitan area. We will also discuss the efficiency implications of fragmented government.

NEIGHBORHOOD EXTERNALITIES

Interactions among neighbors generate neighborhood externalities (Durlauf, 2004). Recall the third axiom of urban economics:

Externalities cause inefficiency

As we saw in Chapter 1, an externality is an unpriced interaction, and it can be positive or negative. A positive externality occurs when a person is not compensated for an action that benefits someone else. A negative externality occurs when a person does not pay for an action that imposes a cost on someone else. Social interactions at the neighborhood level generate several types of externalities for both children and adults.

Consider first the externalities relevant for children. Children imitate adults, and a neighborhood of educated and successful adults provides good role models for kids. Successful adults don't get a dollar every time they inadvertently encourage kids to stay in school, so there is a positive externality. In schools, the most important factor in student learning is the peer group: Kids learn more when they are surrounded by other kids who are motivated and focused. Motivated kids don't get a dollar every time they do their homework, so there is a positive externality. Troublesome kids don't pay a dollar every time they disrupt class, so there is a negative externality.

An important facet of children's externalities is imitative or self-reinforcing behavior. A person who joins a group benefits from social interaction but also tends to imitate the behavior—good or bad—of the members of the group. Imitation occurs for three reasons:

- There is a psychological payoff from behaving like others.
- A group provides a wider set of opportunities. For example, a chess club or a drama club will provide opportunities to interact with other high achievers.
- A group generates better information about future opportunities. For example, college recruiters target high-achieving students in chess and drama clubs.

Although the social aspects of a chess club or a drama club may be similar to those of a street gang, the self-reinforcing or imitative aspects generate very different long-term employment prospects.

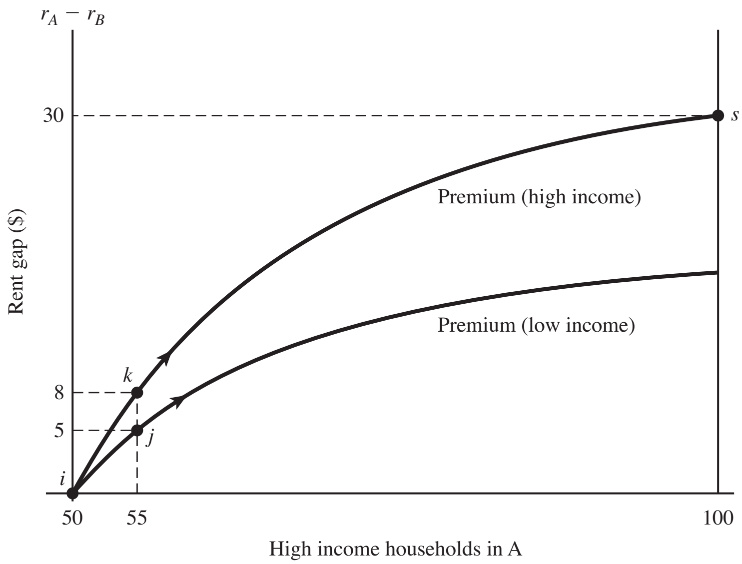
Consider next the externalities relevant for adults. In addition to regular social interactions, adults may get better information about job opportunities from their neighbors. Much of the information about employment opportunities comes from informal sources such as neighbors and friends. On the negative side, drug abuse among neighbors generates an unpleasant living environment. These are externalities because neighbors don't charge each other for information that leads to job prospects, and drug abusers don't compensate their neighbors for the unpleasant environment.

These neighborhood externalities affect a household's choice of a neighborhood. Most households have the same preferences with respect to adult role models and school peers, so they all would prefer the same sort of neighborhood. The positive externalities generated by a household generally increase with income and education level, so people generally prefer neighborhoods with large numbers of high-income, educated households. Of course, the number of such households is limited, so who gets them as neighbors?

NEIGHBORHOOD CHOICE

Households compete for places in a desirable neighborhood by bidding for housing and land in the neighborhood. In this part of the chapter, we focus on competitive bidding for neighborhoods that differ in their income mixes. We assume that the

FIGURE 8–2 Segregation Equilibrium



If high-income households have a steeper premium curve, the integrated outcome (point *i*, with 50 of each type of household in A) is unstable. Segregation (shown by point *s*) is the equilibrium, with all 100 high-income households in A. The equilibrium rent gap is $30: Rent is $30 higher in neighborhood A.

positive externalities from neighbors increase with income. This means that the attractiveness of a neighborhood increases with the number of high-income households. Will neighborhoods be integrated, with a mixture of high-income and low-income households, or will they be segregated?

We will use a model developed by Becker and Murphy (2000). Consider a city with two neighborhoods (A and B) and two income groups (high income and low income), each with 100 households. The only difference between the two neighborhoods is in their income mixes and the resulting neighborhood externalities. Each household occupies one unit of land, and there is a fixed amount of land in each neighborhood—just enough to accommodate 100 households. In Figure 8–2, the horizontal axis measures the number of high-income households in neighborhood A. Since each neighborhood has a total of 100 households, the number of low-income households in a neighborhood is 100 minus the number of high-income households.

The vertical axis in Figure 8–2 measures the difference in land rent between the two neighborhoods. Specifically, it measures the rent in the high-income neighborhood minus the rent in the low-income neighborhood. The horizontal axis starts at 50 high-income households, so under the assumption that high-income households generate positive externalities from role models and school peers, the rent premium is always positive, and the larger the number of high-income households, the larger the premium.

Segregation Equilibrium

The positively sloped curves in Figure 8–2 show the rent premiums for the two income groups. For example, suppose neighborhood A has 55 high-income households and 45 low-income households, and the numbers are reversed in neighborhood B. As shown by point *k*, a high-income household is willing to pay $8 more to live in neighborhood A rather than B. Similarly, point *j* shows that the low-income household is willing to pay a $5 premium for neighborhood A. At point *i*, the two neighborhoods are identical (50 high-income and 50 low-income households in each), so the rent premium is zero. The premium curves are positively sloped, reflecting the positive externalities from the high-income population.

Equilibrium requires that all households in a particular neighborhood pay the same rent. If they didn't, landowners with low prices would have an incentive to raise them, and households on expensive land would have an incentive to change locations. In our model of two household types, equilibrium requires that high-income and low-income households in a particular neighborhood pay the same rent.

The integrated outcome shown by point *i* is a symmetric but unstable equilibrium. It is symmetric because the two neighborhoods are identical. It is an equilibrium because the two premium curves intersect, meaning that both types pay the same rent premium for neighborhood A (zero when the neighborhoods are identical). Point *i* is unstable because a small movement of population will generate a different equilibrium. Suppose a group of five high-income households moves from neighborhood B to A, displacing five low-income households who move the other direction. Neighborhood A, with 55 high-income households, now has a more favorable mix of households than B, and both types of households are willing to pay more to live in A.

- High-income households are willing to pay an $8 premium (point *k*).
- Low-income households are willing to pay a premium of only $5 (point *j*).

High-income households will outbid low-income households, so after the first five high-income households move into the neighborhood, others will join them, displacing low-income households and leading us away from point *i*, not back toward it.

A deviation from the integrated outcome triggers self-reinforcing changes. In Figure 8–2, the arrows on the premium curves indicate the direction of movement. Whenever the high-income premium curve is above the low-income curve, high-income households will outbid low-income households, and the high-income population will increase (we move to the right). The high-income curve lies everywhere above the low-income curve, so the high-income population will continue to increase at the expense of the low-income population until neighborhood A has only high-income households (shown by point *s*). Recall the second axiom of urban economics:

Self-reinforcing changes lead to extreme outcomes

In this case, the self-reinforcing change is an increase in the number of high-income households in neighborhood A. This change makes the neighborhood even more attractive to high-income households, and the extreme outcome is that all high-income households locate in one neighborhood.

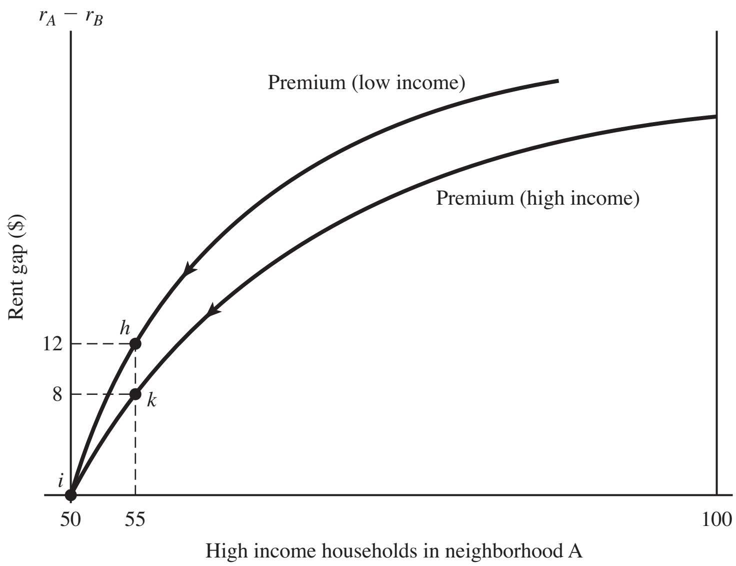
Point *s* represents income segregation because all the high-income households are in neighborhood A and all the low-income households are in neighborhood B. Although low-income households value proximity to high-income households, the high-income households value proximity to high-income households even more, so they outbid low-income households for the limited number of places in the more desirable neighborhood. In graphical terms, segregation happens because high-income households have a steeper premium curve, reflecting a larger marginal benefit of living close to high-income households. A one-unit increase in the number of high-income households increases the premium of high-income households by a larger amount, so they outbid low-income households for the more desirable neighborhood.

Integration as a Stable Equilibrium

We've seen an example in which income integration is an unstable equilibrium. Under what circumstances would the integrated equilibrium be stable? The stability of integration is determined by the slopes of the premium curves.

Figure 8–3 shows integration as a stable equilibrium. In this case, low-income households have a steeper premium curve. Suppose we start at point *i*, and five high-income households move to neighborhood A. What happens next?

FIGURE 8–3 Integration Is a Stable Equilibrium



If low-income households have a steeper premium curve, the integrated outcome (point *i*, with 50 of each type of household in neighborhood A) is stable. Any deviation from the integrated outcome is self-correcting. In equilibrium, rent is the same in the two neighborhoods.

- As shown by point *h*, a low-income household is willing to pay a premium of \$12 for a neighborhood with 55 high-income households.
- As shown by point *k*, a high-income household is willing to pay a premium of only \$8 for a neighborhood with more high-income households.

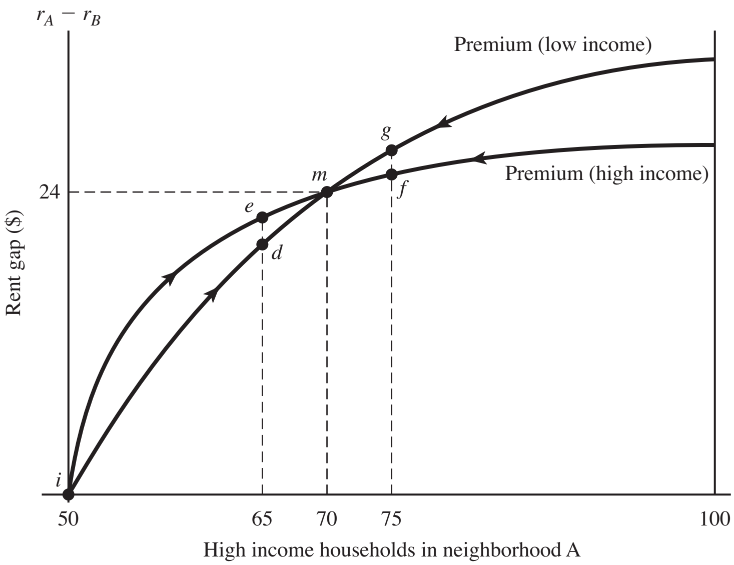
Low-income households will outbid high-income households, so the number of low-income households will increase at the expense of high-income households.

The arrows along the premium curves indicate the direction of the change if we start at any point other than the integrated outcome (point *i*). If high-income households outnumber low-income households (if the number of high-income households exceeds 50), low-income households will outbid high-income households, and we will move back toward point *i*. In other words, any deviation from the integration outcome is self-correcting, not self-reinforcing. Integration is stable because low-income households are willing to pay more than high-income households to live in a neighborhood with more high-income households.

Mixed Neighborhoods

A third possibility is a mixed neighborhood, with an income mix between perfect segregation and perfect integration. For example, we could have one neighborhood with 70 percent high-income households and a second with 70 percent low-income households. This is shown in Figure 8–4, where the two premium curves for

FIGURE 8–4 Mixed Neighborhood Equilibrium



If the two premium curves cross, with a steeper curve for low-income households, the intersection point is a stable equilibrium. At point *m*, 70 of 100 households in neighborhood A are high-income, and the rent gap is \$24.

neighborhood A intersect at a quantity of 70 high-income households (and 30 low-income households). That leaves 30 high-income households and 70 low-income households for neighborhood B.

Point *m* is a stable equilibrium. It is an equilibrium because the two premium curves intersect there, indicating that in neighborhood A, both types of households pay the same premium. Point *m* is stable because the premium curve for the low-income households is steeper at that point. To test for stability, consider two deviations from point *m*.

- **More high-income households.** Suppose the number of high-income households increases from 70 to 75. As shown by points *f* and *g*, low-income households are now willing to pay a bigger premium, so they will outbid high-income households, and the high-income population will shrink back to 70.
- **Fewer high-income households.** Suppose the number of high-income households shrinks from 70 to 65. As shown by points *d* and *e*, high-income households are now willing to pay a bigger premium, so they will outbid low-income households, and the high-income population will expand back to 70.

In the equilibrium with mixed neighborhoods, the equilibrium premium is $24. Recall the first axiom of urban economics:

Prices adjust to generate locational equilibrium

For equilibrium in the market for neighborhoods, each type of household, high-income and low-income, must be indifferent between the two neighborhoods. With a premium of $24, each household in neighborhood A pays $24 extra to live in the neighborhood with the more favorable income mix. This premium exactly offsets the benefits of the more favorable income mix, so we have a locational equilibrium.

Lot Size and Public Policy

Up to this point, we have assumed that each household occupies one unit of land. Of course, land is a normal good, and consumption increases with income. What are the implications of variable lot sizes for neighborhood choice and income segregation?

As we'll see, when high-income households consume more land, integration is more likely. As a starting point, consider the situation shown in Figure 8–2. A high-income household is willing to pay a premium of $8 to live in a neighborhood with 55 high-income households, while the premium for the low-income household is only $5. If both types of households occupy one unit of land, high-income households outbid low-income households, leading to segregation.

Things are different when the high-income household consumes more land. Suppose high-income households occupy two units of land, compared to only one unit for low-income households. The high-income premium of $8 translates into only a $4 premium *per unit of land*. As shown in Table 8–3, the high-income household has a lower premium per unit of land, so low-income households will outbid high-income households for land in the more desirable neighborhood. As a result,

TABLE 8–3 Lot Size and Integration

	Premium for High-Income Community	Lot Size	Premium per Unit of Land
Low income	$5	1	$5
High income	$8	2	$4

any deviation from the integrated equilibrium (e.g., 55 high-income households) will cause self-correcting (not self-reinforcing) changes as low-income households outbid high-income households. The result is integration, the symmetric equilibrium with 50 households of each type in each neighborhood.

Another way to think about the effects of lot size is to take the perspective of landowners, who of course maximize their rental income. If you have two units of land to rent, you can either rent to a single high-income household for $8 or to a pair of low-income households, each paying $5 for a total of $10. Obviously the pair of low-income households is a better choice. A high-income household loses the bidding battle because it competes against two low-income households.

This example illustrates the importance of land consumption in neighborhood choice and diversity. If the difference in land consumption between the two types of households is large relative to the difference in the premium, the low-income household will have a larger premium per unit of land, and integration will occur. On the other hand, if the difference in land consumption is relatively small, segregation will occur. For example, if the high-income household occupied only 1.33 units of land, its premium per unit of land would be $6 per unit of land ($8/1.33), and segregation will persist.

Minimum Lot Size Zoning and Segregation

Some local governments use minimum lot size zoning to control land use. Under this policy, the government specifies a minimum lot size for residential development and outlaws higher density. As we'll see in a later chapter, one motivation for such a policy is to exclude people whose tax contributions fall short of the costs they impose on local government.

One of the consequences of this zoning policy is income segregation. In Table 8–3, integration is a stable equilibrium when high-income households occupy twice as much land as low-income households. Suppose the government specifies a minimum lot size of two units of land, the quantity chosen by high-income households. This policy imposes an extra cost on low-income households—they have to buy twice as much land as they want—and it decreases their premium per unit of land to $2.5, now less than the $4 premium of the high-income household. Once low-income households are forced to consume the same amount of land as high-income households, they lose the bidding war for land in the more desirable neighborhood. The result is that integration (the market-equilibrium outcome) is replaced by segregation.

NEIGHBORHOOD CHOICES: THE ROLES OF EDUCATION AND CRIME

We've seen that income segregation occurs when high-income households outbid low-income households for slots in neighborhoods with more favorable neighborhood effects. In this part of the chapter we'll look at two sources of differences in neighborhood characteristics: schools and crime. In both cases, if high-income households are willing to pay more for a more favorable environment, income segregation occurs.

Education and Neighborhood Choice

In the typical metropolitan area, educational achievement varies across schools. Table 8–4 shows data on student performance, graduation rates, and socioeconomic characteristics for the eight high schools in the Portland, Oregon, school district. Student performance is measured as the percent of students who meet or exceed the state standard for mathematics. There is substantial variation in performance across the high schools, with the percentage ranging from 44 to 80 percent. The second column of numbers shows the percentage of students from households with relatively low income. The percentage of students who are economically disadvantaged ranges from 12 to 70 percent. The last column shows the percentage of students from racial minorities, which ranges from 19 to 74 percent.

A family's choice of a neighborhood affects the educational level of its children. The substantial variation in performance across schools could result from a number of factors, including differences in teachers and discipline, as well as the composition of the student body. Later in the book we will explore the education production function, which identifies the inputs that determine the educational achievement of students. As we'll see, a key input to the education of a student is his or her peer group (classmates and schoolmates). A student learns more when he or she is surrounded by fellow students who are smart, motivated, and not disruptive. Good peers come from home environments that encourage achievement.

TABLE 8–4 Variation in High School Achievement, Portland Public Schools

High School	Percent Meeting or Exceeding in Math	Percent Economically Disadvantaged	Percent Nonwhite
Lincoln	80	12	21
Franklin	73	50	46
Wilson	72	18	19
Cleveland	69	30	28
Grant	68	26	36
Benson	49	61	71
Madison	45	70	66
Jefferson	44	68	74

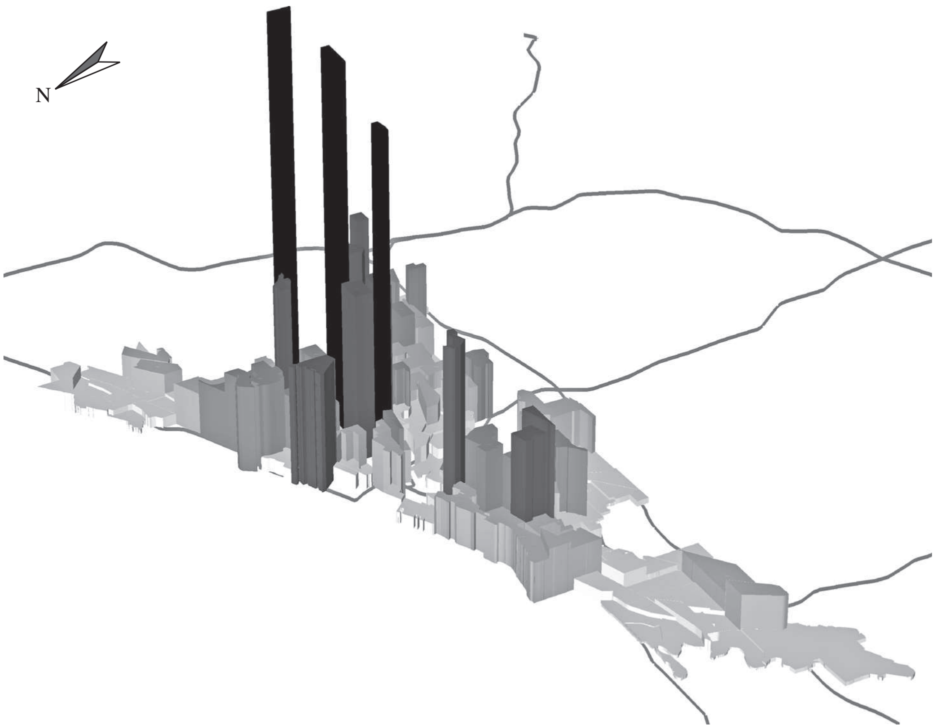
Source: Oregon Department of Education Report Cards, 2009–2010.

The variation in the performance of schools generates competition among households for the right to enroll in high-performance schools with favorable school peers. All parents want good schools and good peers for their kids, and the question is, who gets the best schools and peers? As we saw earlier, income segregation occurs if high-income households are willing to pay more than low-income households for neighborhoods with higher quality schools and better peers.

Crime and Neighborhood Choice

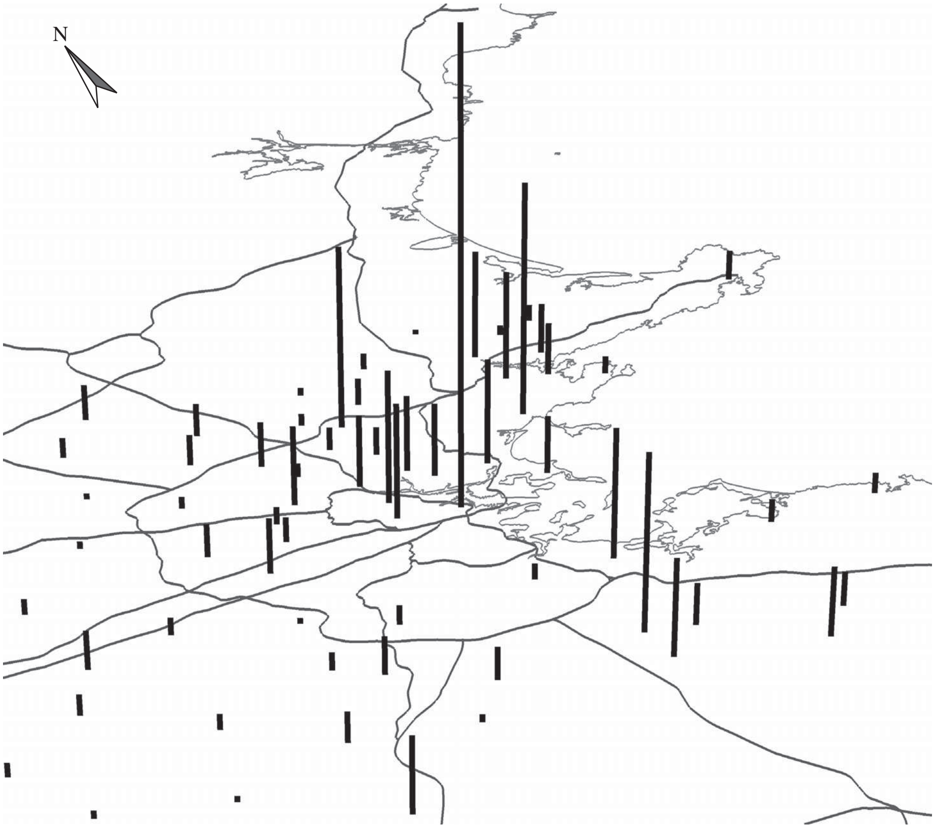
Another factor in neighborhood choice is crime. Map 8–3 shows the spatial variation in the costs of crime across census tract in the Cleveland municipality (not metropolitan area). The victim costs of crime include the opportunity cost of lost work time, monetary losses, and the costs of injuries. Miller, Cohen, and Wiersema (1996) estimated victim costs for different crimes as follows: $370 per larceny, $1,500 per burglary, $4,000 per auto theft, $13,000 per armed robbery, $15,000 per

MAP 8–3 Crime Costs: City of Cleveland



The map shows the annual victim cost per capita across census tracts in the municipality of Cleveland (not the entire metropolitan area). The victim cost reaches a maximum of $12,973 near the center. The lines represent freeways.

MAP 8–4 Spatial Variation in the Cost of Crime: Boston Metropolitan Area



The map shows the annual victim cost per capita across municipalities in the Boston metropolitan area. The victim cost reaches a maximum of $587 in the municipality of Boston. The lines represent freeways.

assault, $70,000 per rape, and $2.4 million per homicide. In Map 8–3, the jigsaw pieces representing census tracts are extruded by the annual per-capita victim cost. The cost reaches its maximum of $12,973 near the center, and there is substantial variation across census tracts in the municipality.

Map 8–4 takes a broader look at the spatial variation in the cost of crime, showing per-capita crime costs across municipalities in the Boston metropolitan area. Each bar represents one of the dozens of municipalities in the metropolitan area, and the height of the bar equals the annual per-capita cost of crime. The crime cost reaches its maximum of $587 in the municipality of Boston.

The substantial variation in crime rates mean that a household's choice of a neighborhood is influenced by crime. Households are willing to pay a premium to live in low-crime neighborhoods. A study of property crime (Thaler, 1977) estimates an elasticity of property values with respect to the crime rate of −0.067: A

10 percent increase in the crime rate decreases the market value of housing by about 0.67 percent, or $1,340 on a house worth $200,000.

What are the implications of crime for income segregation? Crime rates are generally lower in high-income neighborhoods, providing another attraction for all households, high-income and low-income. Which type of household is willing to pay more for low-crime neighborhoods? A recent study suggests that the willingness to pay for crime reduction increases with income (Cohen, Rust, Steen, and Tidd, 2004). Cullen and Levitt (1999) conclude that high-income households are more sensitive to crime: They flee in larger numbers when crime rates increase. These studies suggest that high-income households are willing to pay more than low-income households to live in low-crime (high-income) neighborhoods, so crime encourages income segregation.

RACIAL SEGREGATION

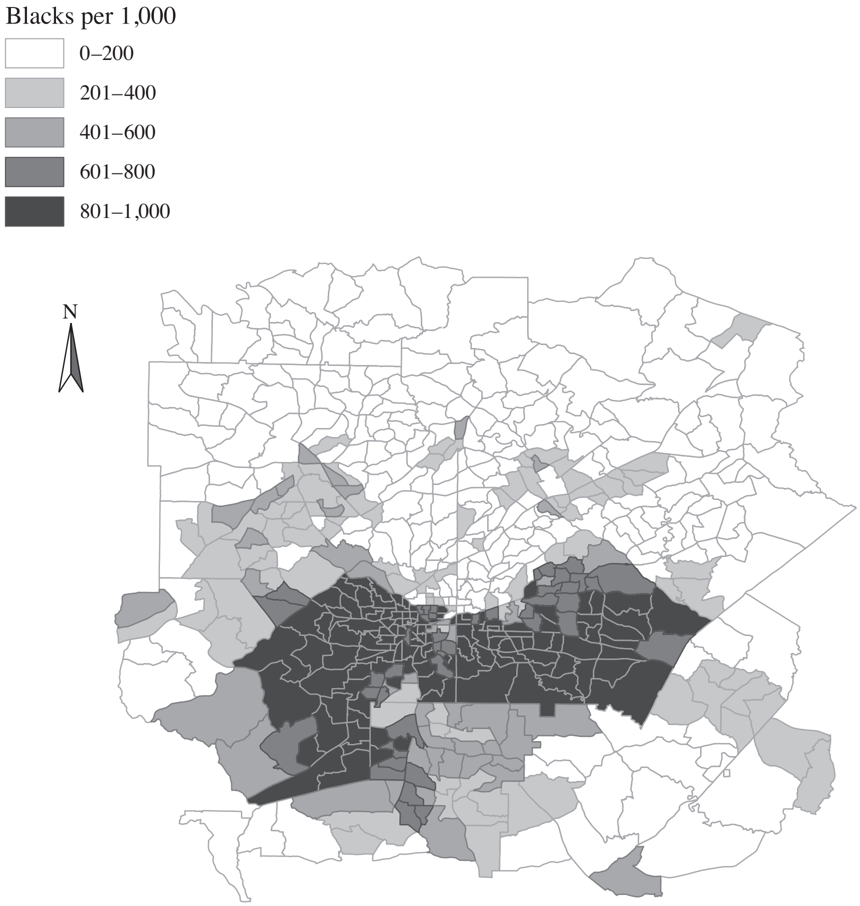
The framework we've used to explore income segregation can also be used to explore the issue of racial segregation. In the United States, more than two-thirds of the blacks living in metropolitan areas reside in central cities, leaving one-third of metropolitan blacks for suburban areas. For whites, the fractions are reversed: One-third live in central cities, leaving two-thirds for the suburbs.

One way to quantify the degree of racial segregation in a metropolitan area is the index of dissimilarity. This index shows the proportion of one race (e.g., blacks or African Americans) that must relocate to achieve racial integration, with each census tract in the metropolitan area having the same racial mix. For the United States, the index of 0.64 indicates that to achieve complete integration, 64 percent of blacks (or whites) would need to relocate. The dissimilarity index is highest for metropolitan areas in the Northeast (0.75) and lowest in the West (0.48).

Between 1980 and 2000, racial segregation as measured by the dissimilarity index decreased in 203 of 220 metropolitan areas. The average reduction was 12 percent, and 13 of the 43 largest metropolitan areas experienced reductions of at least 15 percent, while 6 experienced reductions of at least 20 percent. In 2000, the six most segregated metropolitan areas were Detroit, Milwaukee, New York, Newark, Chicago, and Cleveland. Between 1980 and 2000, these metropolitan areas experienced relatively small reductions in segregation.

Maps 8–5 and 8–6 show the black share of population by census tract in two metropolitan areas, Atlanta and Pittsburgh. In Atlanta, the median black share is 256 per 1,000, with a range of 1 to 992. The city has a dissimilarity ratio of 0.69, and the high degree of segregation is apparent from the map. In Pittsburgh, with a dissimilarity ratio of 0.73, the median black share across census tracts is 46, and the range is 0 to 982. The map also shows the rivers that flow through the city, the Allegheny and Monongahela, which join to form the Ohio River. Most of the census tracts with relatively high black shares are along the rivers.

MAP 8–5 Racial Segregation: Atlanta



Racial Preferences and Neighborhood Choice

What causes racial discrimination? One factor is household preferences for the racial mix of neighborhoods. A recent survey asked white and black respondents to state their ideal neighborhood mixes (Vigdor, 2009). The survey revealed fundamental differences in preferences.

1. For 19 percent of whites, the ideal mix is all white.
2. For 81 percent of white respondents, the ideal mix is less than 20 percent black.
3. The median ideal share of black neighbors (with half of respondents stating a larger share and half stating a smaller share) was 13 percent for white respondents, compared to 33 percent for black respondents.

MAP 8–6 Racial Segregation: Pittsburgh

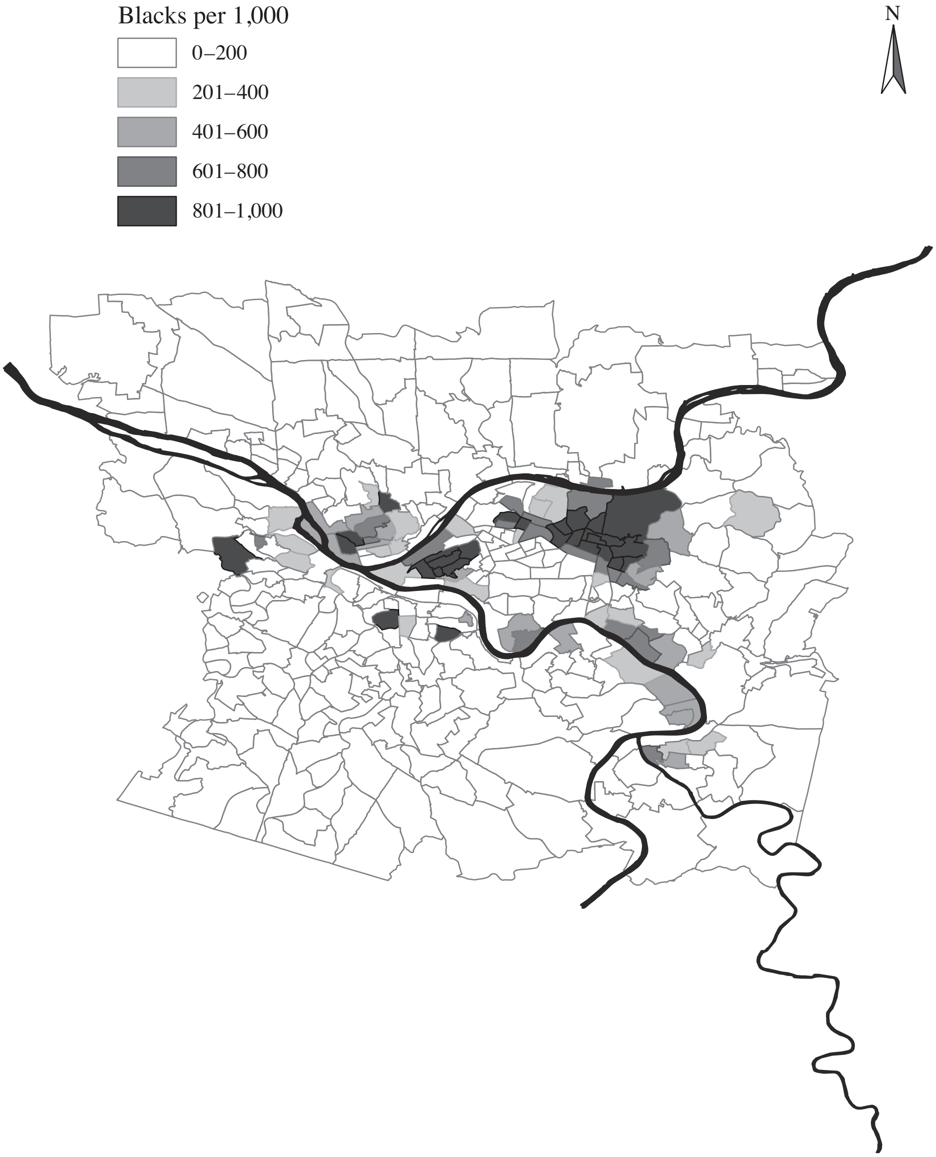
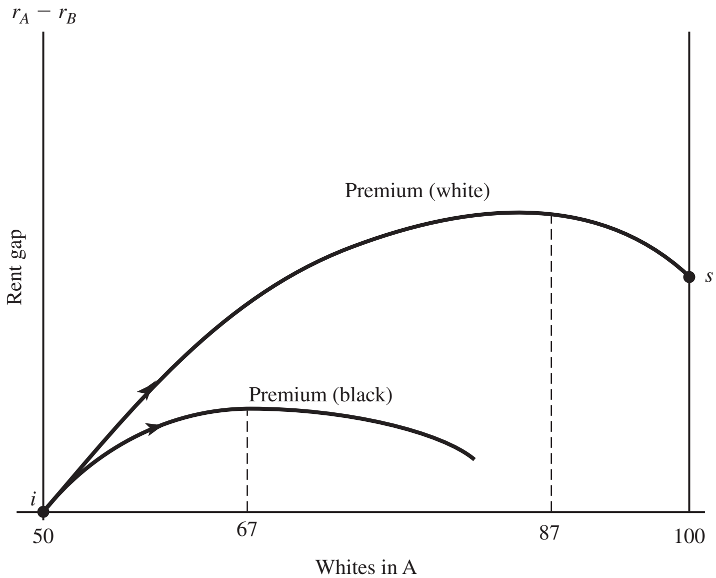


Figure 8–5 applies the neighborhood choice model to the issue of racial preferences and segregation. The higher curve reflects the general preferences of whites, with the ideal neighborhood having a mix of 87 percent white and 13 percent black. In contrast, the premium curve for blacks peaks at a mix of 67 percent white and 33 percent black. Because the white curve lies everywhere above the black curve, the only stable equilibrium involves racial segregation. As we saw earlier, although

FIGURE 8–5 Racial Segregation



The premium curves reflect racial preferences: For white households, the ideal mix is 87 percent white. For black households, the ideal mix is 67 percent white. Because the white curve lies everywhere above the black curve, the equilibrium in this case is segregation (point *s*).

the integration outcome (point *i*), is an equilibrium, it is not stable. If one or more white households moved to neighborhood A, the premium of whites would exceed the premium of blacks, so the white population would grow at the expense of the black household. The segregation equilibrium shown by point *s* is the only stable equilibrium.

What would be required to generate a stable integrated equilibrium? The slope of the black premium curve must be steeper than the white curve at the origin (50 white and 50 black households in neighborhood A). In this case, a deviation from the integrated outcome would be self-correcting because if the white population increased above 50, blacks would outbid whites for the limited number of places in the neighborhood.

Other Reasons for Racial Segregation

We've seen that racial preferences can lead to the extreme outcome of complete segregation. A second reason for racial segregation is income segregation. Black households have lower incomes on average. As explained by Mills and Lubuele (1997) housing is less expensive in the inner city, and this attracts low-income households, some of which are black. Several studies have shown that income segregation explains only a part of racial segregation.

A related factor is exclusionary zoning by suburban governments. As we saw earlier in the chapter, minimum lot-size zoning encourages income segregation, and so do other policies such as the prohibition of multifamily units, maximum densities, requirements for two-car garages, and development fees. Since black households have lower incomes, on average, than white families, exclusionary zoning has a larger effect on black households.

Another factor in racial segregation is racial discrimination by real-estate brokers. One technique is *racial steering:* The broker directs individual buyers away from predominantly white neighborhoods. According to Yinger (1998), blacks are treated differently from otherwise identical whites: Blacks are shown fewer dwellings, steered into certain neighborhoods, and given less advice and assistance on financing options. For example, 1 in 10 black renters is denied access to housing made available to white renters, and one in four learns about fewer vacant dwellings.

Until recently, federal housing policies have indirectly encouraged segregation. Historically, most public housing was concentrated in low-income areas, and until recently, housing vouchers (coupons used to help pay for private housing) could be used only in the city where the recipient lived when the voucher was issued. More recently, the federal government made most housing vouchers portable, so more recipients can use them to rent suburban housing.

THE CONSEQUENCES OF SEGREGATION

Why should we be concerned about segregation with respect to income or race? We've seen that diversity in tastes for local public goods and housing causes households to sort themselves into communities of like households. In addition, diversity in demand for some consumer goods causes sorting. For example, some households are attracted to communities with fast-food restaurants and retail strips, while others are attracted to areas with neighborhood coffee shops and bookstores. Spatial variation in crime causes sorting with respect to the willingness to pay for safe neighborhoods. In general, diversity in demand for goods that are tied to geography leads naturally to a sorting of households with respect to education and income. The question is whether sorting with respect to race and income limits the economic opportunities of some households, decreasing their incomes and causing poverty.

Unfavorable Neighborhood Effects

One reason to be concerned about sorting with respect to income is that the concentration of low-income households could generate harmful neighborhood effects. For example, in a neighborhood with a relatively low employment rate, information about job openings will be scarce, making it more difficult for the unemployed to find work. In a neighborhood with few high-school and college graduates, there will be few role models to inspire students to perform well in school and thus prepare themselves for the working world.

A number of studies have explored the effects of racial segregation on young black adults. Cutler and Glaeser (1997) show that an increase in segregation increases the probability of dropping out of high school and increases the likelihood of being idle (neither in school nor working). Segregation also leads to lower employment rates and higher rates of single parenthood. An important factor in these adverse outcomes is that young people living in a highly segregated, low-income environment have fewer contacts with positive role models—educated and successful people—and thus are less likely to be successful themselves.

A more recent study explores the consequences of segregation for immigrants to the United States (Cutler and Glaeser, 2008). The general conclusion is that immigrants with low levels of human capital (low educational attainment) tend to suffer in a segregated environment, while immigrants with high levels of human capital actually do better when they live in a large concentration of immigrants.

Limited Access to Jobs: The Spatial Mismatch

The second reason to be concerned about racial and income segregation is related to the geography of segregation. Location matters in an urban labor market, and a segment of the workforce that is concentrated in a residential area far from jobs will have a relatively low employment rate. In other words, a spatial mismatch between residence and workplaces causes relatively low employment rates.

A number of studies have explored the spatial-mismatch problem for black workers. Ihlanfeldt and Sjoquist (1990) tested the mismatch hypothesis in 50 metropolitan areas and came to the following conclusions.

- Overall, inferior access to jobs explains about 25 percent of the gap between the employment rates of black and white workers and about 31 percent of the gap between Hispanic and white employment rates.
- The spatial mismatch is more important in larger metropolitan areas. The mismatch explains only 3 percent of the black–white gap in small cities, compared to 14 percent in medium-sized cities and 25 percent in large cities.

A recent study by Hellerstein, Neumark, and McInerney (2008) reveals the subtle interactions between race and space (location). The authors measure job accessibility as the number of jobs in a particular zip code and the adjacent zip codes. Their first conclusion is that, contrary to a casual statement of the spatial mismatch, black workers actually have greater access to jobs than whites. Specifically, the job density (the number of accessible jobs divided by the number of workers) is 0.77 for black workers, compared to 0.73 for white workers. In contrast, the job density of jobs that require only a high-school education is 0.64 for blacks, compared to 0.73 for whites. In other words, black workers have less access than whites to low-skill jobs. This is a problem because, on average, black workers have less human capital. Similarly, for jobs that do not require a high-school education, the job density is 0.50 for blacks, compared to 0.66 for whites. In general, the accessibility problem for black workers occurs for low-skilled jobs, not jobs in general.

The authors look at the relationship between employment rates and job density, testing the notion that higher job density leads to higher employment rates.

1. For jobs in general, an increase in job density did not have a significant effect on the employment rates of blacks.
2. For low-skill jobs (requiring high-school graduation or less), there is a positive relationship between job density and black employment rates. The elasticity of employment with respect to job density is 0.06, meaning that a 10 percent increase in job density increases the employment rate by 0.6 percent.
3. If low-skill jobs are divided into jobs filled by black workers and those filled by other workers, the relationship between job density and employment rate is stronger for the black-worker jobs. For black residents, proximity to jobs filled by blacks is more important than proximity to white-worker jobs.

The authors conclude that race plays a key role in urban labor markets. The problem for black workers is not a lack of nearby jobs, but instead a lack of nearby jobs for which they will be hired. This lack of hiring into nearby jobs could result from discrimination, inferior information networks, or low productivity resulting from weak educational background or unfavorable neighborhood effects.

Moving to Opportunity

The policy experiment called Moving to Opportunity (MTO) was designed to explore the effects of increasing the mobility of low-income households. Under the experiment, several thousand households living in public housing in high-poverty neighborhoods were given housing vouchers to spend on private housing in other neighborhoods. The idea was that the recipients would move to better neighborhoods, with more favorable neighborhood externalities and better access to jobs. What happened?

By all accounts, the MTO program had relatively small effects on the economic circumstances of the recipients. The voucher households moved from very low-income neighborhoods to moderately poor neighborhoods (Quigley and Rafael, 2009). On average, a voucher recipient moved from a neighborhood at the 96th percentile of poverty (96 percent of neighborhoods had higher income) to a neighborhood at the 88th percentile. The new neighborhoods had lower crime rates, and adolescent girls experienced a substantial reduction in "female fear," the term for the fear of sexual harassment and coercion. Similarly, adults reported improved mental health. In contrast, the new neighborhoods were not significantly different in terms of access to jobs. As a result, the voucher households did not experience significant changes in earnings or employment. For youths, test scores were unaffected by moves to new neighborhoods. Also unaffected was youth behavior in aggregate, although girls behaved better, while boys behaved worse.

What are the implications of the MTO experiment? As explained by Rafael and Quigley (2009), the experiment did not generate a sufficiently large change in living conditions to provide a reliable test of either neighborhood effects or the spatial mismatch. They note that the magnitude of the treatment is too small to offset the disadvantages experienced by low-skilled black workers, and so the experiment is uninformative.

SUMMARY

This chapter discusses the conditions that lead households to sort themselves into neighborhoods according to their demand for local public goods, their demand for taxed goods, income, and race. Here are the main points of the chapter:

1. If there is diversity in demand for local public goods, households will sort into municipalities and school districts that provide different levels of the public good.
2. If there is variation in the demand for a locally taxed good, people with relatively high demand for the good will have an incentive to form new municipalities with other high-demand households.
3. Neighborhood externalities occur because neighbors provide role models, share classrooms, and provide information about job prospects.
4. Income segregation occurs when high-income households are willing to pay a bigger premium than low-income households for high-income neighbors.
5. Large-lot zoning promotes income segregation by requiring low-income households to consume more land.
6. Racial segregation increases the frequency of unfavorable outcomes, such as dropping out of school and idleness, and increases poverty.

APPLYING THE CONCEPTS

For exercises that have blanks (_____), fill each blank with a single word or number. For exercises with ellipses (. . .), complete the statement with as many words as necessary. For exercises with words in square brackets ([increase, decrease]), circle one of the words.

1. **Property Tax versus Square-Foot Tax**
   In Metro, there are three types of houses: *E* (expensive), *M* (medium), and *C* (cheap). Each household prefers the same spending on public education. Schools are initially financed with the property tax based on market value, and initially there are three school districts. Suppose the property tax is replaced by a square-foot tax (for example, $2 per square foot of living space per year). The equilibrium number of school districts will decrease to one if. . . because. . . .
2. **How Many School Districts?**
   Households in Metro differ in their demand for public schools: Half are low demanders (*L*) and half are high demanders (*H*). In addition, houses differ in value: Half are expensive (*E*) and half are cheap (*C*). The location choices of households are based on property taxes and spending on public schools.
   a. Normally, we would expect _____ school districts as households sort themselves with respect to _____ and _____.
   b. There will be only two school districts if. . . .

3. **Age Sorting**
   Consider a city of 200 people (100 aged and 100 young) and two neighborhoods (100 people in each). People generally prefer to live close to aged people. To draw the rent-premium curves, put the number of aged people in neighborhood A (from 50 to 100) on the horizontal axis. The premium curve of aged people is concave from below, and in a neighborhood of 100 aged people, the premium is $30. The premium curve of young people is linear, and in a neighborhood of 100 aged people, the premium is $50. The two premium curves intersect at Aged = 70, with the rent premium = $20.
   a. Draw the two premium curves.
   b. Integration (50 aged, 50 young) [is, isn't] a stable equilibrium because . . .
   c. A mixed neighborhood (70 aged, 30 young) [is, isn't] a stable equilibrium because. . . .
   d. Segregation (100 aged, 0 young) [is, isn't] a stable equilibrium because. . . .
   e. Some communities have a minimum age for their residents, for example 50 years old. What is the rationale for minimum age?
   f. Design a policy that would generate an age-segregated community even without an explicit age limit. Show the effects of the policy on your graph.
4. **Integration and Segregation**
   Consider a city of 200 people (100 tall and 100 short) and two neighborhoods (100 people in each). People generally prefer to live close to short people. To draw the rent-premium curves, put the number of short people in neighborhood A (from 50 to 100) on the horizontal axis. The premium curve of tall people is concave from below, and in a neighborhood of 100 short people, the premium is $30. The premium curve of short people is linear, and in a neighborhood of 100 short people, the premium is $50. The two premium curves intersect at S = 70 and premium = $20.
   a. Draw the two premium curves.
   b. Integration (50 short, 50 tall) [is, isn't] a stable equilibrium because. . . .
   c. A mixed neighborhood (70 short, 30 tall) [is, isn't] a stable equilibrium because. . . .
   d. Segregation (100 short, 0 tall) [is, isn't] a stable equilibrium because. . . .
   5. **Allocating Space on the Space Plane**
      Consider a firm that provides rides on a space plane, an aircraft that takes off like an airplane, flies up and away from the earth until it is just about to go into orbit, and then returns to the earth. There will be a single trip, and the total weight limit for passengers is 20,000 pounds. The firm's objective is to maximize its total revenue. There are two income groups, low and high, and the willingness to pay for a trip increases with income.
   a. All the space plane riders will come from the low-income group if low-income people have a larger _____.
   b. Defend your answer by completing the following numerical example: The willingness to pay for a ride is $1,200 for a high-income person, compared to $600 for a low-income person. . . .
   c. So what? Who cares? How is this exercise related to the material in the chapter?

6. Allocating Land in Desirable Communities: MLS

Consider a city that has vacant land (6,000 square feet of space) ready for development in a neighborhood that is desirable because of its superior schools. There are two types of households, low income (*L*) and high income (*H*). Each *L* is willing to pay $40,000 to live in the neighborhood, and each *H* is willing to pay $60,000. The preferred lot size is 1,000 square feet for type *L*, compared to 3,000 square feet for type *H*.

a. If there are no restrictions on lot sizes, the vacant land will be allocated to type [*L, H*] because. . . .

b. Suppose the city sets a minimum lot size. The vacant land will be allocated to type *H* if the minimum lot size is at least _____ square feet because. . . .

7. Teacher Isoquant and Cost Minimization

Consider a teacher isoquant, which shows different input combinations that produce a target quantity of output. Suppose the inputs for teacher productivity are graduate course work (on the horizontal axis, ranging from 0 to 10 years) and verbal ability (on the vertical axis, measured as the teacher's SAT verbal score). Under the typical pay structure, an additional year of graduate course work increases teacher pay by about $4,000. Suppose that a one-unit increase in verbal ability increases teacher pay by $500.

a. Use the empirical evidence cited in the book to draw an education isoquant. The marginal rate of technical substitution between graduate education and verbal ability is _____ SAT points per year of graduate coursework.

b. Suppose the MRTS is small, but positive, with a value of 2 SAT points per year of graduate course work. Draw an isoquant and identify the cost-minimizing combination of verbal ability and graduate course work. To minimize the cost of a given output level, a school district should hire teachers with _____ years of graduate course work.

8. Equalizing Educational Achievement

The objective of the state of Egalitaria is to equalize educational achievement across schools, where school achievement equals the average test score of students in the school. Comment on the following: "The state can realize its objective by equalizing expenditures across schools, by spending the same amount per student in each school."

9. Multiple Equilibria with Race

Using Figure 8–5 as a starting point, suppose the premium curve for black households is an inverted U and lies above the white curve for a white population in the range 50–60 whites, but below the white curve for a larger white population. In other words, the two premium curves intersect at 60 white households.

a. Draw the premium curves.

b. Integration (50 white, 50 black) [is, isn't] a stable equilibrium because. . . .

c. A mixed neighborhood (60 white, 40 black [is, isn't] a stable equilibrium because. . . .

d. Segregation (100 white, 0 black) [is, isn't] a stable equilibrium because. . . .

10. Commuting and Employment Rate

Consider the following statement from the book's discussion of the spatial mismatch: "One reason for the relatively low employment rate (for blacks) is that the average commute time of blacks was 26 minutes, compared to only 19 minutes for whites."

a. Use a supply-demand graph of the urban labor market to show the economic logic of this statement. The vertical axis measures the monetary wage paid by the employer. Draw a single demand curve, relevant for both white and black workers, and two supply curves, one for whites (labeled S19) and one for blacks (labeled S26).

b. Arrows up or down: An increase in commuting time from 19 minutes to 26 minutes _____ the supply of labor, which _____ the equilibrium wage and _____ the equilibrium quantity of labor.

REFERENCES AND ADDITIONAL READING

1. Becker, G., and K. Murphy. *Social Economics.* Cambridge: Harvard University Press, 2000.
2. Card, David, and Alan Krueger. "Labor Market Effects of School Quality: Theory and Evidence." In *Does Money Matter?* ed. Gary Burtless. Washington DC: Brookings, 1996.
3. Cohen, M., R. Rust, S. Steen, and S. Tidd. "Willingness-to-pay for Crime Control Programs." *Criminology* 42 (2004), pp. 89–109.
4. Cullen, J. B., and S. D. Levitt. "Crime, Urban Flight, and the Consequences for Cities." *Review of Economics and Statistics* 81 (1999), pp. 159–69.
5. Cutler, David M., and Edward L. Glaeser. "Are Ghettos Good or Bad?" *Quarterly Journal of Economics* (1997), pp. 827–72.
6. Durlauf, Steven. "Neighborhood Effects." Chapter 50 in *Handbook of Regional and Urban Economics 4: Cities and Geography,* eds. Vernon Henderson and Jacques-Francois Thisse. Amsterdam: Elsevier, 2004.
7. Gabriel, Stuart A., and Stuart S. Rosenthal. "Household Location and Race: Estimates of a Multinomial Logit Model." *Review of Economics and Statistics,* 1989, pp. 240–49.
8. Heckman, James, Anne Layne-Farrar, and Petra Todd. In "Does Measured School Quality Really Matter?" Chapter 7 in *Does Money Matter?* ed. Gary Burtless. Washington DC: Brookings, 1996.
9. Hellerstein, Judith, David Neumark, and Melissa McInerney. "Spatial Mismatch or Racial Mismatch." *Journal of Urban Economics* 64 (2008), pp. 464–79.
10. Ihlanfeldt, Keith R., and David L. Sjoquist. "Job Accessibility and Racial Differences in Youth Employment Rates." *American Economic Review* 8 (1990), pp. 267–76.
11. Institute on Race and Poverty. *Examining the Relationship between Housing, Education, and Persistent Segregation: Final Report.* Minneapolis: Institute on Race and Poverty, 1998.
12. Kain, John F. "Housing Segregation, Negro Employment, and Metropolitan Decentralization." *Quarterly Journal of Economics* 82 (1968), pp. 175–97.

13. Krueger, Alan. “Experimental Estimates of Education Production Functions.” *The Quarterly Journal of Economics* 114 (1999), pp. 497–532.
14. Krueger, Alan, and Diane Whitmore. “The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR.” *The Economic Journal* 111 (2001), pp. 1–28.
15. Meldrum, Christina, and Susan Eaton. “Resegregation in Norfolk, Virginia: Does Restoring Neighborhood Schools Work?” Report of the Harvard Project on School Desegregation (1994).
16. Miller, T., M. A. Cohen, and B. Wiersema. *Victim Costs and Consequence—A New Look*. Washington DC: National Institute of Justice, 1996.
17. Mills, Edwin S., and Luan Sende Lubuele, “Inner Cities.” *Journal of Economic Literature* 35 (1977), pp. 727–56.
18. O’Regan, Katherine M., and John M. Quigley. “Cars for the Low-income.” *Access* 12 (Spring 1998), pp. 20–25.
19. Orfield, Gary. “Desegregation, Resegregation, and Education.” Part I of *In Pursuit of a Dream Deferred: Linking Housing and Education*. Minneapolis: Institute on Race and Poverty, 1995.
20. Popkin, Susan, Tama Leventhal, and Gretchen Weismann. “Girls in the ‘Hood: How Safety Affects the Life Chances of Low-Income Girls.” *Urban Affairs Quarterly* 45 (2010), pp. 715–44.
21. Quigley, John, and Steven Rafael. “Neighborhoods, Economic Self-Sufficiency, and the MTO Program.” Brookings-Wharton Papers on Economic Affairs 2008. Washington DC: Brookings, 2009.
22. Rafael, Steven. “The Spatial Mismatch Hypothesis and Black Youth Joblessness: Evidence from the San Francisco Bay Area.” *Journal of Urban Economics* 43 (1998), pp. 79–111.
23. Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. “Teachers, Schools, and Academic Achievement.” NBER Working Paper Number 6691 (1998).
24. Schiller, Bradley. *The Economics of Poverty and Discrimination,* 6th ed. Englewood Cliffs, NJ: Prentice-Hall, 1995.
25. Thaler, Richard. “Econometric Analysis of Property Crime—Interaction between Police and Criminals.” *Journal of Public Economics* 8 (1977), pp. 37–51.
26. Vigdor, Jacob. “Race: The Perplexing Persistence of Race.” Chapter 7 in *Making Cities Work*, ed. Robert P. Inman. Princeton NJ: Princeton University Press, 2009.
27. Yinger, John. “Evidence on Discrimination in Consumer Markets.” *Journal of Economic Perspectives* 12, no. 2 (Spring 1998), pp. 23–40.

CHAPTER 9

Zoning and Growth Controls

Nimby (not in my back yard) is for wimps. The new acronym is Banana: build absolutely nothing anywhere near anybody.

—*The Economist* 17 April 1993

A tranquil city of good laws, fine architecture, and clean streets is like a classroom of obedient dullards, whereas a city of anarchy is a city of promise.

—Mark Helprin

This chapter explores the government's role in the urban land market. So far we have assumed that land is allocated to the highest bidder. In fact, cities regulate land use, using land-use zoning plans to segregate different types of land use—commercial, industrial, and residential—into separate zones. Residential zones are typically divided into separate zones for low-density and high-density housing. We'll explore the causes and consequences of zoning, focusing on who wins and who loses.

Municipalities use a variety of policies to limit their population growth. Some cities tax new development to discourage growth, and others limit the number of building permits. Some cities limit the amount of land that can be developed, either by limiting the extent of urban services such as roads, sewers, and water or by establishing an urban growth boundary. In the Portland area, a metropolitan authority oversees an urban growth boundary over the entire metropolitan area. We'll explore the trade-offs associated with growth controls, showing who wins and who loses.

LAND-USE ZONING

Urban land-use zoning dates back to 1870 in Germany and was first implemented in the United States in New York City in 1916. The basic idea of zoning is to separate land uses that are "incompatible" in some sense. As we'll see, local governments have adopted a flexible definition of "incompatible."

The Early History of Zoning

Fischel (2004) summarizes the history of land-use zoning in the United States. Before comprehensive zoning, many cities used ordinances to control land use in specific areas. For example, to address concerns that skyscrapers would block views and light, cities regulated tall buildings. New York City implemented the first comprehensive zoning plan in 1916, and eight other cities adopted zoning plans in the same year. By 1936, zoning had spread to over 1,300 cities.

Why didn't zoning develop earlier? Fischel argues that the urban transportation technology of the late 19th and early 20th centuries made zoning unnecessary, at least from the perspective of suburban homeowners. As we saw earlier in the book, manufacturers transported their output on horse-drawn wagons, a slow and expensive mode that required firms to locate close to the city's central port or railroad terminal. The main form of public transit was the hub-and-spoke streetcar system. Low-income households lived in apartments close to the city center or along the spokes of the streetcar system. Commercial activities and apartments located along the streetcar lines, generating neighborhoods with mixed land use. Most homeowners lived a few blocks from the streetcar lines—inside the spokes—in neighborhoods separated from industry, commerce, and apartments. Homeowners placed a high value on their quiet, low-density neighborhoods and organized to prevent the extension of streetcar lines that would disturb their peace.

Innovations in transportation increased the location options for business, setting the stage for industrial zoning. Before the intracity truck, the externalities generated by industry and commerce (pollution, noise, congestion) were confined to the central areas of the city, far from the homes of suburban homeowners. The intracity truck (which dates to 1910), allowed firms to move away from the city's central export node and closer to their suburban workers. Once firms became more footloose, cities implemented zoning to separate industry from homes. A headline from the *New York Times* in 1916 read "Zoning Act Removes Fear of Business Invasion."

Innovations in mass transit increased the location options for workers, setting the stage for residential zoning. The motorized passenger bus, developed in 1920, allowed low-income workers to live between the spokes of the streetcar systems, where homeowners had been insulated from high-density housing. Cities developed residential zoning to keep apartments out of homeowner neighborhoods. In the leading case on the constitutionality of zoning (*Euclid* v. *Ambler* [1924]), U.S. Supreme Court Justice Sutherland wrote that apartments are "a mere parasite, constructed in order to take advantage of the open spaces and attractive surroundings created by the residential character of the district."

Zoning as Environmental Policy?

Zoning can be used to separate pollution sources from residential areas. Recall the third axiom of urban economics:

Externalities cause inefficiency

Industrial firms generate all sorts of externalities, including noise, glare, dust, odor, vibration, and smoke. Zoning is appealing as an environmental policy because it is simple: The easiest way to reduce the exposure to pollution is to put a buffer between a polluter and its potential victims. The problem with zoning as an environmental policy is that it doesn't reduce pollution to its socially efficient level but simply moves it around.

The economic approach to pollution is to impose a tax on pollution equal to the external cost it imposes on society. A pollution tax forces a firm to pay for pollution in the same way that it pays for raw materials, capital, and labor. As a result, a firm has an incentive to economize on pollution in the same way that it economizes on raw materials and labor. In other words, a pollution tax reduces pollution to its socially efficient level.

What would happen if we simply replaced zoning with a pollution tax that didn't vary across space? We would expect some polluting firms to move closer to residential areas, so the exposure to pollution in some areas would increase. One approach would be to combine pollution taxes with zoning: By placing polluters in industrial zones and also imposing a tax, pollution could be reduced to the socially efficient level and exposure could be controlled. In fact, if zoning reduces the exposure to pollution, it also decreases the external cost of pollution and thus the pollution tax.

Retailers generate a number of externalities that affect nearby residents. Traffic generates congestion, pollution, noise, and parking conflicts. A traditional zoning plan deals with these externalities by confining retailers to special zones. A more flexible approach is to give retailers more location options while enforcing performance standards for parking, traffic, and noise. For example, a city can require retailers to provide adequate off-street parking, pay for improvements in the transportation infrastructure to handle extra traffic, and design the retail site to control noise and other externalities.

High-density housing may generate externalities. Like retailing activity, apartment and condominium complexes increase traffic, causing noise, congestion, and parking problems. In addition, tall buildings may block views and sunlight. An alternative to traditional zoning, which simply bans high-density housing in certain areas, is to use performance standards to prevent these externalities. Traffic problems can be prevented by street improvements, and parking problems can be avoided by requiring off-street parking. In addition, buildings can be designed to deal with the issues of lost views and sunlight.

Fiscal Zoning

Another motivation for zoning is to ensure that households or firms generate a fiscal surplus, not a deficit. A fiscal surplus occurs when a land user's tax bill exceeds the cost of public services provided. For example, a large retailer may pay a lot in property and sales taxes but not require much in local public services. If a commercial or industrial land user generates a fiscal surplus, the surplus will at least partly offset any negative externalities such as noise, traffic, or pollution. Some communities eagerly host firms and use the fiscal surplus to cut tax rates and spend more on local public services. In general, low-income communities more frequently make this trade-off between environmental quality and fiscal benefits (Evenson and Wheaton, 2003).

A fiscal deficit occurs when a firm or household pays less in taxes than it gets in public services. Local governments get about three-fourths of their tax revenue from the property tax, so a household's local tax liability is determined largely by the value of its house or apartment. A household's use of local public services such as education, recreation, and public safety depends in part on the number of people in the household. A large household in a small dwelling is more likely to generate a fiscal deficit for the local government.

One way to decrease the likelihood of fiscal deficits is to zone for minimum lot sizes. Housing and land are complementary goods, and in general the larger the lot, the higher the market value of the property (dwelling and land combined). A minimum lot size excludes some households that would generate fiscal deficits. For example, suppose a city breaks even on a family living in a house worth \$200,000: The family's tax liability equals the cost of providing public services. One rule of thumb is that the market value of a property (for land and dwelling) is about five times the value of land. Therefore, requiring a household to live on a lot worth at least \$40,000 is roughly equivalent to requiring the household to live in a \$200,000 house.

We can use a simple formula to determine the minimum lot size. Given the rule of thumb that the property value is five times the land value,

$$v^* = 5 \cdot r \cdot s$$

where v^* = the target (breakeven) market value of the property, r is the price of land per acre, and s is the lot size (in acres). Rearranging the expression, we can solve for the target lot size:

$$s = \frac{v^*}{5 \cdot r}$$

For example, if the target value is \$200,000 and the price of land is \$80,000 per acre, the target lot size is 1/2 acre:

$$s = \frac{\$200{,}000}{5 \cdot \$80{,}000} = 0.50$$

Minimum Lot Zoning and the Space Externality

Another motivation for minimum lot-size zoning is to internalize an externality associated with residential space (Evenson and Wheaton, 2003). People value space between houses, and a larger lot size means more space between houses and higher utility for everyone in the neighborhood. Your neighbors benefit from your lot-size decision but don't pay for it, so there is an externality. Recall the third axiom of urban economics:

Externalities cause inefficiency

In this case, the positive space externality means that lots will be smaller than the socially efficient size. In picking a lot size, people ignore the benefits experienced by their neighbors, so lots are too small.

One response to the space externality is minimum-lot zoning. It establishes a minimum space between houses, with equal contributions of space by each household. For example, if zoning has the effect of establishing a 100-foot gap between houses, each household buys 50 feet of space between its house and the property line. In addition to increasing lot size, minimum-lot zoning enforces reciprocity in space decisions.

Provision of Open Space

Local governments provide open space in two ways. First, they provide public land in parks and greenbelts. Second, they use zoning to restrict the use of private land. For example, zoning may prevent agricultural land from being subdivided into small parcels for residential or commercial development.

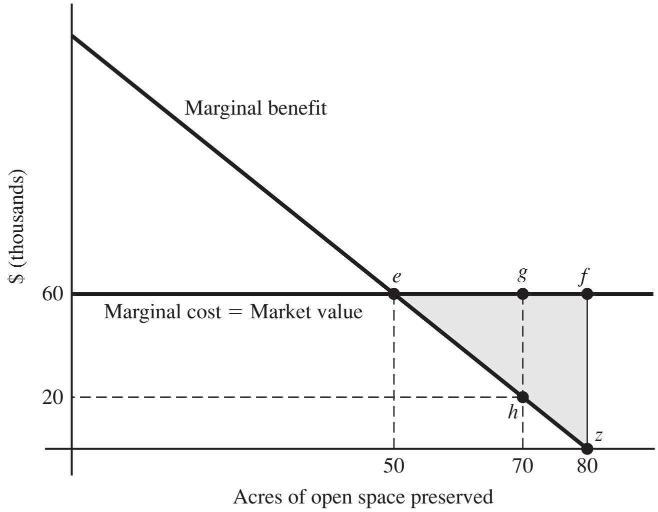
Is zoning for open space efficient? The alternative is to purchase the land for public use. When a city pays for open space in the same way that it pays for fire trucks and schools, it bears the full cost of providing it. As a result, the city weighs the costs and benefits of open space and picks the socially efficient quantity. In contrast, when a city simply zones land as open space, the cost of the public good is shifted to the property owner. The government and voters face less than the full cost of open space and have an incentive to provide too much of it.

We can use Figure 9–1 to illustrate the inefficiency of open-space zoning. The marginal-benefit curve is negatively sloped, reflecting the assumption that citizens' willingness to pay for open space diminishes as the amount of space increases. The opportunity cost of open space is the market value of land, the amount that developers are willing to pay for it. In Figure 9–1, the market value of land is $60,000 per acre, and that's the marginal cost of open space.

We can use the marginal principle to show the socially efficient quantity of open space. The marginal principle is reviewed in Section 1.1 of "Tools of Microeconomics," the appendix at the end of the book. The socially efficient quantity is the quantity at which the marginal benefit equals marginal cost and in this case is 50 acres (point *e*). If the goverment buys the land at the market price, we reach the efficient point *e*. The city outbids developers for 50 acres, so that's how much open space the city buys.

Suppose the city zones land for open space and does not compensate landowners for the loss in market value resulting from the restriction on land use. From the perspective of the city and its voters, the marginal cost of open space is now zero. The rational response is to pick the quantity at which the marginal benefit of open space equals zero. In Figure 9–1, the city chooses point *z*, with 80 areas of open

FIGURE 9–1 Open-Space Zoning



The marginal cost of open space is the market value of land in an alternative use. Point *e* shows the socially efficient quantity of open space, where the marginal benefit equals the marginal cost. If land is simply zoned for open space without compensation, the city will choose point *z*. The shaded triangle shows the deadweight loss.

space. When the government does not pay the market value of open space, it provides more than the socially efficient quantity.

We can use the logic of the marginal principle to compute the welfare loss from open-space zoning. As reviewed in Section 1.2 of "Tools of Microeconomics," the appendix at the end of the book, the surplus associated with an activity is shown by the area between the marginal-benefit and marginal-cost curves up to the chosen quantity. The open-space zoning policy goes too far, and the welfare loss from the policy is shown as the shaded triangle *efz*. For example, the 70th acre has a cost of $60,000 and a benefit of only $20,000, so the loss from using the acre as open space is $40,000. To compute the welfare loss from the policy, we do the same calculation for the other 29 acres of excess open space and add the losses for each acre to get the welfare-loss triangle.

THE LEGAL ENVIRONMENT OF ZONING

Local governments are creatures of state governments and derive their power to control land use from the states. In most states, the enabling legislation is patterned after the Standard State Zoning Enabling Act, developed by the U.S. Department of Commerce in 1926. Section 1 of the act states the following:

> Grant of Power. For the purpose of promoting health, safety, morals, or the general welfare of the community, the legislative body of cities and incorporated villages is hereby

> empowered to regulate and restrict the height, number of stories, and size of buildings and other structures, the percentage of the lot that may be occupied, the size of yards, courts, and other open spaces, the density of population, and the location and use of buildings, structures, and land for trade, industry, residence, or other purposes.

Zoning is considered a legitimate exercise of the police power of local government if it promotes the public health, safety, and welfare.

Current zoning laws are the result of over 60 years of legal decisions. In the last six decades, individuals affected by specific zoning laws have sued local governments, forcing state and federal courts to rule on the constitutionality of zoning ordinances. If a particular type of zoning is declared unconstitutional, all cities get the message from the courts and rewrite their zoning ordinances to drop the illegal practices. On the other hand, if a zoning practice is upheld as constitutional, the practice spreads to other local governments. Court decisions have established three criteria for the constitutionality of zoning: substantive due process, equal protection, and just compensation.

Substantive Due Process

The case of *Euclid* v. *Ambler* (1924) established the standards for **substantive due process.** According to the due-process criterion, zoning must be executed for a legitimate public purpose using reasonable means. In the early 1920s, the city of Euclid, Ohio, enacted a zoning ordinance that restricted the location, size, and height of various types of buildings. Ambler Realty had purchased some property between the railroad tracks and a major thoroughfare and expected to sell the land to an industrial developer. When the city zoned its land for residential use, Ambler sued. The Supreme Court ruled against Ambler, concluding that the zoning ordinance satisfied the standards for substantive due process because it had some “reasonable relation” to the promotion of “health, safety, morals, and general welfare.” In other words, zoning to separate different land uses is a legitimate use of the city’s police power because it promotes public health and safety.

One interpretation of the *Euclid* v. *Ambler* decision is that a zoning ordinance is constitutional as long as it generates some benefit for the local community. The court did not say that the benefit of zoning must exceed its cost, only that the benefit must be positive. Fischel (1985) calls this benefit analysis, as opposed to benefit-cost analysis. The court defined the possible social benefits from zoning in broad terms, to include monetary, physical, spiritual, and aesthetic benefits.

As an example of the use of the benefit criterion to justify zoning, consider San Francisco’s treatment of its Chinese population. In the 1880s, the city passed laws that explicitly segregated the Chinese. When explicit segregation was declared unconstitutional, the city passed a zoning law that banned laundries from certain neighborhoods. The zoning law did not violate the Constitution because it promoted public welfare by keeping an undesirable land use (laundries) out of some residential areas. Because the Chinese operated most of the city’s laundries, the zoning law provided a legal means of ethnic segregation.

Equal Protection

The equal-protection clause of the Fourteenth Amendment requires that all laws be applied in an impersonal (nondiscriminatory) fashion. Zoning is exclusionary in the sense that it excludes some types of people from a city, for example, people who live in apartments instead of single-family dwellings. The plaintiffs in recent court cases have argued that zoning laws violate the equal-protection clause because they are not applied in an impersonal manner, but instead treat some people differently than others.

The federal courts have upheld the constitutionality of exclusionary zoning. In the *Euclid* v. *Ambler* decision, the Supreme Court suggested that although a zoning ordinance must generate some benefit for insiders (citizens of the community), the effects of zoning on outsiders are unimportant. In *Warth* v. *Selden* (1975), the court dismissed the claims of outsiders because they did not prove that the zoning ordinance caused specific personal damage. In *Village of Arlington Heights* v. *Metropolitan Housing Corporation* (1977), the court dismissed the claims of outsiders because they did not prove discriminatory intent on the part of zoning officials. In *Ybarra* v. *Town of Los Altos Hills,* the court ruled that although zoning laws that discriminate on the basis of race are unconstitutional, zoning laws that discriminate on the basis of income are legal. In general, the federal courts have adopted a noninterventionist approach to exclusionary zoning.

Some state courts have adopted a more activist role. In *Southern Burlington County NAACP* v. *Mount Laurel* (1975), the New Jersey Supreme Court ruled that Mount Laurel's exclusionary zoning harmed low-income outsiders. The court directed the city to develop a new zoning plan under which the city would accommodate its "fair share" of low-income residents. The court established quotas for communities to provide enough housing for low- and moderate-income workers to live within reasonable commuting distances of their jobs. The effects have been minor, in part because the state legislature modified the quotas and even allowed communities to buy and sell up to half their quotas (Mills and Lubuele, 1997).

Other states have also ruled on exclusionary zoning. The implication from *Associated Home Builders Inc.* v. *City of Livermore* (California Supreme Court, 1976) is that the courts will judge zoning on the basis of its effects on both insiders and outsiders. If a zoning ordinance does not represent a reasonable accommodation of the competing interests of insiders and outsiders, it may be declared unconstitutional. In Oregon, state law requires municipalities to plan and zone land for a diversity of housing types and income levels.

Just Compensation

The third criterion for the constitutionality of zoning is just compensation. The Fifth Amendment states ". . . nor shall private property be taken for public use, without just compensation." This is the taking clause: If the government converts land from private to public use, the landlord must be compensated. Most zoning ordinances do not actually convert land to public use, but merely restrict private use. For example, industrial zoning prevents a landowner from building a factory in a residential area, and minimum lot-size zoning prevents high-density housing. By restricting the use of private land, zoning decreases the market value of the property.

The policy issue is whether landowners should be compensated for the loss of property value caused by zoning. For example, if large-lot zoning decreases a landowner's property value by $5,000, should the local government pay $5,000 in compensation? According to Fischel (1985), the courts have provided mixed and confusing signals to local zoning authorities. The courts routinely uphold zoning laws that cause large losses in property values, suggesting that as long as the landowner is left with some profitable use of his land, compensation is not required. The courts have developed several rules to determine whether compensation is required.

1. **Physical invasion.** Compensation is required if the government physically occupies the land. The invasion rule is applicable only when the government actually occupies the land. It does not apply to most zoning actions in which the government merely restricts private land use.
2. **Diminution of value and reasonable beneficial use.** The origin of this rule is *Pennsylvania Coal* v. *Mahon* (1922), in which Justice Holmes states, "… while property may be regulated to a certain extent, if regulation goes too far it will be recognized as a taking." In other words, compensation is required if zoning decreases the property value by a sufficiently large amount. Unfortunately, the courts have not indicated how far zoning must go before compensation is required. A related rule is reasonable beneficial use: If zoning leaves the landowner with options that provide a reasonable rate of return, no compensation is required.
3. **Harm prevention.** According to this rule, compensation is not required if the zoning ordinance prevents a harmful use of the land. In other words, zoning is not a taking if it prevents the landowner from using land in ways that are detrimental to the general public. The harm-prevention rule suggests that a landowner, like a car owner, has limited property rights. The car owner has the right to drive her car, but she must stop at red lights. Should the driver be compensated for the opportunity cost of time spent waiting for the light to turn green? Since the traffic lights prevent a harmful use of the car, compensation is not required. Similarly, landowners have limited property rights: If zoning prevents the landlord from building a polluting factory in a residential district, compensation is unnecessary because zoning prevents a harmful use of the land. Most zoning ordinances are judged by a broad interpretation of the harm-prevention rule: If an ordinance promotes public health, safety, or welfare, compensation is usually not required.

A CITY WITHOUT ZONING?

What would an unzoned city look like? Would glue factories and pizza parlors invade quiet residential neighborhoods? Would land use be disorderly and ugly? Some tentative answers to these questions come from Siegan's (1972) analysis of

Houston, the only metropolitan area in the United States without zoning. Land use in the city is controlled by restrictive covenants, voluntary agreements among landowners that limit land uses and structures. The covenants governing residential subdivisions (over 7,000 in number) are typically more strict than conventional zoning. They have detailed restrictions on architectural design, external appearance, and lot maintenance. The covenants for industrial parks limit the activities that can locate in the park.

How does Houston compare to zoned cities? Although a rigorous comparison of land-use patterns may be impossible, some tentative observations can be made:

1. **Dispersion of industry.** The spatial distribution of Houston's industrial firms is similar to that of zoned cities. As in other cities, Houston's industrial firms locate close to the transportation network (near railroads and highways) and tend to cluster to exploit localization economies.
2. **Retailers.** Few retailers locate in quiet residential neighborhoods. Like retailers in most cities, most retailers in Houston locate along major thoroughfares in strip developments and shopping centers to take advantage of large volumes of foot and auto traffic.
3. **Strip development.** Houston appears to have more strip development (retail and commercial establishments along arterial routes) than do zoned cities.
4. **Apartments.** Low-income housing is more plentiful and relatively inexpensive. There is a wide range of densities in apartment projects; the projects occupied by the wealthy have more open space and lower density, while the projects occupied by the poor have higher density.

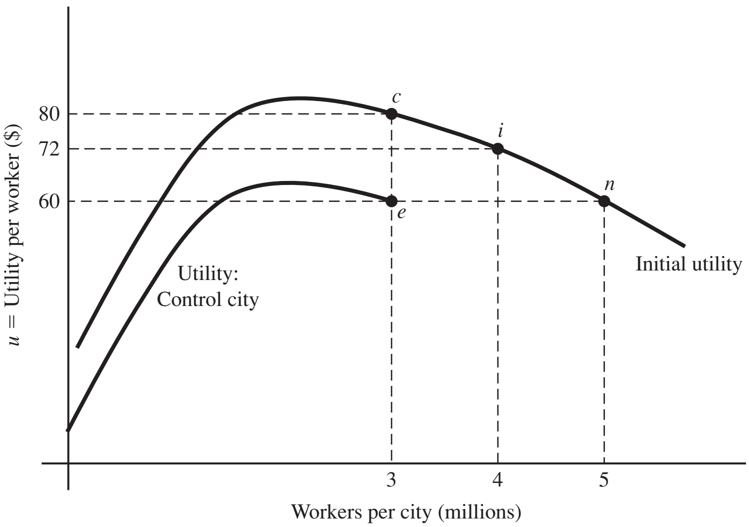
There are two lessons from Houston's experience without zoning. First, in the absence of zoning, landowners have the incentive to negotiate restrictions on land use. It seems that neighborhood externalities are large enough to justify the cost of developing and enforcing restrictive covenants. This is the Coase solution to externalities (named after Ronald Coase): The parties affected by externalities negotiate a contract to solve the externality problem. Second, in the absence of zoning, most industrial firms cluster in locations accessible to the transportation network, and most retailers cluster in shopping centers and retail strips.

GROWTH CONTROL: URBAN GROWTH BOUNDARIES

Cities use a number of policies to restrict the amount of developed land and thus their population. A city can outlaw development beyond a growth boundary or restrict urban services such as roads, water, and sewers to certain areas. A survey completed in 1991 found that roughly a quarter of cities used urban service boundaries to limit their land areas. In this part of the chapter, we explore the consequences of these sorts of land restrictions.

We can use the utility curve developed in Chapter 4 to show the effects of growth boundaries on cities in a regional economy. We'll use the theoretical framework developed by Helsley and Strange (1995). Suppose we start with two identical

FIGURE 9–2 Effects of Precise Growth Control in a Two-City Region



The initial equilibrium is shown by point *i*: Each of the two cities has a population of 4 million. A growth-control policy that reduces the population of the control city to 3 million (point *c*) and increases the population of the uncontrolled city to 5 million (point *n*), opens a utility gap of \$20 (= \$80 − \$60). The resulting increase in the price of land in the control city shifts the city's utility curve downward until equilibrium is restored at points *e* and *n*.

cities in a regional economy. To simplify matters, we will assume that all city residents are renters, and land is owned by absentee landowners. Later in the chapter we'll explore the implications of changing this assumption. Figure 9–2 shows the initial utility curve for the typical city. Point *i* shows the initial equilibrium with 4 million workers in each city. The utility level is \$72 in each city, so we have a locational equilibrium: No worker has an incentive to change cities.

Precise Growth Control: Limiting Land Area and Lot Size

Consider first the effects of precise growth control. Suppose one city caps the number of workers in the city by (a) specifying a minimum lot size per person and (b) fixing the total land area of the city. If the workforce of the controlled city is capped at 3 million, the workforce of the uncontrolled city must increase to 5 million to accommodate the fixed workforce of the region. Growth control in one city displaces workers to the other city.

The immediate effect of the policy is to generate a utility gap between the two cities. In Figure 9–2, utility in the control city rises from \$72 (point *i*) to \$80 (point *c*) as population decreases. In contrast, the increase in population in the other city decreases utility from \$72 (point *i*) to \$60 (point *n*). Workers are mobile between the two cities, so this utility gap will not persist. Recall the first axiom of urban economics:

Prices adjust to generate locational equilibrium

In this case, the workers of the region compete for a fixed number of lots in the control city, and the price of land in the control city will rise until the two cities have the same utility level, making workers indifferent between the two cities.

The increase in land rent in the control city shifts its utility curve downward. Recall that workers are renters, not property owners. An increase in land rent decreases the amount of money available to spend on consumer products, so the utility level drops. How low will utility go? Given the cap of 3 million workers in the control city, the uncontrolled city has a workforce of 5 million and a utility level of \$60 (point *n*). In other words, the common (regionwide) utility level is anchored by point *n*. Land rent in the control city must rise to the point at which utility is \$60. This is shown by point *e* on the lower utility curve for the control city. In the new locational equilibrium, the benefits of living in a smaller city (shorter commutes and less noise, dirt, and congestion) are fully offset by higher land rent.

Our analysis of growth control could be applied to a region with more cities. If the control city is one of 11 cities in a region rather than one of two, the workers displaced by growth control would be spread among 10 other cities. Each uncontrolled city would experience a smaller increase in its workforce (100,000 instead of 1 million) and thus a smaller decrease in utility. Spreading out the displaced workers would mean a smaller utility loss per capita. This is sensible because when the control city is a small part of the regional economy, the per-capita effect of its policy will be relatively small.

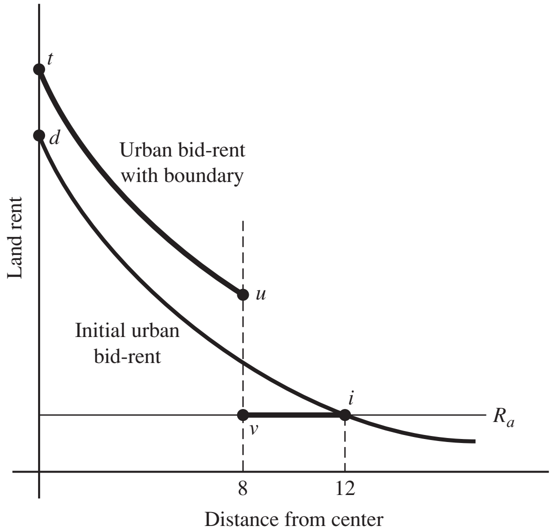
Winners and Losers from Growth Boundaries

The growth-control policy decreases the utility of workers throughout the region. The workers in the uncontrolled city lose because their city grows, moving further downward along the negatively sloped utility curve. Both cities are initially too big, and the uncontrolled city moves even further from the optimum size. The workers in the controlled city lose too because locational equilibrium generates a common utility level, and utility in the control city is dragged down by the lower utility in the uncontrolled city.

The decrease in the common utility level reflects the basic inefficiency of replacing two identical cities with a large city and a small one. Recall that the immediate effect of the growth-control policy is shown by points *c* and *n*. Utility is higher in the smaller city (\$80, compared to \$60), so an efficient policy would move workers from the larger city to the smaller one. The growth-control policy prevents the efficient movement of workers, so the common utility is lower than it would be with two cities of equal size.

Consider next the effects of the growth boundary on landowners in the control city. In Figure 9–3, the thin negatively sloped curve is the initial urban bid-rent curve, a composite of the residential and business bid-rent curves. The initial

FIGURE 9–3 Urban Growth Boundary and the Land Market



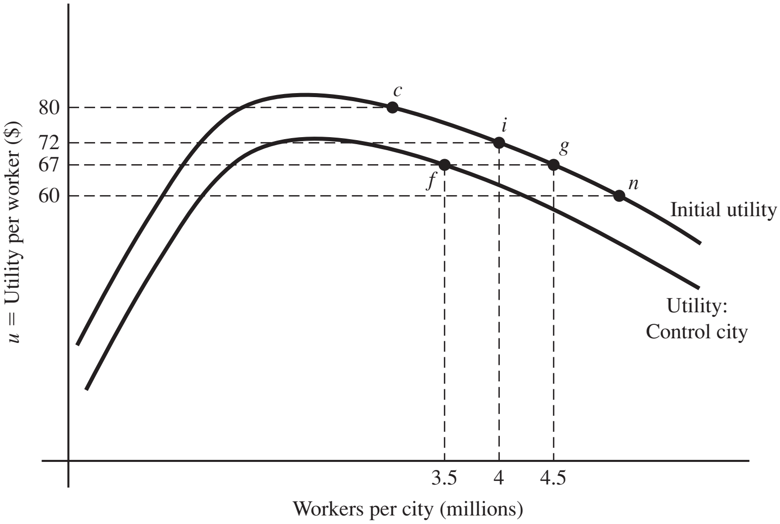
The initial equilibrium is shown by point *i*: The urban bid-rent curve intersects the agricultural bid-rent curve at 12 miles. An urban growth boundary at eight miles from the center increases urban rent within eight miles (the curve connecting points *t* and *u*) and decreases rent between eight and 12 miles (the line connecting points *v* and *i*).

equilibrium is shown by point *i*, where the urban bid-rent curve intersects the horizontal agriculture bid-rent curve. The initial radius of the city is 12 miles. A growth boundary at eight miles cuts off the last four miles from urban development, so rent in this area drops to the agriculture rent (shown as the horizontal line between points *v* and *i*). Obviously, landowners who own land just outside the boundary are losers. In contrast, people who own land inside the boundary are winners. As we saw earlier, competition between workers for the fixed number of lots in the controlled city bids up land rent. In Figure 9–3, the urban bid-rent curve shifts upward to the thick line, indicating higher rent on land inside the boundary.

Urban Growth Boundary and Density

So far, we have considered a growth-control policy that precisely controls a city's population by limiting both the land area and lot size of the city. Suppose the city uses a growth boundary but does not restrict lot size. The immediate effects of a growth boundary are the same as the effects of the earlier policy, as shown in Figure 9–2 and reproduced in Figure 9–4 (page 243). We go from point *i* (4 million workers in each city) to points *c* in the growth-control city (3 million workers) and

FIGURE 9–4 Effects of an Urban Growth Boundary in a Two-City Region



The initial equilibrium is shown by point i: Each of the two cities has a population of 4 million. If the control city does not limit lot size, the increase in the price of land resulting from a growth boundary will increase density, generating a population of 3.5 million (point f). Regional equilibrium is restored at points f and g, with a common utility level of $67 and a total of 8 million people in the two cities, with 3.5 million in the control city and 4.5 million in the other city.

n in the other city (5 million workers). The utility gap is $20 (equal to $80 – $60). What happens next?

As before, the utility gap increases the price of land in the control city. When the price of land increases, density will increase as firms and people economize on land. Single-family homes will be built on smaller lots, and more people will live in high-density apartments and condominiums. Firms will occupy smaller lots and taller buildings. This increase in density weakens the growth boundary as a population-control policy: Higher density will partly offset the loss of urban land. In Figure 9–4, the new equilibrium is shown by points f and g, with 3.5 million workers in the growth-boundary city and 4.5 million in the other city. The new utility level is $67, compared to $60 with the earlier policy. Utility is higher with flexible density because the distortionary effects of growth control can be partially "undone" by increases in density.

Since an urban growth boundary increases density, is it an appropriate response to the problem of urban sprawl? As we saw earlier in the book, a number of public policies contribute to sprawl (low density) in U.S. cities, including the underpricing of travel and local public services, subsidies for housing, and exclusionary zoning. The efficient response to these distortions is to eliminate them directly, allowing individuals to make socially efficient location and density decisions. A growth boundary is a blunt tool to deal with the distortions for two reasons.

1. Although a growth boundary may change density in the correct direction, it may go too far or not far enough.
2. A growth boundary creates distortions of its own.

Portland's Urban Growth Boundary

The urban growth boundary in Portland, Oregon, is a metropolitan boundary that is periodically expanded to accommodate growth. By law, there must be a 20-year supply of vacant land within the boundary, which was expanded by 4,000 acres in 1998 and by 1,940 acres in 2004. The recent expansion was implemented to increase the supply of land suitable for industrial development.

The Portland growth boundary differs from the growth-boundary policy we have discussed in two important respects. First, the Portland boundary is extended as the population of the metropolitan area increases. Second, the boundary is combined with a number of policies that promote rather than inhibit increases in density. The objective is to direct development to locations that promote the efficient utilization of public infrastructure such as schools, roads, and highways. In other words, the growth boundary is an integral part of urban planning, the set of policies that determines the spatial arrangement of activities in the metropolitan area.

Municipal versus Metropolitan Growth Boundaries

Our discussion of growth boundaries considers the case of a metropolitan boundary. The analysis is applicable to the two metropolitan areas in the United States—Portland and Minneapolis–St. Paul—that use growth and service boundaries to control the population of a metropolitan region. Urban service boundaries typically apply to individual municipalities rather than an entire metropolitan area. The basic logic of growth boundaries doesn't change with the level of geography. If one municipality adopts a growth boundary, it will displace households to other municipalities in the metropolitan area, triggering the same sort of changes in the common utility level and land rent.

There are two differences between a municipal and a metropolitan boundary. First, people are more mobile between municipalities than between metropolitan areas, so we would expect a quicker response to a municipal growth boundary. Second, some of the people displaced by a municipal boundary will relocate to other municipalities in the same metropolitan area, so some congestion and pollution will simply move to other parts of the metropolitan area. If the residents of the control city travel to neighboring municipalities to work, shop, or socialize, they will encounter some of the congestion and pollution displaced from their own municipality.

Trade-offs with Growth Boundaries and Open Space

We've seen that a growth boundary decreases utility levels throughout a region and increases land rent in the control city. Renters are harmed by higher prices for land and housing, while landowners who own land within the boundary obviously benefit from higher land prices. This raises two questions.

How does a growth boundary affect homeowners? As land owners, homeowners benefit from higher land prices. So a policy that increases the price of land in a city benefits people who own homes at the time the growth control policy is implemented. In contrast, newcomers must pay higher housing prices, so they are harmed.

How do the benefits of a growth boundary compare to the cost? This question is very difficult to answer, and the answer is likely to vary from one city to another. A recent study suggests that in one English city, the cost outweighs the benefit (Cheshire and Sheppard, 2002). The key feature of a growth boundary—or a green belt or open space within the city—is that it provides public space at the expense of private space. The public open space in and around a city provides a bucolic atmosphere and opportunities for outdoor recreation and views. The trade-off is that the limited supply of developable land leads to higher prices and higher density—less private space. The authors conclude that a modest relaxation of the open space and boundary policies of Reading, England, would generate a net gain of $384 per household per year, or about 2 percent of annual income.

OTHER GROWTH-CONTROL POLICIES

In addition to growth and service boundaries, cities use a number of policies to control their growth. In this part of the chapter, we explore the effects of two alternative policies, limits on building permits and development taxes.

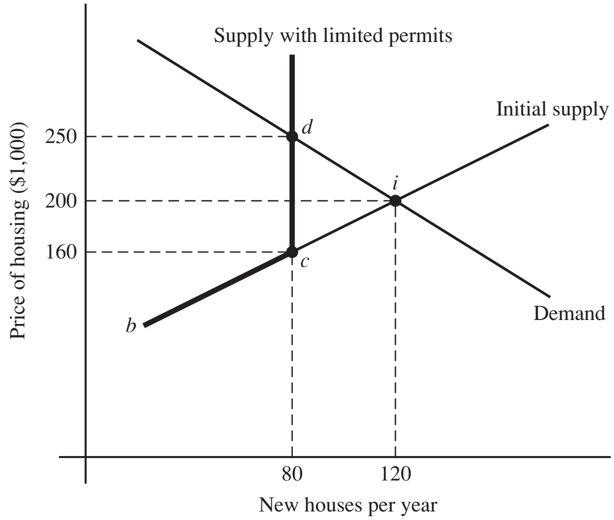
Limiting Building Permits

Consider a city that limits the number of building permits for new housing and business facilities. If the number of permits issued is less than the number demanded by developers, the policy decreases the number of people who can live and work in the city. Like a growth boundary, a limit on building permits displaces households from one city to another, triggering changes in the two cities that generate lower utility in both cities. This is shown by Figure 9–4, with a shift of 0.5 million people from the control city to the other city.

Figure 9–5 shows the implications of a permit limit on the price of new housing. The initial equilibrium is shown by point *i*, with a price of $200,000 and 120 new houses per year. If the city limits the number of building permits to 80 per year, the new supply curve for housing is the kinked curve that includes points *b, c,* and *d*: The maximum number of new houses is 80, so the supply curve goes vertical at 80 houses. The new supply curve intersects the demand curve at an equilibrium price of $250,000, meaning that the permit policy increases the price of housing by $50,000.

The permit policy also decreases the cost of producing housing. As reviewed in Section 2.2 of "Tools of Microeconomics," the appendix at the end of the book, a supply curve is also a marginal-cost curve. Like other supply curves, the housing supply curve shows the marginal cost of production. For example, at the initial equilibrium price of $200,000, firms supply 120 houses. The 120th house would not be

FIGURE 9–5 Market Effects of Limit on Building Permits



If the number of building permits is limited to 80, the supply curve for housing is the kinked curve *bcd*. The new equilibrium is shown by point *d* (price of housing = $250,000). Point *c* shows the marginal cost of production ($160,000). The difference between the price and the marginal cost is the developer's willingness to pay for the building permit ($90,000).

supplied at a price of $199,000 because it costs more than $199,000 to produce; a price of $200,000 is just high enough to cover the production cost, so the marginal cost is just under $200,000. Farther down the supply curve at point *c*, the cost of supplying the 80th house is $160,000. The permit policy decreases the number of houses built from 120 to 80, so the marginal cost of production decreases.

The permit policy decreases the marginal cost of production because it decreases the demand for land and its price. For example, if houses are built on quarter-acre lots, the permit policy decreases the demand for vacant land from 30 acres per year (120 houses times 0.25 acres per house) to 20 acres per year. The decrease in the demand for land decreases the market price of land, decreasing the cost of producing housing. A comparison of point *i* to point *c* in Figure 9–5 shows that the permit policy decreases the marginal cost of production by $40,000.

The city must decide how to allocate its 80 building permits among its developers. One option is to auction the permits to the highest bidders. What is the monetary value of a building permit? A person with a building permit can make a profit equal to the difference between the market price of a house ($250,000, as shown by point *d*) and the cost of producing the house ($160,000, as shown by point *c*), so the monetary value of a permit is $90,000. If the city auctions the permits to the highest bidders, the market price would be $90,000. A second option is to allocate building permits to developers that promote the city's development objectives. A city could allocate its permits to high-density housing or to a project in an area targeted for development. Or a city could stage a "beauty contest," giving the permits to the development project that is most appealing to planning officials.

Development Taxes

Another way to limit a city's population is to impose a development tax on new dwellings. As explained in detail later in the book, local governments use various taxes to finance local public goods. When the annual taxes from a property owner fall short of the costs of providing public services, one response is to impose a one-time development tax to cover the gap. In this case, a development tax is simply solving a fiscal problem and is not a growth-control policy.

Some cities impose impact fees on commercial and industrial developers, using the revenue from the fees to expand local transportation networks. For example, in the Westchester area of western Los Angeles, developers pay a one-time fee of $2,010 for each additional rush-hour trip generated by new office buildings. The revenue from the impact fee is used to widen the roads used by the employees of the new office buildings. Impact fees can reduce the fiscal burden of new development, decreasing the opposition to development.

HOUSING REGULATIONS AND HOUSING PRICES

Local governments use a wide variety of policies to regulate residential development. The policies, which are implemented by zoning boards, city councils, and environmental review boards, regulate building design, limit lot size, collect fees for infrastructure, and acquire land for open space. On average, it takes about six months to get a building permit for a project. The housing regulations decrease the price elasticity of supply of housing and generate higher housing prices.

A recently developed index of the housing regulatory environment shows that the stringency of housing regulations varies across cities (Gyourko, Saiz, and Summers, 2008). The first column of Table 9–1 shows the Wharton regulatory index (also known as WRLURI for "Wharton Residential Land Use Regulatory Index") for several metropolitan areas. The index is normalized to a value of zero: for the city with the average stringency of housing regulations, the value of the index is zero. A positive index value indicates above-average regulation, and the larger the number, the more stringent the regulations. The value of the index is relatively large for a city where

- rezoning and subdivision permits take a long time to be approved;
- impact fees for infrastructure are relatively high;
- permitting requires the approval of several review bodies;
- the number of building permits is capped;
- the minimum lot size is relatively large;
- builders are required to provide open space.

TABLE 9–1 Housing Regulations and Housing Prices

	Metropolitan Area		Municipality	
	Wharton index	Price Premium ($)	Wharton index	Price Premium ($)
Atlanta	0.04	5,120	0.70	89,606
Boston	1.54	197,132		
Chicago	0.06	7,680	−1.15	−147,209
Dallas	−0.35	−44,803	−0.14	−17,921
San Francisco	0.90	115,207	1.96	250,896
Seattle	1.01	129,288	2.39	305,939
Premium per unit index		128,008		128,008

Source: Gyourko, Saiz, and Summers (2008); Gyourko (2009).

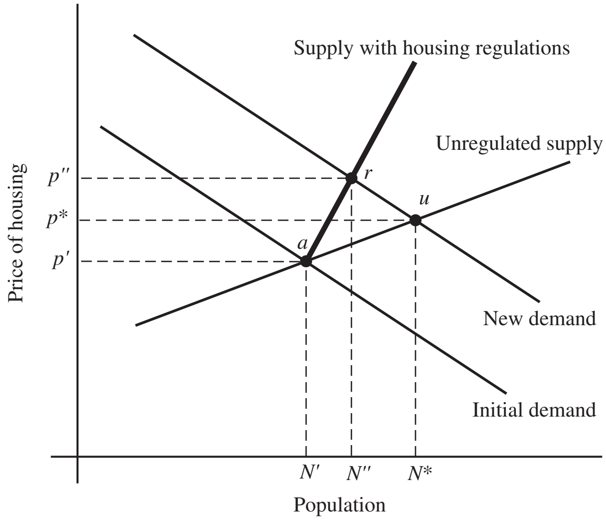
The second column of numbers in Table 9–1 shows, for each metropolitan area, the housing premium generated by residential building regulations. A one-unit increase in the index translates into an increase of $128,008 in the price of the typical house. The index suggests that the relatively lax regulations in Dallas generate a cost savings of almost $45,000, while the relatively stringent regulations in Boston generate a premium of over $197,000. There is a positive correlation between the value of the regulatory index and household wealth: regulation is more stringent in wealthier cities, so wealthier cities have higher housing prices.

Figure 9–6 shows the economics behind the premium associated with housing regulations. Starting from the initial equilibrium at point *a*, the housing supply curve is kinked, and is relatively steep for increases in quantity. In other words, supply is relatively inelastic in the positive direction. An increase in demand shifts the demand curve to the right, and as we move along the kinked supply curve, the equilibrium price increases by a relatively large amount (from p' to p'' rather than from p' to p^*). In contrast, the equilibrium quantity increases by a relatively small amount, from N' to N'' rather than from N' to N^*. In other words, stringent housing regulations generate smaller and more expensive cities.

Glaeser, Gyourko and Saks (2005) compute the numbers behind the analysis shown in Figure 9–6. In a metropolitan area with stringent regulations, a 10 percent increase in population increases housing prices by about $60,000. In contrast, in a metropolitan area with light regulations, a 10 percent increase in population increases housing prices by only $5,000. The more restrictive city experiences a larger increase in housing prices because it has a relatively inelastic supply.

The third and fourth columns of Table 9–1 show the Wharton index and housing premiums for municipalities (political or central cities). The Chicago municipality has relatively lax regulations, generating a cost savings of about $147,000 relative to a city with average regulations. In contrast, Seattle has very stringent regulations, generating a premium of almost $306,000.

FIGURE 9–6 Housing Regulation and Housing Prices



Housing regulation generates a relatively inelastic supply and a kink in the supply curve at the initial equilibrium (point *a*). An increase in demand generates a relatively large increase in price (to p'') and a relatively small increase in quantity (to N'').

SUMMARY

This chapter describes the history of zoning and its legal foundations, and it also explores the effects of various growth-control policies. Here are the main points of the chapter.

1. Zoning is a blunt policy to control pollution because it just moves pollution around. An alternative policy is to combine pollution taxes with zoning.
2. Local governments can use minimum lot-size zoning to exclude land users who would generate a fiscal deficit—paying less in taxes than they get in public services.
3. The use of zoning to provide open space generates excessive amounts of open space because voters don't bear the full cost of the public good.
4. In a two-city region, an urban growth boundary in one city decreases utility in both cities and increases land rent in the city with the boundary.
5. A limit on building permits increases the equilibrium price of housing and decreases the price of vacant land.
6. In a city with relatively stringent housing regulations, growth pressure generates a relatively large increase in housing prices and a relatively small increase in population.

APPLYING THE CONCEPTS

For exercises that have blanks (_______), fill each blank with a single word or number. For exercises with ellipses (…), complete the statement with as many words as necessary. For exercises with words in square brackets ([increase, decrease]), circle one of the words.

1. **Voting for Open Space**
   The city of Medianville has 100 citizens with income ranging from $20,000 to $70,000. The average income is $50,000 and the median income is $60,000. An individual's demand for open space depends only on income, and the individual demand curve (marginal benefit curve) has a vertical intercept equal to 1/1,000 of income and a slope of −$0.50 per acre of open space. The marginal social cost of open space is $2,000 per acre. The city picks the quantity of open space with majority rule, using a head tax to pay for open space.
   a. The decisive voter—the one who will determine the election outcome—is the _______ voter, with income = _______.
   b. Use a graph like Figure 9–1 to show the situation for the decisive voter. The city will provide _______ acres of open space because …
   c. If the city switches to zoning to provide open space, without any compensation to landowners, it will provide _______ acres of open space because . . .
2. **Landowners Vote on a Growth Boundary**
   Consider a city where the owners of vacant lots will vote on a proposed growth boundary. There are 12 vacant lots (one per owner), four inside the proposed boundary and eight outside the boundary. The initial price of land is $20 per lot. The growth boundary will reduce the price of land outside the boundary to zero.
   a. If the growth boundary doubles the price of lots within the boundary, the policy [increases, decreases] the total value of land from _______ to _______.
   b. If the growth boundary quadruples the price of lots within the boundary, the policy [increases, decreases] the total value of land from _______ to _______. The vote tally will be _______ landowners in favor and _______ opposed.
   c. Suppose the city combines the proposed growth boundary with a capital-gains tax equal to 80 percent of the change in the price of land. The revenue from the tax will be redistributed in equal shares to landowners outside the boundary. The tax per inside landowner would be _______ and the compensation per outside landowner would be _______. The vote tally will be _______ landowners in favor and _______ opposed.
3. **Compensation for a Growth Boundary**
   Consider a residential city that initially has no controls on urban growth. Its radius is expected to increase from six miles to nine miles. Suppose the city announces a new growth boundary at its current radius (six miles). Your job is to develop a self-financing program under which landowners who gain from the boundary compensate landowners who lose. For each of the following

individuals, indicate whether he or she will be a compensator or compensatee, and whether the payment will be relatively large or relatively small. Illustrate your answers with a graph.

- **a.** Bennie owns land three miles from the city center. He is a compensat [or, ee], and the payment will be relatively [large, small].
- **b.** Remus owns land 6.5 miles from the center. He is a compensat [or, ee], and the payment will be relatively [large, small].
- **c.** Margie owns land 8.5 miles from the city center. She is a compensat [or, ee], and the payment will be relatively [large, small].

4. Growth Boundary and the Urban Labor Market

Consider the effects of a growth boundary on the urban labor market. Assume the boundary directly affects only residential land, not commercial or industrial land.

- **a.** Use a supply-demand graph of the urban labor market to show the effects of the growth boundary on the city's equilibrium wage and total employment.
- **b.** Arrows up or down: The policy _______ the equilibrium wage and _______ equilibrium employment.
- **c.** We would expect the owners of commercial and industrial land to [support, oppose] the boundary because . . .

5. Permit Queue: Hey, No Cuts!

Consider the building-permit policy depicted in Figure 9–5. Suppose the city announces on January 1 that 300 days later (October 28) it will give the 70 permits to the first 70 licensed building contractors through the planning office door. The police chief announces the following queuing rules:

- **i.** No cuts: A person who joins the queue goes to the end of the queue.
- **ii.** No substitutions: No one can reserve a place in line for anyone else.

The price of a permit is the time a recipient spends in the queue. There are four types of licensed contractors, with 25 contractors of each type: For type A, the opportunity cost of queue time is $300 per day; for type B, the cost is $500 per day; for type C, the cost is $1,000 per day; for type D, the cost is $2,000 per day.

- **a.** Draw a supply-demand graph of the permit market, with the price measured as the days spent in the queue.
- **b.** The equilibrium queue time is _______ days because . . .
- **c.** If the city eliminates the no-substitution rule, the equilibrium queue time will [increase, decrease, not change] because . . . The line will form immediately if . . .
- **d.** If the city eliminates the no-cuts rule, the allocation of permits will be based on _______, and the price will be in terms of . . .

6. Decrease in Demand and Permit Price

In Figure 9–5, the equilibrium price of building permits is $90,000. Consider the effects of changes in demand.

- **a.** If the demand curve shifts downward by $25,000, the equilibrium price of housing is _______ and the equilibrium price of a permit is _______ = _______ minus _______.
- **b.** If the demand curve shifts downward by $100,000, the equilibrium price of the permit is _______ because . . .

7. **Permit and Leftover Principle**
   In the analysis of limits on building permits illustrated by Figure 9–5, the price of housing increases while the price of land decreases. In other words, housing and land prices change in opposite directions.
   a. Housing and land prices change in opposite directions because the policy ______ in the housing market and ______ in the land market. Illustrate with two graphs, one for each market.
   b. Recall the leftover principle: Land rent = total revenue − nonland cost. If 80 permits are handed out free of charge and can be resold, the changes in prices are consistent with the leftover principle because . . .
8. **Permit Restrictions and the Urban Labor Market**
   Consider the labor-market effects of a policy that limits residential building permits.
   a. Use a supply-demand graph of the urban labor market to show the effects of the permit policy on the city's equilibrium wage and total employment.
   b. Arrows up or down: The policy ______ the equilibrium wage and ______ equilibrium employment.
   c. We would expect the owners of commercial and industrial land to [support, oppose] the permit policy because . . .
9. **Incidence of Development Tax**
   Leapfrog city has two rings of vacant land suitable for housing, one that is three miles from the center and a second that at the edge of the city, six miles from the center. Both rings were expected to be developed in the next year. Suppose the city imposes a new development tax of $20,000 per new house. The tax will be paid, in legal terms, by the firm that builds the house.
   a. Use a supply-demand graph to show the effects of the development tax on the city's housing market.
   b. Arrows up, down, or horizontal: The tax ______ the equilibrium price of new housing, ______ the equilibrium quantity of housing, ______ the demand for vacant land, and ______ the price of land.
   c. According to Ms. Wizard, "The tax will prevent the development of the outer ring of vacant land (six miles from the center)." Draw a graph that is consistent with Wizard's statement.
10. **Land-Use Policies and the Price of Land**
    Consider the following: "Depending on the variable controlled by a land-use policy, the policy may either increase or decrease the price of undeveloped land." The "variable" is the supply of or demand for a particular type of land. Consider the following land-use policies: Growth boundary (*GB*); Limit on building permits (*BP*); Development tax (*DT*). For each policy, fill the blanks in the following statement. For a policy that controls two variables, there will be two statements.
    A ______ policy [increases, decreases] the [demand for, supply of] ______ land, and [increases, decreases] the equilibrium price.

REFERENCES AND ADDITIONAL READING

1. Brueckner, Jan K. "Strategic Control of Growth in a System of Cities." *Journal of Public Economics* 57 (1995), pp. 393–416.
2. Cheshire, Paul, and Stephen Sheppard. "The Welfare Economics of Land Use Planning." *Journal of Urban Economics* 52 (2002), pp. 242–69.
3. Engle, Robert, Peter Navarro, and Richard Carson. "On the Theory of Growth Controls." *Journal of Urban Economics* 32 (1992), pp. 269–83.
4. Evenson Bengte, and William C. Wheaton. "Local Variation in Land Use Regulations." *Brookings-Wharton Papers on Urban Affairs: 2003.* Washington DC: Brookings, 2003.
5. Fischel, William. *The Economics of Zoning Laws.* Baltimore: Johns Hopkins, 1985.
6. Fischel, William. "An Economic History of Zoning and a Cure for its Exclusionary Effects." *Urban Studies* 41 (2004), pp. 317–340.
7. Glaeser, Edward, Joseph Gyourko, and Raven Saks. *American Economic Review* 95 (2005), pp. 329–33.
8. Gyuorko, Joseph. "Housing: Urban Housing Markets." Chapter 5 in *Making Cities Work*, ed. Robert P. Inman. Princeton NJ; Princeton University Press, 2009.
9. Gyourko, Joseph, and Edward Glaeser. "Urban Growth and Housing Supply," *Journal of Economic Geography* 6 (2006), pp. 71–89.
10. Gyourko, Joseph, Albert Saiz, Albert Summers, and Anita Summers. "A New Measure of the Local Regulatory Environment for Housing Markets: The Wharton Residential Land Use Regulatory Index." *Urban Studies* 45 (2008), pp. 693–729.
11. Helsley, Robert W., and William C. Strange. "Strategic Growth Controls." *Regional Science and Urban Economics* 25 (1995), pp. 435–60.
12. Lillydahl, Jane H., Arthur C. Nelson, Timothy V. Ramis, Antero Rivasplata, and Steven R. Schell. "The Need for a Standard State Impact Fee Enabling Act." *Journal of the American Planning Association* 54 (Winter 1988), pp. 7–17.
13. Mills, Edwin S., and Luan Sende Lubuele. "Inner Cities." *Journal of Economic Literature* 35 (1997), pp. 727–56.
14. Siegan, Bernard. *Land Use without Zoning*. Lexington, MA: D.C. Heath, 1972.

PART THREE

Urban Transportation

One of the advantages of an urban location is its proximity to the many activities within a metropolitan area. This part of the book examines the two main components of the urban transportation system. Chapter 10 considers the automobile/highway system, focusing on three externalities caused by automobiles: congestion, environmental degradation, and collisions. The chapter explores various policy responses to these externalities. Chapter 11 explores the economics of urban mass transit, focusing on the commuter's choice of a mode of travel (e.g., automobile versus mass transit) and a city planner's choice of a mass-transit system (e.g., buses versus light rail versus heavy rail). The chapter explains why so few commuters in the United States use mass transit and why light-rail and heavy-rail transit systems are usually less efficient than bus systems.

CHAPTER 10

Autos and Highways

The home is where part of the family waits until the others are through with the car.

—HERBERT PROCHNOW

I started to slow down but the traffic was more stationary than I thought.

—FROM AN AUTOMOBILE INSURANCE CLAIM FORM

This first chapter on urban transportation discusses the automobile. In the United States, the automobile is the travel mode for 88 percent of commuting and over 90 percent of all travel. We explore three externalities generated by the automobile—congestion, air pollution, and vehicle collisions—and discuss policy responses to the externalities. Recall the third axiom of urban economics:

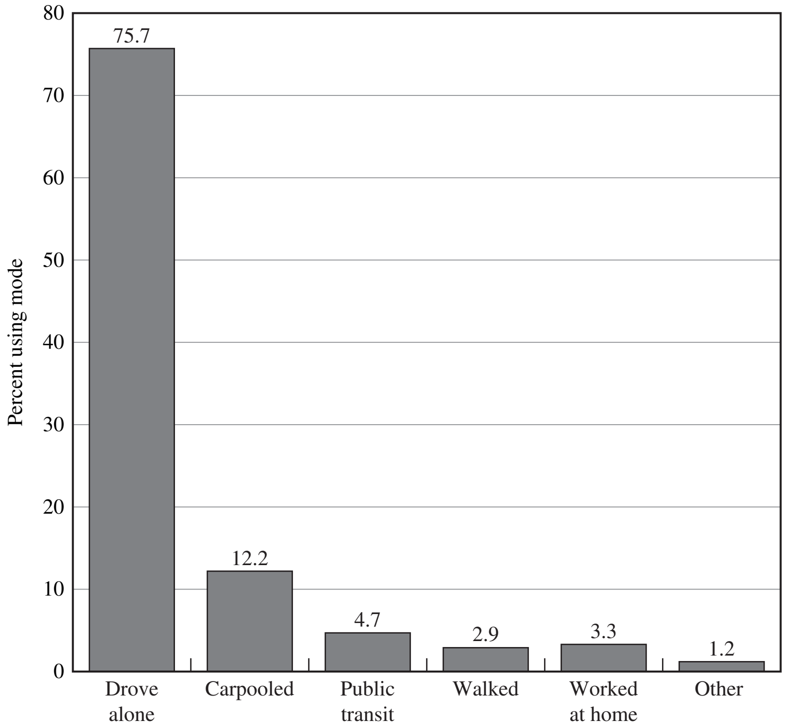
Externalities cause inefficiency

The economic approach to externalities is to internalize them, imposing a tax equal to the marginal external cost. We also explore the merits of alternative policies to address the externalities, including subsidies for mass transit, mileage charges, and gasoline taxes.

Figure 10–1 (page 258) shows the distribution of travel modes for U.S. workers. About three-fourths drive alone, and another 12 percent carpool in private vehicles. The drive-alone shares are highest in cities in Ohio and Alabama, which together have 6 of the top 10 drive-alone cities. Carpooling rates are highest in California and Texas, which together have 8 of the top 10 carpooling cities, with rates between 18 percent and 20 percent. The share of commuting by automobile is highest among workers who commute within suburban areas and lowest among workers who commute within a central city.

At the national level, 5 percent of commuters use public transit, but the transit share varies substantially across cities. The transit share is above 10 percent in just two metropolitan areas, New York (25 percent) and Chicago (12 percent). Seven metropolitan areas have transit shares above 6 percent: San Francisco, Washington, Boston, Philadelphia, Honolulu, Seattle, and Pittsburgh. The rest of the U.S. metropolitan areas have transit shares less than 6 percent. The transit

FIGURE 10–1 Modal Choice for U.S. Commuters



Source: U.S. Census Bureau. *Journey to Work: 2000*. Washington DC: U.S. Census Bureau, 2004.

TABLE 10–1 Purposes of Travel

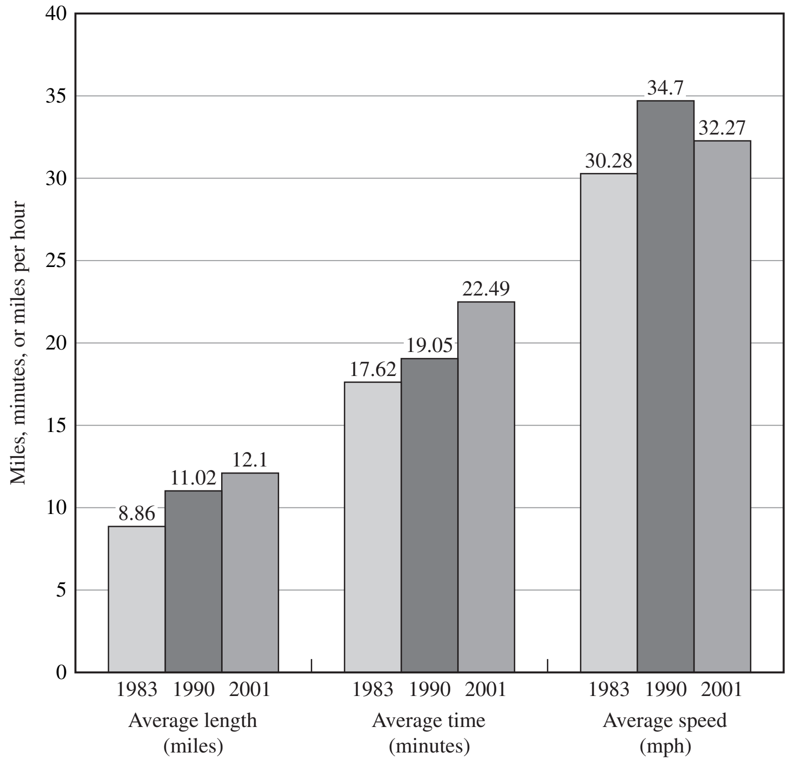
	Share of Travel (percent)	Average Trip Length (miles)
Social and recreational	30	11.36
To/from work	19	12.11
All other family and personal business	19	7.84
Shopping	14	7.02
Work-related business	9	28.26
School/church	6	6.00
Other	4	43.08

Source: U.S. Department of Transportation. *Summary of Travel Trends, 2001 National Household Travel Survey* (2004).

share is 11 percent for central-city residents, compared to 2 percent for suburban residents.

As shown in Table 10–1, commuting (driving to and from work) is responsible for only one-fifth of travel by private vehicles. Social and recreational travel is at the top of the list, with a 30 percent share. The average annual mileage per

FIGURE 10–2 Commuting Distance, Time, and Speed



Source: U.S. Department of Transportation. *Summary of Travel Trends, 2001 National Household Travel Survey.* Washington, DC: U.S. Department of Transportation, 2004.

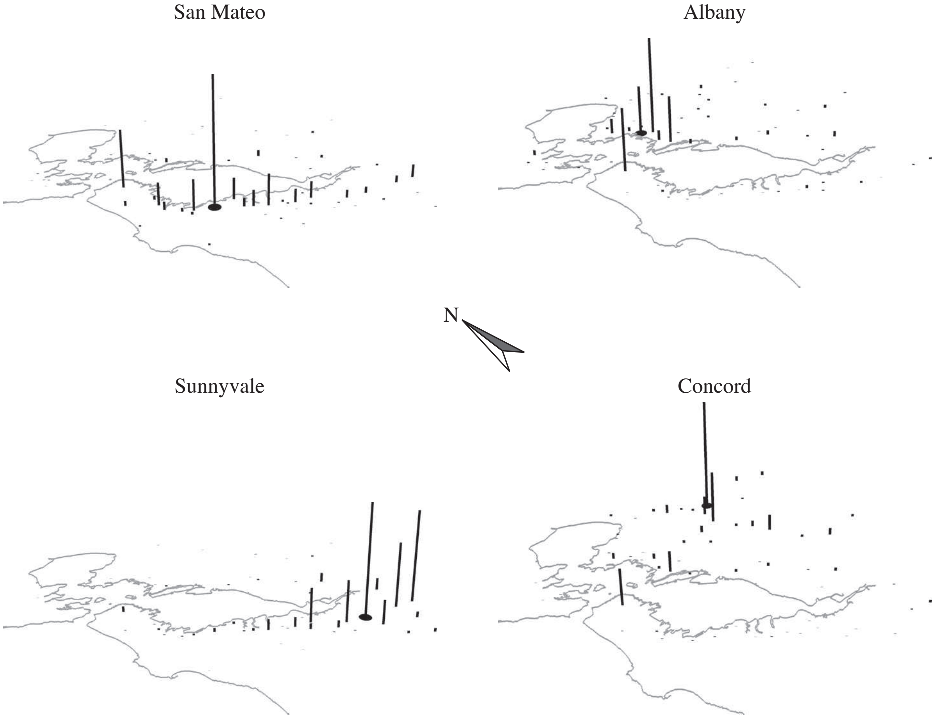
household is over 35,000, with about 6,700 for commuting. Table 10–1 shows the average trip lengths: 12 miles for commuting and 11 miles for social and recreational trips.

Figure 10–2 shows trends in the length, time, and speed of commuting (the numbers are for one-way travel). Between 1983 and 2001, trip length increased by about 37 percent, while travel time increased by a smaller percentage (28 percent) because travel speed increased. Between 1990 and 2001, speed decreased while the trip length increased, so travel time increased by 18 percent.

As we saw earlier in the book, the most frequent commuting trip is between suburban municipalities. As shown in Figure 7–2, about 44 percent of commuter trips occur outside the central city, while 29 percent are within central cities.

Map 10–1 shows some examples of commuting in the San Francisco Bay area. Each panel shows the number of commuters from a particular municipality (marked with a disk) to other municipalities in the metropolitan area. Panel A shows commuting from San Mateo: The tallest bar shows the number of workers who commute within San Mateo, and the second tallest bar shows the number of workers

MAP 10–1 Commuting Patterns in San Francisco Bay Area



who commute from San Mateo to the municipality of San Fransisco to the north. Commuting to the east of San Mateo is spread among a large number of municipalities. Panel B shows commuting from Sunnyvale, close to the southern end of San Francisco Bay. Roughly 23 percent of workers commute within Sunnyvale, and the most of the rest commute to jobs in other municipalities the South Bay area. Panel C shows commuting from Albany, in the East Bay area. Roughly 16 percent of commuters cross the bay to jobs in the city of San Francisco, while 40 percent commute within Albany or to nearby Berkeley. Finally, Panel D shows commuting from Concord, about 23 miles east of San Francisco. Roughly 44 percent of workers commute to jobs in Concord or nearby Walnut Creek (shown by the second tallest bar) while 10 percent commute to jobs in the city of San Francisco.

CONGESTION EXTERNALITIES

According to the Texas Transportation Institute, the typical U.S. commuter in 2003 wasted about 47 hours because of traffic congestion. In some cities, the time lost by the typical commuter is much higher: 93 hours in Los Angeles, 72 hours in San Francisco, 69 hours in Washington DC, 67 hours in Atlanta, and 63 hours in Houston. In addition to time lost, $5 billion worth of gasoline and diesel fuel is wasted each year because of delays and slow travel. Adding the value of lost time to the wasted fuel, the annual cost is $63 billion per year. This is about five times the congestion cost experienced in 1982.

We'll use a simple model to explain congestion externalities and evaluate some alternative public policies to deal with it. Consider a travel route within a metropolitan area with the following characteristics:

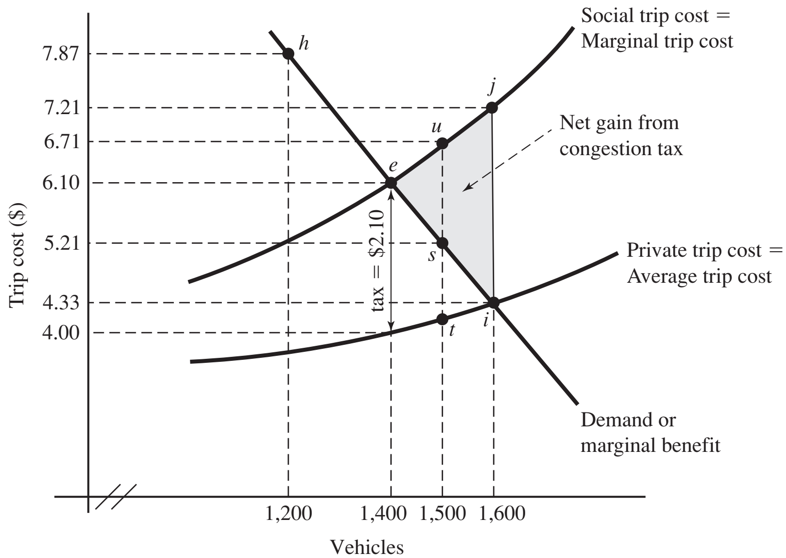
- **Distance.** The travel route is 10 miles long and could be along a radial highway into the city center or along a circumferential highway linking suburbs.
- **Monetary travel cost.** The monetary cost of auto travel is 20 cents per mile, or $2.00 for the 10-mile route.
- **Time cost.** The time cost of a trip is the time times the opportunity cost per minute ($0.10).

The total cost of a trip is the $2.00 monetary cost plus the time cost, which depends on how long the trip takes. There is one person per vehicle, so we can use "vehicles" and "drivers" interchangeably.

The Demand for Urban Travel

Consider first the demand side of urban travel. In Figure 10–3, the horizontal axis measures the number of vehicles per lane per hour and the vertical axis measures the cost of the commuting trip, the sum of the monetary and time

FIGURE 10–3 Congestion Externalities and the Congestion Tax



The equilibrium is shown by point *i*: When drivers pay the private trip cost, traffic volume is 1,600. The optimum is shown by point *e*, where the marginal benefit (shown by the demand curve) equals the marginal cost (the Social trip cost), generating a volume of 1,400 vehicles. The net gain from congestion tax is shown by the shaded area.

costs. The demand curve shows the number of drivers who make the trip, which depends on its cost. For example, if the trip cost is \$7.87, point *h* shows that there are 1,200 people for whom the benefit of the trip exceeds the cost, so the traffic volume is 1,200 vehicles per lane per hour. As the cost of the trip decreases, the benefit exceeds the cost for more people, so we move downward along the demand curve to 1,400 vehicles at a cost of \$6.10 and 1,600 vehicles at a cost of \$4.33.

As explained in Section 2.1 in "Tools of Microeconomics," the appendix at the end of the book, a demand curve is also a marginal-benefit curve. The demand curve in Figure 10–3 shows how much the marginal traveler is willing to pay for a trip. For example, at a price of \$7.87, 1,200 people make the trip. The 1,200th traveler makes the trip if the cost is \$7.87, but wouldn't make the trip if the cost were any higher, say \$7.88. This tells us that the benefit of the 1,200th trip is just below \$7.87. Similarly, the marginal benefit for the 1,400th traveler is \$6.10 and the marginal benefit for the 1,600th vehicle is \$4.33. As we move downward along the demand curve, people with progressively lower marginal benefits use the highway.

The Private and Social Costs of Travel

Table 10–2 shows the relationships between traffic volume and travel time. Column B lists the trip time for different volumes. For up to 400 vehicles, there is no congestion: Everyone travels at the legal speed limit of 50 miles per hour, and the trip takes 12 minutes. But after we cross the congestion threshold of 400 vehicles, the time required to make the trip increases. For example, the trip time increases to 12.48 minutes with 600 vehicles, to 17.28 minutes with 1,200

TABLE 10–2 Traffic Volume, Travel Time, and the Congestion Externality

A Volume (vehicles per lane)	B Trip Time (minutes)	C Private Trip Cost (\$)	D Increase in Time per Vehicle (minutes)	E Increase in Total Travel Time (minutes)	F External Trip Cost (\$)	G Social Trip Cost (\$)	H Marginal Benefit (demand)
200	12.000	3.20	0.000	0.00	0.00	3.20	16.73
400	12.000	3.20	0.000	0.00	0.00	3.20	14.96
599	12.476						
600	12.480	3.248	0.004	2.40	0.24	3.49	13.19
1,199	17.268						
1,200	17.280	3.728	0.012	14.40	1.44	5.17	7.87
1,399	19.985						
1,400	20.000	4.000	0.015	21.00	2.10	6.10	6.10
1,599	23.262						
1,600	23.280	4.328	0.018	28.80	2.88	7.21	4.33
1,799	27.100						
1,800	27.120	4.712	0.020	36.00	3.60	8.31	2.56

vehicles, to 27.12 minutes with 1,800 vehicles. As the highway becomes crowded, the space between vehicles decreases, and drivers slow down to maintain safe distances between cars.

Column C shows the private trip cost, defined as the cost incurred by the typical driver. The time cost equals the trip time (shown in column B) times the opportunity cost ($0.10 per minute), for example, $1.20 for a volume of 400 vehicles, $1.248 for 600 vehicles, and so on. Adding the $2 monetary cost to this time cost, we get the numbers for the private trip cost in column C. The private trip cost increases from $3.20 for volume up to 400, to $3.248 for 600 vehicles, $3.728 for 1,200 vehicles, and so on up to $4.712 for 1,800 vehicles.

Columns D, E, and F show the numbers behind the congestion externality. Column D shows the increase in travel time per vehicle from one additional vehicle. For example, the travel time for 599 vehicles is 12.476 minutes, and the travel time for 600 vehicles is 0.004 minutes longer, 12.480 minutes. In other words, when the 600th vehicle enters the roadway, it slows down each other vehicle by 0.004 minutes. Multiplying this by the 599 vehicles, the increase in others' travel time when the 600th vehicle enters is 2.40 minutes. Finally, multiplying the increase in travel time by the opportunity cost of travel time ($0.10 per minute), we get the external cost of the 600th vehicle, $0.24. This tells us that the 600th vehicle imposes a cost on other drivers of $0.24. Repeating the same calculations, we get an external cost of $1.44 for the 1,200th vehicle, $2.10 for the 1,400th vehicle, and so on. The external cost increases with traffic volume.

Column G shows the social trip cost, the sum of the private cost and the external cost. When there is no congestion (when the volume is less than 400 vehicles), the external cost is zero, so the social cost equals the private cost (column C). But once we pass the congestion threshold, the social cost of the trip exceeds the private cost. For example, at a volume of 1,400 vehicles, the social cost is $6.10 compared to a private cost of $4.00.

There are some alternative labels for the private trip cost and the social trip cost. The private trip cost is the cost per driver, so we could call it the average cost of travel. The social trip cost is the social cost associated with the last or marginal vehicle, so we could call it the marginal cost of travel.

Equilibrium versus Optimum Traffic Volume

What is the equilibrium number of vehicles? A person will use the road if his or her willingness to pay for the trip (the marginal benefit) exceeds the private trip cost. Figure 10–3 shows the demand curve and the private trip cost curve. The demand curve intersects the private-cost curve at point *i*, indicating that the equilibrium volume is 1,600 vehicles and the equilibrium trip cost is $4.33. For the first 1,600 people, the willingness to pay is greater than or equal to the private trip cost, so they use the roadway. The 1,601st vehicle does not use the roadway because the willingness to pay is less than the private trip cost.

What is the optimum number of vehicles? We can use the marginal principle to identify the socially efficient number of vehicles. The marginal principle is reviewed in Section 1.1 of "Tools of Microeconomics," the appendix at the end of the book. According to the marginal principle, we should increase the level of an activity until the marginal social benefit equals the marginal social cost. No positive externalities are associated with travel, so the demand curve shows the marginal social benefit of travel. The marginal social cost is shown by the social trip cost curve in Figure 10–3. The demand curve intersects the social-cost curve at point *e*, so the optimum volume is 1,400 vehicles. For the first 1,400 vehicles, the social benefit of travel (the willingness to pay) is greater than or equal to the social cost, so their use of the roadway is socially efficient. In contrast the social cost of the 1,401st vehicle exceeds the social benefit, so its use of the highway is not socially efficient.

The equilibrium volume exceeds the optimum volume because each driver ignores the congestion cost imposed on others. An additional vehicle slows traffic, forcing other drivers to spend more time on the road. Suppose that Lois, the 1,500th driver, has a willingness to pay of \$5.21 (shown by point *s*). With 1,500 vehicles, the private trip cost is \$4.16 (shown by point *t*), and the social trip cost is \$6.71 (shown by point *u*). She will use the road because her willingness to pay exceeds her private trip cost (\$5.21 > \$4.16). But her use of the road is inefficient because her willingness to pay is less than the social trip cost (\$5.21 < \$6.71). The burden she imposes on society equals the gap between the social benefit (her benefit) of \$5.21 and the social cost of \$6.71, or \$1.50. Lois ignores the external cost of her decision, so she makes an inefficient choice.

THE CONGESTION TAX

The simple solution to the congestion problem is to use a congestion tax to internalize the externality. In Figure 10–3, a congestion tax of \$2.10 per trip would shift the private trip cost curve upward by \$2.10, decreasing the equilibrium number of vehicles from 1,600 to 1,400. For Lois (in vehicle 1,500), the benefit of the trip is still \$5.21, but if 1,500 vehicles use the road, her cost is the sum of the private trip cost of \$4.16 (point *t*) and a tax of \$2.10, or \$6.26. Her cost now exceeds her willingness to pay, so she doesn't use the highway. Similarly, for the 1,401st through the 1,600th vehicles, the willingness to pay is now less than the cost of making the trip, so they stay off the road. The congestion tax ensures that decision-makers face the full social cost of travel so highways will be used efficiently.

Benefits and Costs of the Congestion Tax

From the perspective of the individual traveler, the imposition of congestion taxes generates good news and bad news. Consider first the people who pay the tax and continue to use the highway. In Figure 10–3, Hiram is at point *h* on the demand curve. The bad news is that he pays a congestion tax of \$2.10. There are two bits of good news:

- **Decrease in time cost.** The tax decreases traffic volume, so travel speed increases and travel time decreases. In Figure 10–3, the tax decreases the private trip cost from $4.33 to $4.00, a savings of $0.33 for Hiram and every other driver.
- **Lower income tax.** The government can use the revenue from the congestion tax to cut other local taxes, so the congestion tax is revenue-neutral. Suppose the government divides the congestion tax revenue equally among the 1,600 people who initially used the roadway, cutting each person's income tax by $1.84.

As shown in the first row of Table 10–3, Hiram has a net benefit of $0.07 from the congestion tax, equal to a benefit of $2.17 minus the $2.10 congestion tax.

Consider next the people like Lois who don't use the roadway after the congestion tax is imposed. The good news is that her income tax is cut by $1.84, just like everyone else who initially used the road. The bad news is that she loses a consumer surplus. (For a review of the concept of consumer surplus see Section 2.6 of "Tools of Microeconomics," the appendix at the end of the book.) Before the tax, her consumer surplus from using the highway was the gap between her willingness to pay ($5.21) and the private trip cost when 1,600 vehicles use the road ($4.33), or $0.88. In the second row of Table 10–3, Lois's tax cut exceeds her loss of consumer surplus, so the congestion tax makes her better off too.

We can use the marginal approach to measure the welfare gain to society from moving from the market equilibrium to the optimum. The relevant concepts are reviewed in Section 1.2 of "Tools of Microeconomics," the appendix at the end of the book. The shaded area in Figure 10–3 shows the welfare gain to society. To explain the logic of the welfare gain, consider a small move from the equilibrium toward the optimum. If we persuade the 1,600th driver not to use the road, what are the benefits and costs?

- **Benefit:** The total travel cost for society decreases by the social trip cost associated with the 1,600th driver ($7.21 at point *j*).
- **Cost:** The driver loses the benefits of the highway trip; the willingness to pay for the trip is shown by the demand curve ($4.33 at point *i*).

By diverting this vehicle, society saves $7.21 in travel costs and sacrifices only $4.33 in foregone travel benefits, for a net gain of $2.88. This is shown in Figure 10–3 as the gap between the social trip cost curve and the demand curve at 1,600 vehicles.

TABLE 10–3 Benefits and Costs of the Congestion Tax

	Cost		Benefit		
	Tax Paid	Lost Consumer Surplus	Decrease in Time Cost	Lower Income Tax	Net Benefit
Hiram	$2.10	—	$0.33	$1.84	$0.07
Lois	—	$0.88	—	$1.84	$0.96

To compute the welfare gain to society from moving to the optimum, we repeat this thought experiment for the 1,599th driver, the 1,598th driver, and so on down to the 1,401st driver. The net gain from diverting the 1,599th driver is slightly lower than the gain from diverting the 1,600th driver because the social trip cost is lower (we are lower on the cost curve) and the willingness to pay is higher (we are farther up the demand curve). As we decrease the number of vehicles, the net gain from diverting a vehicle decreases as the gap between the social trip cost and the demand curve shrinks. The gain to society (the welfare gain) from moving all the way to the optimum level is the sum of the net gains from the diverted vehicles, shown as the shaded area between the social cost curve and the demand curve.

Congestion Taxes and Urban Growth

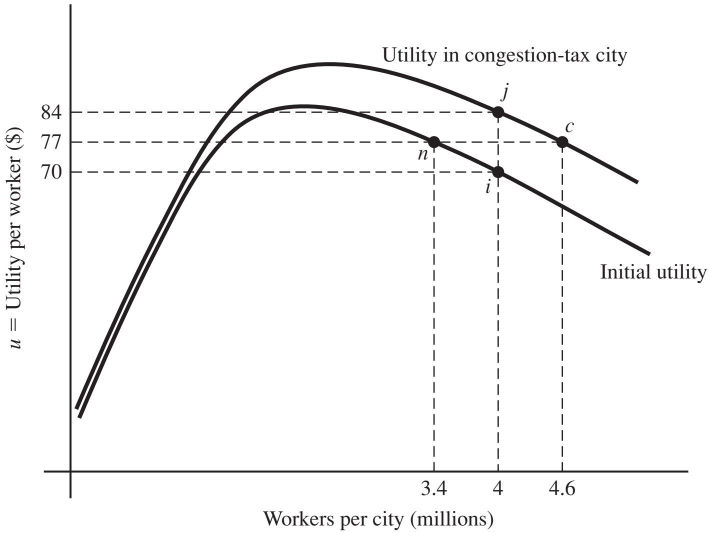
We've seen that because the congestion tax internalizes an externality, it improves the efficiency of the economy and generates a welfare gain to society. We can use the utility curves derived earlier in the book to show the implications of a congestion tax for urban growth. As we'll see, a city that implements a congestion tax will grow at the expense of other cities in the region.

Figure 10–4 shows the utility curves for a two-city region. Suppose that initially congestion is unpriced in both cities, generating point *i* as the initial equilibrium: Each city has a population of 4 million and a utility level of $70. Suppose one city implements a congestion tax and uses the revenue from the congestion tax to cut income taxes. In Figure 10–4, the congestion tax shifts the city's utility curve upward because the internalization of congestion externalities reduces the diseconomies of urban size. Recall that these diseconomies (more noise, pollution, and congestion as population increases) pull utility down as a city grows. The reduction of congestion affects the utility curve in two ways:

- The curve is positively sloped over a larger population range because agglomeration economies dominate diseconomies over a larger population range.
- For the negatively sloped portion of the utility curve, the curve is not as steep: Diseconomies are weaker, meaning that utility falls less rapidly as population increases.

The upward shift of the utility curve causes the congestion-tax city to grow at the expense of the other city. The immediate effect is a utility gap shown by points *i* and *j*: With a population of 4 million each, utility is $14 higher in the congestion-tax city, reflecting the efficiency gains of internalizing the externality. Workers will migrate to the congestion-tax city, causing movement downward along its utility curve (from point *j* toward point *c*). As workers leave the other city, we move upward along the initial utility curve (still relevant for the other city) from point *i* toward point *n*. Equilibrium is restored at points *n* and *c*: The congestion-tax city gains population at the expense of the other city. In addition, utility increases in both cities.

FIGURE 10–4 A Congestion Tax Causes Urban Growth



The initial equilibrium is shown by point *i*: Each of the two cities has a population of 4 million. Internalizing congestion externalities reduces the diseconomies of urban growth, shifting the utility curve upward and opening a utility gap (shown by points *i* and *j*) between the two cities. Migration to the congestion-tax city eliminates the utility gap, restoring equilibrium at points *n* (no congestion tax) and *c* (congestion tax). The congestion-tax city grows at the expense of the other city, and utility increases in both cities.

PRACTICALITIES OF THE CONGESTION TAX

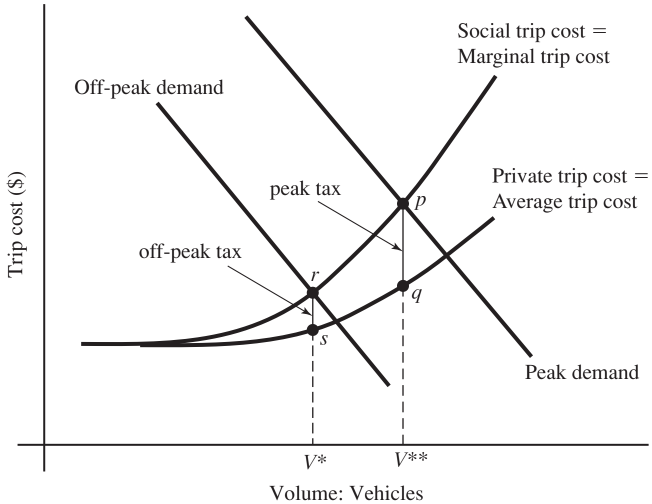
We've seen that the congestion tax internalizes an externality, leading to efficiency gains and urban growth. In this part of the chapter, we'll discuss several practical issues concerning the implementation of a system of congestion taxes. We address three questions.

1. How would the perfect congestion tax vary by time of day?
2. How high would a congestion tax be?
3. What sort of experiences do cities have with pricing urban travel?

Peak versus Off-Peak Travel

The congestion tax equals the gap between the private and social cost of travel, and it varies across time and space. As shown in Figure 10–5, during peak travel times, the demand curve is relatively high, generating a large gap between the private and social cost of travel and thus a large congestion tax. In contrast, during off-peak periods when demand is relatively low, the congestion tax is low as well. Historically,

FIGURE 10–5 Congestion Tax in Peak versus Off-Peak Periods



During the peak travel period, traffic volume is relatively high, generating a large gap between the private and social cost of travel (shown by points *p* and *q*) and thus a higher congestion tax. During the off-peak period, the gap between the social and private cost of travel is lower (points *r* and *s*), so the congestion tax is lower.

commuting trips have been concentrated during the morning peak period (6:30–8:30 a.m.) and the evening peak period (4:30–6:30 p.m.).

The term "rush hour" refers to peak travel periods, but the phrase is ironic rather than literal. Although people may be in a rush, they move slowly. In modern cities, traffic volume is high and traffic is slow for most of the workday, not just for a couple of hours. In cities with populations of at least 1 million, there is no midday break in congestion: Travel speeds start to drop early in the morning and continue to drop during the day until they rise after about 7 p.m. In medium-sized cities (populations of 500,000 to 1 million), there is a midafternoon lull in traffic (1 p.m. to 4 p.m.). The term "rush hour express" applies to subways, and in Japan, rush hour workers called "fanny pushers" (euphemistic translation) are hired to pack people into rush-hour trains.

Estimates of Congestion Taxes

The efficient congestion tax varies across space and time. According to Parry and Small (2009), for U.S. metropolitan areas as a whole, the efficient tax in 2005 was about $0.056 per mile. The efficient tax is higher for peak-period travel: $0.085 per mile. These overall averages obscure substantial differences across metropolitan areas in the efficient congestion tax.

TABLE 10–4 Congestion Taxes in Selected Metropolitan Areas

	Washington, DC	Los Angeles, CA	London
Peak period tax (per mile)	$0.21	$0.26	$1.23
Off-peak tax (per mile)	$0.02	$0.03	$0.49

Source: Adapted from Parry and Small (2009).

Table 10–4 shows estimated congestion taxes for three metropolitan areas, Washington DC, Los Angeles, and London. For the peak travel period, the tax per mile is $0.21 in Washington, $0.26 in Los Angeles, and $1.23 in London. As expected, the off-peak taxes are lower. In highly congested London, the difference in traffic volumes between peak and off-peak travel is relatively small, so the off-peak tax is $0.49 per mile.

Implementing Road Pricing: Tolls and HOT Lanes

Modern technology allows the efficient and convenient collection of taxes for using congested roads. Under a vehicle identification system (VIS), each car is equipped with a transponder—an electronic device that allows sensors along the road to identify a car as it passes. The system records the number of times a vehicle uses a congested highway and sends a congestion bill to the driver at the end of the month. For example, if the congestion tax is $0.21 per mile, a driver who travels 10 miles along the highway 20 times per month would pay a monthly congestion bill of $42.00 (20 times $2.10). An alternative approach, which avoids issues of privacy, is to use anonymous debit cards to charge for driving on congested roads.

Singapore was the first city to use prices to control the volume of traffic. Under the Area Licensing System (ALS) implemented in 1975, drivers were charged about $2 per day to travel in a toll zone in the central area of the city. In 1998, Singapore switched to Electronic Road Pricing (ERP), a smart-card system with charges that increase with the level of congestion. The system features 28 gantries that charge users for entering the central area during the daytime, and 14 tolled highways that are subject to tolls during the morning peak period. There are no charges for travel on the weekends.

Many U.S. cities have designated highway lanes for use by high-occupancy vehicles—buses and carpools (two or more riders). The idea behind an HOV (high-occupancy vehicle) lane is to encourage carpooling, but if not many solo drivers switch to carpooling or buses, the HOV lane will be underutilized and the other lanes will become more congested. A recent response to this problem is to designate lanes for use by either high-occupancy vehicles or solo drivers who are willing to pay a toll. These express lanes or HOT lanes (for high-occupancy or toll) are used in Los Angeles, San Diego, Houston, and Minneapolis-St. Paul, and are under consideration in many other cities. In San Diego, the toll varies with the level of congestion in order to maintain a target speed. The toll is typically between $0.50 and

$4.00, but has been as high as $8.00 (Small and Verhoef, 2007). The use of HOT lanes is relatively high for commuters, high-income people, highly educated people, women, and people between 35 and 45 years of age.

There are important trade-offs in pricing express lanes. If the price is relatively high, the express volume will be relatively low, and congestion in the other lanes will be relatively high, as in the case of a traditional HOV lane. As the price of the express lane decreases, some travelers will switch to the express lane, decreasing congestion in the regular lanes. In general, an express lane sorts travelers with respect to their willingness to pay for speed: travelers with a relatively high opportunity cost of travel time will use the express lane, while those with a relatively low opportunity cost will use the slower regular lane.

A recent study explored the effects of express lanes in Orange County, California (Small, Winston, and Yan, 2005, 2006). A 10-mile stretch of State Route 91 has four regular lanes and two express lanes in each direction. The toll for the express lanes varies with traffic volume, and there are discounts for carpools. During the study, the peak toll was $3.30, and the average time saved for a 10-mile trip in the express lane was about 3.4 minutes. The choices of the travelers revealed the following.

- Travelers vary in time cost of travel, with a median value of $21.46 per hour, or about 93 percent of the average wage rate.
- The price elasticity of demand for the express lane is –1.59: a 10 percent increase in price decreases the number of travelers using the express lane by 15.9 percent.
- The elasticity of demand with respect to travel time in the regular (free) lanes is 0.73: a 10 percent increase in travel time increases the number of express travelers by 7.3 percent.

ALTERNATIVES TO A CONGESTION TAX

A number of alternative congestion policies have been proposed. To set the stage for a discussion of the alternatives, consider the four ways that a congestion tax decreases traffic volume:

1. **Modal substitution.** The tax increases the cost of single-driver travel relative to carpooling and mass transit (buses, subways, light rail), causing some travelers to switch to other travel modes.
2. **Time of travel.** The tax is highest during the peak travel periods, causing some travelers to travel at different times. Because work and school schedules are relatively inflexible, commuters and students would be less likely to change their travel times than other travelers (e.g., shoppers). Nonetheless, firms would have an incentive to change work schedules to allow their workers to avoid costly travel during the peak periods.
3. **Travel route.** The congestion tax is highest on the most congested routes, causing some travelers to switch to alternative routes.

4. **Location choices.** The congestion tax increases the unit cost of travel (travel cost per mile), causing some commuters to decrease their commuting distances. Some workers may move closer to their jobs, and others may switch to jobs closer to their residences.

These four responses cause us to move up the travel-demand curve as the cost of travel increases. In Figure 10–3, the congestion tax decreases traffic volume from 1,600 to 1,400 because it changes travel modes, times, routes, and distances.

Gasoline Tax

One alternative to the congestion tax is a gasoline tax. The simple idea is that if the cost per mile of travel increases, people will drive less. The problem is that a gas tax increases the cost of all automobile travel, not just travel along congested routes during peak periods. A gas tax decreases the relative cost of alternative travel modes, causing modal substitution in the right direction (#1 above). It also increases the cost per mile traveled, affecting location choice in the right direction (#4). But the gas tax fails to affect the time of travel or the travel route (#2 and #3), except to the extent that congestion generates lower gas mileage.

It may be tempting to conclude that getting two responses (mode and location) out of four isn't so bad. But consider the gasoline tax required to internalize congestion externality for peak-period congestion. If the appropriate congestion tax is $0.21 per mile and the average vehicle gets 20 miles per gallon of gasoline, the required gas tax would be $4.20 per gallon. The problem is that the tax would apply to all gasoline purchased, not just the gasoline used during the peak period on congested roads. As we'll see later in the chapter, there are some environmental benefits from taxing gasoline, but the appropriate tax is much less than $4.20 per gallon. It is worth noting that a tax of $4.20 would cause the price of gasoline in the United States to be close to the prices in several countries in Western Europe.

Subsidies for Transit

Another alternative to a congestion tax is to subsidize mass transit. The basic idea is to match the underpricing of car travel with equivalent underpricing of buses, subways, commuter trains, and light rail. Transit subsidies change modal choice (#1) in the right direction but don't directly affect the time of travel, travel routes, or location choice. Although a transit subsidy will decrease the volume of automobile travel and reduce congestion, it will never be as efficient as a congestion tax. A fundamental problem of a matching subsidy is that it causes transportation in general to be underpriced, leading to excessive amounts of travel. In the next chapter of this book, we will return to the issue of subsidies for mass transit, exploring the rationale for the subsidies, one of which is to counteract the distortions caused by the underpricing of automobile commuting.

The Pricing of Parking

The conventional pricing of workplace parking generates a relatively low cost of driving to work. Many firms provide free parking to their employees: a firm absorbs the cost of parking facilities as a cost of doing business. An alternative approach is to charge employees for parking, and to use the money raised in parking fees to increase the wages of all workers, including workers who carpool, ride mass transit, walk, or bike. This approach, known as "cashing out" free parking, increases the cost of driving to work and encourages workers to use alternative commuting modes. In Figure 10–3, eliminating this price distortion would shift the demand curve for solo driving to the left, decreasing the equilibrium traffic volume and decreasing the cost associated with congestion.

There is evidence that commuters respond to changes in the price of parking. Shoup (1998) estimates that employer-paid parking shifts 25 percent of all commuters into solo driving and increases the number of cars driven to work by 19 percent. When the city of Ottawa, Canada, increased parking rates for government employees from zero to 70 percent of the commercial rate, the number of workers driving to work decreased by 23 percent, the automobile occupancy rate increased from 1.33 to 1.41, and bus ridership increased by 16 percent (DiRenzo, Cima, and Barber, 1981). At four workplaces in Los Angeles that eliminated free parking, the number of solo drivers decreased by 19 percent to 81 percent (Small and Verhoef, 2007). In case studies of the effect of a law requiring California employers to cash out free parking, the number of solo drivers decreased by an average of 17 percent.

What is the actual social cost of parking in urban areas? Small and Verhoef (2007) estimate the daily cost of commuter parking for different sites in a metropolitan area. The estimated daily cost is \$4.44 for a suburban surface lot, \$9.18 for a suburban structure, and \$15.04 for an urban structure. Based on the suburban figures, they estimate that the commuter parking cost per mile traveled is \$0.28. The cost per mile is relatively large because (*a*) the opportunity cost of land used for commuter parking is relatively high, and (*b*) the fixed cost of commuter parking lots and structures are spread over commuting trips, not all trips.

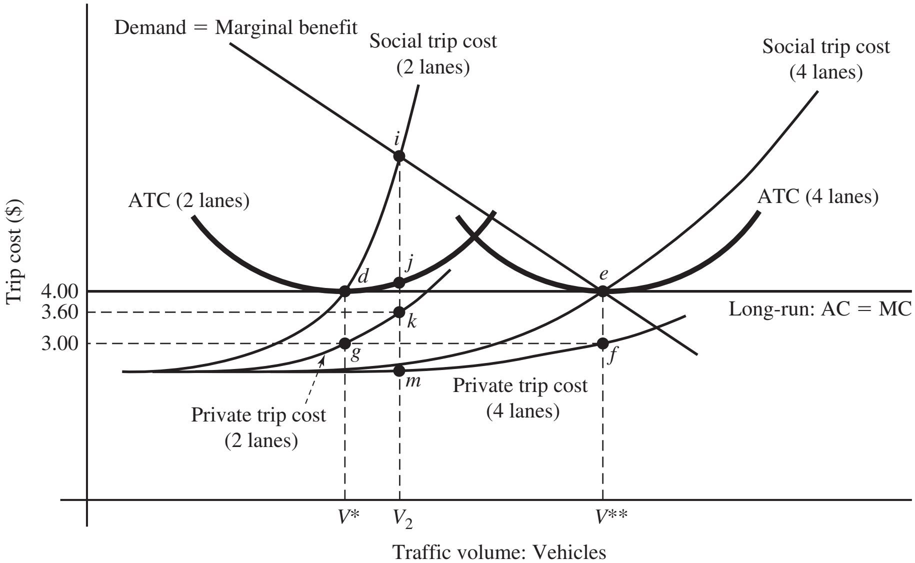
THE ROAD CAPACITY DECISION

How could a government determine the socially efficient capacity of a road? As we'll see, a government that imposes a congestion tax can use a simple rule to decide on road width: If the total revenue from the congestion tax exceeds the cost of building the road, the government should build a wider road. The optimum road width generates just enough congestion-tax revenue to pay for the road.

Interpreting the Spaghetti Cost Curves

The demonstration of this convenient rule requires some background on the different components of travel and road costs. Figure 10–6 shows two sets of cost curves, one for a two-lane road and a second for a four-lane road. At first glace, the figure looks like a pile of unruly spaghetti, but there is some logic to the curves.

FIGURE 10–6 Expand Capacity until Congestion Tax Revenue Equals Road Cost



The congestion tax per vehicle is the gap between the private and social trip cost (the gap between points *i* and *k* for a two-lane road). The average road cost is the gap between the average total cost and the private trip cost (the gap between *j* and *k* for a two-lane road). For a two-lane road, tax revenue per vehicle exceeds the average road cost, so capacity should be expanded. For a four-lane road, the congestion tax is shown by the gap between points *e* and *f*, and so is the average road cost: Congestion taxes pay for the optimum road. The four-lane road is socially efficient because at point *e*, the long-run marginal cost ($4) equals the marginal benefit (from the demand curve).

The curves labeled ATC show the average total cost of travel. These curves include the cost of building the road and the private cost of travel (shown in earlier figures). For example, ATC (two lanes) is a U-shaped curve that reaches a minimum at point *d*, with a volume of *V** and an average cost of $4. As traffic volume increases, two conflicting forces affect average total cost:

- **Decrease in average road cost.** The cost of building a two-lane road is fixed, and the larger the volume, the lower the road cost per vehicle.
- **Increase in private trip cost.** Once the congestion threshold is passed, private trip cost increases.

The ATC curve is U-shaped because the decrease in the average road cost dominates for small volumes, but the rising trip cost dominates for large volumes. The other important feature of the ATC curve is that the gap between ATC and the private trip cost curve is the average road cost (average fixed cost).

The cost curves for the four-lane road show the benefits of building a road with a larger capacity. The average total cost curve reaches the minimum at twice the volume (*V*** is twice as large as *V**). In addition, the private trip cost curve stays horizontal for twice the traffic volume and is lower for all volumes. Similarly, the social trip cost curve for the four-lane road is lower than the associated curve for the two-lane road. However, the four-lane road costs twice as much to build.

Let's start at the equilibrium with a two-lane road. If the government imposes a congestion tax, the equilibrium is shown by point *i*, with a traffic volume of V_2. The congestion tax is shown by the gap between the social trip cost (point *i*) and the private trip cost (point *k*). The average road cost is shown by the gap between the average total cost curve (point *j*) and the private trip cost (point *k*). The congestion tax per vehicle exceeds the road cost per vehicle, meaning that total congestion tax revenue exceeds the cost of building the road.

Widen the Road If Congestion Tax Revenue Exceeds the Cost

The excessive revenue from congestion tax is a signal to the government to build a wider road. To double the capacity, the government could add two lanes, making a four-lane road. The new equilibrium is shown by point *e*, where the social trip cost curve intersects the demand curve. Now the congestion tax is shown by the gap between points *e* and *f*. This is also the gap between the private trip cost and the average total cost, or the average road cost. In other words, for a four-lane road, the congestion tax equals the average road cost, so total tax revenue is just enough to pay for the road. The people who use the road pay the entire cost of building it.

The capacity rule tells the government to widen the road to the point where congestion tax revenue equals the cost of the road. In addition to its appeal in terms of fairness and equity (road users pay), this rule generates the socially efficient road width. Recall that the average cost curve for the four-lane road reaches the same minimum average cost ($4) as the two-lane road but with twice the capacity: V^{**} is twice V^{*}. This is sensible if there are constant returns to scale in road building. If so, we can build twice as many lanes at twice the cost (the average road cost doesn't change) and handle twice as much traffic at the same private trip cost (the trip cost per vehicle doesn't change). In other words, the long-run average cost (including road cost and travel cost) is constant. If the average cost is constant, the average cost equals the marginal cost, as shown by the horizontal cost curve at $4.

Point *e* is the efficient outcome because it satisfies the marginal principle, equating the marginal benefit of travel to the marginal cost. At point *e* on the demand curve, the marginal benefit is $4. The marginal cost is shown by the horizontal long-run marginal-cost curve (equal to average cost), which is constant at $4. If the government were to build a wider road, the marginal cost would still be $4, but the demand curve tells us that for volume above V^{**}, the willingness to pay is less than $4. Additional travelers are not willing to pay the full cost of widening the road to accommodate them, so it would be inefficient to widen the road. In contrast, when we start with a two-lane road (at point *i*), widening the road is socially efficient because the willingness to pay for an additional vehicle exceeds the cost of expanding the road to accommodate the vehicle.

Capacity Expansion and Latent Demand

There are many anecdotes about roads that did not experience less congestion after they were widened. The reason is that the demand for peak-period travel is highly elastic. Many travelers avoid using congested roads because travel is so slow. But when a road is widened and travel initially moves faster, travelers who were deterred by slow speed start using the road. This is the phenomenon of "latent demand." In the language of Small (1992), there is a "reserve army of the unfulfilled" that will switch to a previously congested highway as soon as an increase in capacity increases travel speeds. This latent demand may fill most or all of the new capacity during peak periods.

In Figure 10–6, the widening of the road leads to a moderate reduction in the private trip cost. With an initial volume of V_2, point *k* tells us that the private trip cost is \$3.60. Doubling the road width increases volume to V^{**} and, as shown by point *f*, the private trip cost is \$3.00. If we ignored the demand side of road travel, we might imagine that a doubling of road capacity would lead to a larger reduction in travel time and private trip cost. For example if we assumed that demand was perfectly inelastic (it stays at V_2), we would go from point *k* to point *m*, with a much lower trip cost. But because consumers respond to lower travel costs by traveling in larger numbers, travel time and trip costs decrease by a smaller amount.

Who Pays for Roads?

The use of congestion taxes to pay for roads is appealing for reasons of equity and efficiency. In the United States, roads are actually financed by various user fees. Auto and truck drivers pay federal and state taxes on gas, oil, and auto parts. In addition, truck drivers also pay user fees based on the weight of the truck and the miles traveled. During the 1960s, revenue from user taxes exceeded the highway cost by about 25 percent, but since then gasoline taxes have not kept pace with inflation, so revenue from user fees no longer covers the cost of roads and highways. Historically, urban road users have done better than rural travelers in covering the cost of roads.

AUTOS AND AIR POLLUTION

Automobile use causes two sorts of environmental externalities, air pollution and greenhouse gases. Motor vehicles emit volatile organic compounds (VOC), carbon monoxide (CO), nitrogen oxides (NOx), and sulfur dioxide (SO_2). VOCs react with NOx in the atmosphere to form ozone (O_3, aka smog) and also generate particulate matter. In the United States, transport activities are responsible for about two-thirds of CO emissions, about half of VOC emissions, and about two-fifths of NOx emissions (Small and Kazimi, 1995). Poor air quality can exacerbate respiratory problems and cause premature death. As we saw earlier in the book, urban air quality has generally improved in the last 20 years because lower emissions per mile driven have more than offset increases in mileage. Automobile travel also generates greenhouse gases, contributing to global climate change.

Internalizing the Externality

The economic approach to air pollution is to internalize the externality. A tax equal to the marginal external cost of pollution would cause drivers to incorporate the full costs of driving into their decisions, leading to the socially efficient level of pollution. A pollution tax would encourage people to buy cleaner cars and drive fewer miles. The direct approach would be to install a monitoring device in each car to measure its emissions and then charge the owner for the emissions.

An alternative approach is to impose a one-time pollution tax on each new car. The tax would vary by car model and would equal the estimated lifetime emissions times the external cost per unit of emissions. For example, if a particular model is expected to emit 5,000 units of pollutants over its lifetime and the external cost per unit of pollution is \$0.20, the one-time pollution tax would be \$1,000. Under this system, car buyers would have an incentive to buy cleaner cars, but would not have an incentive to drive less once they buy a car.

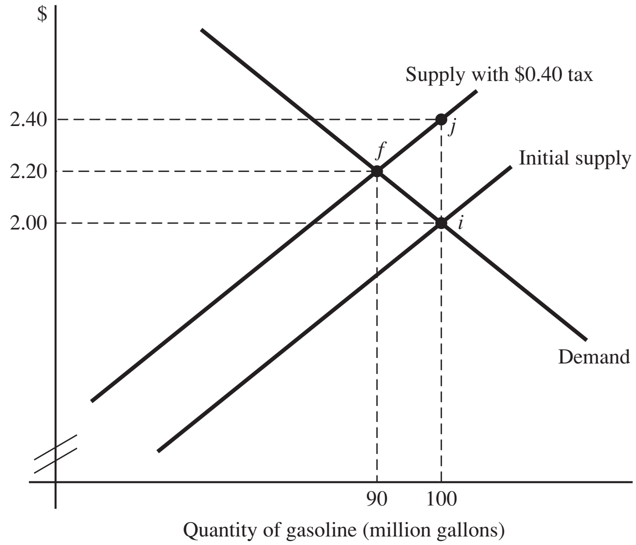
A Gasoline Tax

Another approach is to use a gasoline tax to increase the private cost of auto travel. The tax would increase the cost per mile driven, so it would decrease the total miles driven and decrease air pollution. The problem with a gas tax is that every driver would pay the same tax per gallon, regardless of how much pollution is generated per gallon of gasoline. So a gasoline tax would decrease pollution by decreasing miles driven but would not encourage people to drive cleaner cars. Of course, if government emissions standards generate relatively small differences in emissions per gallon across car models, the lack of incentives to buy clean cars would be less of an issue.

Figure 10–7 shows the effects of using a gasoline tax to internalize the externalities from air pollution. Point *i* shows the initial equilibrium, given a supply curve that does not include any pollution or greenhouse taxes. The equilibrium price is \$2.00 and the quantity is 100 million gallons. Small and Kazimi (1995) estimate that the pollution-related external cost is about \$0.02 per vehicle mile (or about \$2,400 over the life of a car). With an average gas mileage of 20 miles per gallon, this translates into a gas tax of \$0.40 per gallon. In Figure 10–7, this pollution tax shifts the supply curve upward by \$0.40, increasing the equilibrium price from \$2.00 to \$2.20 and decreasing the equilibrium quantity from 100 to 90 million gallons.

In Figure 10–7, the increase in the equilibrium price is half the tax. This is consistent with studies of the effects of gasoline taxes on the price of gasoline (Chouinard and Perloff, 2004). Consumers pay half the tax, and the rest is paid by the people who supply the scarce input used to produce gasoline, crude oil. In Figure 10–7, the tax decreases the quantity of gasoline consumed by 10 percent, so it decreases the demand for crude oil. A decrease in the demand for crude oil decreases the equilibrium price, so part of the gasoline tax is shifted backward onto the people who own oil wells in Texas, Saudi Arabia, and other oil-producing areas.

FIGURE 10–7 Market Effects of a Gasoline Tax



A tax of $0.40 per gallon of gasoline shifts the supply curve upward by $0.40, increasing the equilibrium price and decreasing the equilibrium quantity. Half the tax is borne by consumers, and half is shifted backward onto the people who supply the crude oil used to produce gasoline.

An alternative approach for reducing auto pollution is to subsidize mass transit. A transit subsidy would encourage people to switch to buses and subways, which generate less pollution per passenger. As we saw earlier in the case of congestion externalities, using transit subsidies to reduce auto traffic presents some problems. Travelers are not very responsive to changes in the price of transit, so a transit subsidy wouldn't reduce automobile pollution by very much. In addition, matching the underpricing of automobiles with underpricing of transit leads to a general underpricing of transportation, which in turn leads to a misallocation of resources, with too much labor, capital, and energy allocated to transportation.

Greenhouse Gases and a Carbon Tax

The environmental and economic consequences of accumulating greenhouse gases remain uncertain. As carbon levels rise, scientists expect crop losses as well as substantial costs to protect coastal regions from rising waters, but quantifying these consequences is difficult. In terms of environmental policy, the key number is the external cost per ton of carbon emitted, which determines the appropriate carbon tax. The current estimates are in the range of $25 to $100 per ton. A carbon tax of $50 per ton translates into a gasoline tax of $0.13 per gallon. We could extend Figure 10–7 to show the effects of carbon-based gasoline tax. A $0.13 tax would shift the supply curve upward by \$0.13, increasing the equilibrium price and decreasing the equilibrium quantity. The price would increase by about half the tax (\$0.065), meaning that about half the tax would be shifted backward onto the suppliers of crude oil.

MOTOR VEHICLE ACCIDENTS

A third externality from the use of the automobile is motor-vehicle accidents—collisions with other vehicles, bicycles, and pedestrians. Collisions result in property damage, injuries (3.1 million per year in the United States), and deaths (about 40,000 per year in the United States). A recent study estimates that the annual cost of vehicle collisions is over \$300 billion, or more than \$1,000 per person (Miller, 1993). The externality occurs because when one person's driving decisions lead to a collision, roughly one-third of the costs are borne by someone else.

The accident-related external cost of driving depends on the miles driven. The more you drive, the more likely you are to collide with someone else, generating costs for you and the other person. Of course, both the likelihood of a collision and the consequences depend on traffic conditions as well as how carefully people drive. Parry (2004) suggests that the accident-related external cost of travel is about 4.4 cents per mile driven. By way of comparison, the fuel cost per mile is about 10 cents. In this part of the chapter, we'll explore the effects of policies that improve vehicle safety. We'll also discuss a proposed policy under which people would pay for each mile they drive.

Vehicle Safety Policies: Bikers Beware

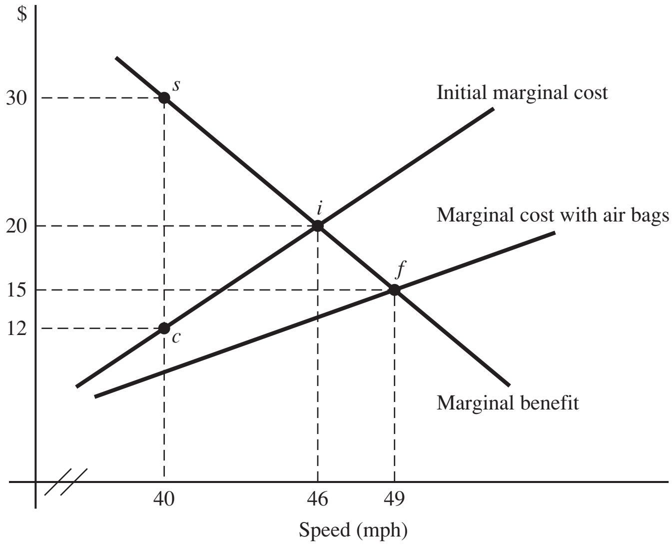
The Vehicle Safety Act of 1966 established safety standards for new cars, and subsequent legislation has extended the standards. Among the mandated features are head restraints, padded dashboards, seatbelts, shatterproof windshields, dual braking systems, collapsible steering columns, and air bags. These safety features add about \$1,000 to the price of a car (Small, 1997).

Around the world, dozens of countries have laws that require car occupants to wear seat belts. Studies of these and other vehicle-safety laws uncover two puzzles:

1. Death rates among car occupants were predicted to drop significantly but instead decreased by a relatively small amount.
2. The death rates for pedestrians and bicyclists increased.

The theory of risk compensation (Peltzman, 1975) explains these puzzles. The idea is that in deciding how fast to drive, a person weighs the benefits and costs and chooses a speed that maximizes his or her utility. A mandated safety feature like seat belts decreases the cost of driving fast—crash injuries are less severe—so people drive faster and experience more collisions. The increase in the frequency of collisions partly offsets the fact that injuries are less severe. In addition, faster driving means higher death rates for pedestrians and bicyclists.

FIGURE 10–8 Speed and Safety Regulations



A traveler chooses the speed at which the marginal benefit (the value of time saved) equals the marginal cost (the expected accident costs). The introduction of a safety regulation such as air bags decreases the marginal cost and increases the speed from 46 mph (point *i*) to 49 mph (point *f*).

We can use a simple example of travel speed to explore the effects of mandated safety features on risk taking. Duke drives every Saturday night to a dance hall in Hazard City and must decide how fast to drive. The benefit of speed is that he spends less time on the road, leaving him more time to dance with Daisy. In Figure 10–8, the marginal benefit curve is negatively sloped, reflecting the diminishing marginal benefit of dancing time. For example, as shown by point *s*, traveling at 40 miles per hour (mph) instead of 39 gives him additional dance time that he values at $30. Farther down the marginal-benefit curve, traveling at 46 mph instead of 45 mph gives him additional dance time that he values at only $20.

The cost of speed is that driving faster increases the likelihood of a collision and the severity of injuries. The positively sloped marginal-cost curve shows that the cost of speed rises at an increasing rate. When Duke speeds up from 39 to 40 mph, the likelihood and severity of injuries increases by a moderate amount, increasing the expected injury cost of the trip by $12 (shown by point *c*). When he speeds up from 45 to 46 mph, the likelihood and severity of injuries are larger, and the expected injury cost increases by $20 (point *i*).

The initial equilibrium is shown by point *i*, where the initial marginal-cost curve intersects the marginal-benefit curve. If Duke tentatively chooses a slower speed, say 39 mph, he could do better. As shown by points *s* and *c*, speeding up to 40 mph generates a bigger benefit ($30 worth of extra dance time) than a cost (an increase of $12 in expected injury costs). He will continue to speed up until he reaches point *i*, with a speed of 46 mph. He doesn't drive faster because the marginal benefit of additional speed is less than the marginal cost: The extra dance time is not worth the large increase in the expected injury cost.

How would mandated safety equipment affect Duke's choice of speed? Suppose the government requires air bags in all cars. The air bags reduce the severity of injuries from a collision, so the expected cost of driving fast decreases. In Figure 10–8, the marginal-cost curve shifts downward, and the marginal principle is now satisfied at point *f*. Without the air bag, the cost of going between 47 and 49 mph was higher than the benefits of the extra dance time; with the air bag, the cost is lower, so Duke drives faster. He compensates for the lower cost of risky behavior (driving fast) by driving faster, accepting a higher likelihood of a collision because he knows that the injuries suffered in a collision will be less severe.

There is evidence of risk compensation in response to mandated vehicle safety features. Peltzman (1975) notes that collision rates were higher than expected in the years following the implementation of safety regulations, and pedestrian death rates were higher too. Crandall et al. (1986) show that the death rates for pedestrians and bicyclists are positively related to an index of safety features, suggesting that drivers in safer cars take more risks and endanger others. Overall, the vehicle safety features have decreased traffic deaths because the decrease in driver deaths exceeds the increase in the deaths of pedestrians and bicyclists.

Pay to Drive Policies

In recent years, a new sort of policy has been proposed to deal with the problem of vehicle collisions. Since the external cost of driving depends on the miles driven, it is natural to consider imposing a per-mile tax on driving, known as a vehicle miles traveled (VMT) tax. Table 10–5 shows the marginal external cost for different types of vehicles and drivers of different ages. The external cost of young drivers is over three times as high as the external cost of middle-aged drivers. The external cost is highest for pick-up trucks and lowest for minivans. The relatively high external cost for small cars reflects the greater usage of small cars by young drivers.

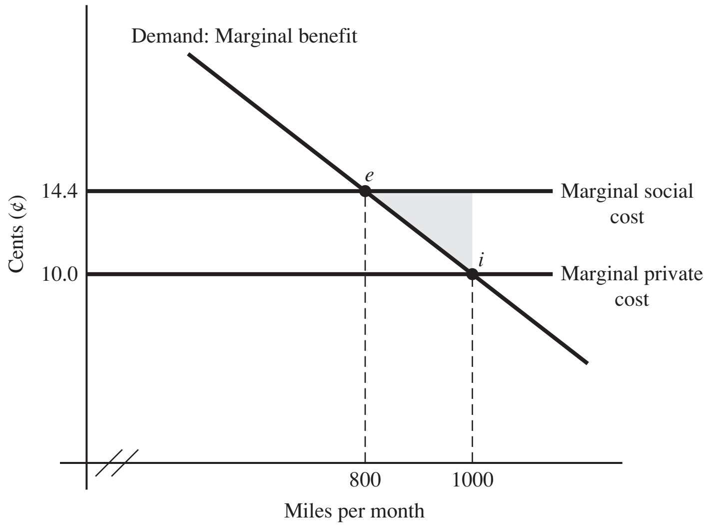
We can use the marginal approach to measure the inefficiency from accident externalities. The relevant concepts are reviewed in Section 1 of "Tools of Microeconomics," the appendix at the end of the book. In Figure 10–9, the demand curve

TABLE 10–5 External Accident Costs for Different Vehicles and Driver Ages

	Small Car	Large Car	SUV	Minivan	Pickup	<25 years	25–70 years	>70 years
Cents per mile	4.8	3.94	3.59	3.04	5.76	10.87	3.42	5.43

Source: Parry, Ian W.H. "Comparing Alternative Policies to Reduce Traffic Accidents." *Journal of Urban Economics* 56 (2004), pp. 346–68.

FIGURE 10–9 Accident Costs and VMT Tax



Automobile travel causes accidents, and the average external cost is 4.4 cents per mile. The individual decision about how many miles to drive is based on private costs, so the equilibrium number of miles (1,000 at point *i*) exceeds the socially efficient number (800 at point *e*). The shaded triangle shows the welfare loss from the accident-related underpricing of travel.

for driving shows the marginal benefit of driving. The lower marginal-cost curve is the private cost of driving, assumed to be a constant 10 cents per mile driven. Among the private costs are the costs of fuel and other costs that depend on miles driven. This does not include insurance premiums, which are fixed on an annual basis and do not increase with mileage. The upper cost curve is the marginal social cost of driving, the sum of the marginal private cost and a marginal external cost of 4.4 cents per mile driven.

The initial equilibrium is shown by point *i*. People base their driving decisions on the costs they bear themselves—the private cost. According to the marginal principle, a driver will pick the number of miles where the marginal benefit equals the marginal private cost, generating 1,000 miles in this example (point *i*). The socially efficient outcome is shown by point *e*, where the marginal benefit equals the marginal *social* cost. For miles 801 through 1,000, the marginal benefit experienced by the driver (shown by the demand curve) is less than the marginal social cost, so driving these miles is socially inefficient.

The shaded triangle shows the welfare loss to society from the underpricing of driving. Starting from point *i*, the last mile driven (1,000) generates a benefit of 10 cents and a social cost of 14.4 cents, so the net loss to society is the difference, or 4.4 cents. Similarly, for mile number 900, the benefit is 12.2 cents and the cost is 14.4 cents, for a net loss of 2.2 cents. Adding up the losses for the 801st through the 1,000th miles, the cost of driving exceeds the benefit by the gap between the

TABLE 10–6 Imperfect VMT Taxes

Type of Differentiation	Percent of Maximum Welfare Gain
None: Uniform tax	76
Age	98
Vehicle Type	78

social-cost curve and the demand curve. The shaded triangle shows the welfare loss from collision externalities.

The obvious solution to the externality problem is to get drivers to pay for the accident costs they impose on society. A tax of 4.4 cents per mile would internalize the externality, moving us from point *i* to point *e*. The gain to society is the welfare loss averted, shown by the shaded triangle. As shown in Table 10–5, the external cost varies across vehicles and driver ages, and a precise driving-tax policy would involve higher taxes on the vehicles and drivers that generate the highest external cost. Specifically, the highest driving taxes would be imposed on young drivers and pickup drivers.

Parry (2004) estimates the welfare gains from alternative pricing schemes for accident externalities. A perfect VMT tax would be differentiated according to driver age (the young pay 10.87 cents per mile) and vehicle type (pickup drivers pay 5.76 cents; minivan drivers pay only 3.04 cents). In this case, the estimated welfare gain is 0.38 cents per mile driven, or $9.4 billion per year at the national level. Table 10–6 shows the welfare gains from tax schemes with different degrees of differentiation. A uniform tax (no differentiation) generates a welfare gain about three-fourths of the maximum gain of $9.4 billion.

Parry also considers two alternatives to the VMT tax. First, if insurance premiums were based strictly on mileage traveled, the effects on travel and efficiency would be about two-thirds of the effects from a perfect VMT tax. Second, a gasoline tax fares poorly as a way to internalize accident externalities, with welfare gains of about one-quarter of the gains generated by a differentiated VMT tax. The basic problem is that a person's gas-tax bill depends on the amount of gasoline consumed, not the number of miles driven. Gas mileage varies across vehicles, so a gas tax is a blunt policy instrument to reduce miles traveled. In addition, consumers would respond to a higher gas tax by purchasing more fuel-efficient cars, further weakening the connection between the tax and accident externalities.

Accidents and Congestion

One of the external costs of traffic accidents is the congestion that results from blocked traffic. Parry (2004) estimates that traffic delays from accidents generate a cost of almost $5 billion per year in the United States. Although a VMT tax would decrease traffic volume and reduce traffic-snarling accidents, local governments would still need policies to respond to accidents that occur.

Cities have developed policies to respond quickly to accidents and clear disabled vehicles to restore traffic flow. Incident-response teams, equipped with strategically placed tow trucks, respond quickly to accidents. In some cities, tow trucks cruise the roads in anticipation of a radio call to clear up a nearby accident. Some cities have installed loop detectors in roadbeds that detect traffic slowdowns and alert officials immediately. Other cities use remote cameras to monitor traffic.

Some cities have developed special plans for dealing with congestion expected to occur during special events. In 1984, the city of Los Angeles anticipated severe congestion during the Olympics and developed a plan to use heavy-lift military helicopters to swoop in and remove disabled vehicles. An early fanciful report speculated that the helicopters would be equipped with large magnets, allowing them to pick up the disabled vehicles and transport them, dangling at the end of a cable, to the wrecking yard. It turns out that traffic was relatively light during the Olympics, so the helicopters were never actually used.

AUTOMOBILES AND POVERTY

In the chapter on neighborhood choice, we discussed the spatial mismatch between central-city workers and suburban jobs. The concentration of low-income workers in central cities, far from suburban jobs, leads to long commutes, low wages, and low employment rates. What is the possible role of automobiles in reducing the spatial mismatch?

Although many black central-city residents commute to suburban jobs, this sort of reverse commuting is costly and time-consuming because most low-income households don't own cars. About 27 percent of urban low-income families (with incomes less than $20,000) do not own a car, compared to 3 percent of urban households with incomes greater than $20,000. Among black workers living in central cities, 45 percent do not have access to a car. For low-skilled workers, having access to a car offers three benefits (O'Regan and Quigley, 1998):

- For a central-city worker who commutes to the suburbs by public transit, switching to an automobile would save about 19 minutes each day.
- Low-skilled workers with cars search for jobs over a much wider area and discover more job opportunities.
- Low-skilled workers with cars are more likely to complete job-training programs and get a job.

These studies have some important implications for welfare policy. In recent years, welfare policy has focused on moving people off welfare and into jobs. O'Regan and Quigley (1998) summarize the possible role of automobiles:

> If potential commute patterns of people coming off public assistance are similar to those of people currently in poor working households, government policy must pay more attention to auto ownership opportunities. . . . So programs that help job takers

obtain a used car—a secured loan for purchase, a leasing scheme, a revolving credit arrangement—may offer real promise, particularly in less dense and less centralized urban areas.

SUMMARY

Automobile travel generates externalities from congestion, pollution, carbon emissions, and accidents. Here are the main points of the chapter.

1. Automobile drivers base their travel decisions on private costs rather than social costs, so the equilibrium traffic volume exceeds the socially efficient volume.
2. A congestion tax internalizes the congestion externality.
3. Internalizing the environmental externalities from automobiles would require a tax of about $0.53 per gallon of gasoline.
4. People in safer vehicles drive less carefully, putting bicyclists and pedestrians at greater risk.
5. On average, the external cost of accidents is 4.4 cents per mile driven. The external cost is highest for young drivers and pickup trucks.

APPLYING THE CONCEPTS

For exercises that have blanks (________), fill each blank with a single word or number. For exercises with ellipses (. . .), complete the statement with as many words as necessary. For exercises with words in square brackets ([increase, decrease]), circle one of the words.

1. **Thump-Thump Data**
   Cities lay cables (plastic-coated sensors) in roadways to gather data on traffic volume and speeds, causing a thump-thump sound. Based on thump-thump data, you have concluded that at a volume (V) of 1,500 vehicles, the time required for a 10-mile trip is 24 minutes, compared to 23.98 minutes with V = 1,499 vehicles. The monetary cost of the trip is $1 and the opportunity cost of time is $0.10 per minute.
   a. At V = 1,500, the private trip cost is ________, computed as . . .
   b. At V = 1,500, the external trip cost is ________, computed as . . .
   c. At V = 1,500, the social trip cost is ________, computed as . . .
   d. Use a figure like Figure 10–3 to show an equilibrium volume of V = 1,500 and an optimum volume of V = 1,300. Provide as many relevant numbers as possible.
   e. On your graph use a curly bracket ({) to show external trip cost for the equilibrium volume and a brace ([) to show the optimum congestion tax.
   f. The optimum congestion tax is [larger, smaller] than the equilibrium external trip cost because . . .

2. Equilibrium versus Optimum

Consider the following table:

Volume	Trip Time	Increase in Travel Time per Driver	Increase in total Travel Time	External Trip Cost	Private Trip Cost	Social Trip Cost	Marginal Benefit (demand)
200	12.00	0	0	0	3.20	3.20	22.12
400	12.00	0	0	0	3.20	3.20	19.64
600	12.60	0.005	3.00	0.30	3.26	3.56	17.16
800	14.00	0.009	7.20	0.72	3.40	4.12	14.68
1,000	16.20	0.013	13.00	1.30	3.62	4.92	12.20
1,200	19.20	0.017	20.40	2.04	3.92	5.96	9.72
1,400	23.00	0.021	29.40	2.94	4.30	7.24	7.24
1,600	27.60	0.025	40.00	4.00	4.76	8.76	4.76
1,800	33.00	0.029	52.20	5.22	5.30	10.52	2.28
2,000	39.20	0.033	66.00	6.60	5.92	12.52	

a. The equilibrium volume = ________.
b. The optimum volume = ________
c. The congestion tax = ________
d. The equilibrium volume exceeds the optimum volume because . . .

3. Am I Stupid?

Upton and Dawn each commute a distance of 10 miles on a highway. The city has decided to implement a congestion tax of $0.50 per mile during commuting hours, and cut payroll taxes to make the change in tax policy revenue-neutral. Consider two responses to the congestion tax:

- Upton: "I will continue to commute on the highway, and I oppose the congestion tax. It would make me worse off by $5 per trip. What do you think I am, stupid?"
- Dawn: "I will stop using the to highway for commuting, and I oppose the congestion tax. It would make me worse off by $5 per trip. What do you think I am, stupid?"

a. If the tax makes Upton worse off, his loss will be [greater, less] than $5 because . . .
b. The tax will actually make Upton better off if . . .
c. If the tax makes Dawn worse off, her loss will be [greater, less] than $5 because . . .
d. The tax will actually make Dawn better off if . . .

4. Chamber of Commerce and the Congestion Tax

Consider Snarlsville, a city that will vote on a proposed revenue-neutral congestion tax. Consider the following statement from the Chamber of Commerce: "The tax will be bad for business, so we urge all business owners to vote NO." Suppose "bad for business" means "sales revenue will decrease."

a. The congestion tax will be [bad, good] for business because . . .
b. A landowner in Snarlsville should vote [no, yes] because . . .
c. A person who owns land in a nearby city should vote [no, yes] because . . .

5. **Substitution Effects of a Gas Tax**
A nation will vote on a new $2 per-gallon tax on gasoline, which will be combined with a $1,000 annual cut on payroll taxes.
   a. The best predictor of a person's vote on the gasoline tax is his or her ________.
   b. Petra purchased 500 gallons of gasoline last year. The tax will [increase, decrease, not affect] Petra's gasoline consumption because . . .
   c. Use the consumer choice model to illustrate your answer to (*b*).
6. **New Sellwood Bridge?**
Consider the issue of replacing the Sellwood Bridge in Portland, Oregon. The existing bridge has a free-flow capacity of 600 vehicles, and the private trip cost at this traffic volume is $2. Given the equilibrium volume of 1,000 vehicles per hour, the private trip cost is $3. The proposed replacement bridge has a free-flow capacity of 1,200 vehicles.
   a. Depict graphically a situation in which the equilibrium volume on the new bridge equals its free-flow capacity.
   b. The outcome shown in the graph will occur if the price elasticity of ________ is equal to ________, computed as . . .
7. **Latent Demand**
Consider the following quote: "According to the Peabody principle, an increase in highway capacity increases traffic volume by an amount sufficient to leave the private trip cost unchanged."
   a. Use a graph to depict a situation in which the Peabody principle is correct.
   b. As the price elasticity of demand for highway travel [increases, decreases] in absolute value, the Peabody principle becomes a more accurate prediction of reality because . . .
8. **Bikers against Seat Belts**
In a column of *The State Paper* (Columbia, South Carolina) on April 5, 2004, columnist John Monk describes the efforts of motorcycle riders to defeat a proposed law that would allow police to issue $25 tickets to automobile drivers and passengers who are not wearing seat belts. The law would not apply to motorcycles, yet the bikers showed up in groups of a dozen or more, some dressed in full biker regalia, to urge legislators to reject the law. It is sensible for bikers to oppose the proposed law because . . .
9. **Vaporville: Ma'am, Step Away from the Car**
Vaporville has helicopters that can deliver a vaporizing beam to instantly vaporize any vehicles involved in an accident, clearing the roadway at a cost of $1,200 per accident. Alternatively, a tow-truck system takes 20 minutes to clear an accident scene at a cost of only $200. To simplify, assume that an accident simply stops traffic until the disabled vehicles are removed. The opportunity cost of travel time is $0.10 per minute. As the vapo-gunner on the helicopter, you must decide when to use the vaporizer and when to wait for the tow truck. The helicopter pilot is a certified appraiser of used (and damaged) vehicles.

a. A peak-period accident stops 4,000 vehicles. The cost of the tow-truck system is $200 plus _______. The cost of the vaporizer system is $1,200 plus _______. If _______ is greater than _______, you should vaporize; if not, you should use the tow truck.

b. An off-peak accident stops 400 vehicles. It will always be sensible to use tow trucks because . . .

c. Suppose that it is possible to instantly determine which driver is at fault for a particular accident. The responsible driver must pay for either a vaporizer or a tow truck to clear up the accident. Describe a simple public policy that would provide the incentive for drivers to freely choose the socially efficient method to clear up the accident. The policy should work during both peak and off-peak periods.

10. Speed, Makeup Violations, and the Invisible Hands

Using Figure 10–8 as a starting point, show the effects of the following changes on the speed chosen by Duke of Hazard City.

a. Daisy is grounded for makeup violations, leaving Duke without his favorite dance partner. Duke's speed [increases, decreases, doesn't change] because . . .

b. The normal country band is replaced by the punk group Adam Smith and the Invisible Hands. For Duke, slam-dancing generates twice the utility as country-western dancing. Duke's speed [increases, decreases, doesn't change] because . . .

c. The legal speed limit is set at 40 mph, and the fine on a speeding ticket increases with the gap between the driver's speed and 40 mph. Specifically, the fine equals the speed gap times $100. The probability of being caught and fined is 0.10. Duke's speed [increases, decreases, doesn't change] because . . .

11. Youngsters Pay to Drive

The demand curve for automobile travel by the typical young driver (25 years or younger) has a vertical intercept of 100 cents per mile and a horizontal intercept of 200 miles per week. Initially, the cost of automobile insurance is a fixed weekly sum, independent of mileage. The average operating cost for driving (for gasoline, oil, maintenance, and repair) is constant at 20 cents. Use the data in Table 10–5 for the external accident cost (rounded to the nearest cent).

a. Use a graph like the one in Figure 10–9 to show the choices per mile the typical young driver.

b. In the initial market equilibrium, the distance traveled is _______ miles.

c. The marginal social cost of travel is _______ cents, and the socially efficient travel distance is _______ miles.

d. If the insurance company switches to a per-mile fee equal to the marginal external accident cost, the net gain to society per young driver is _______, computed as . . .

REFERENCES AND ADDITIONAL READING

1. Adams J. G. U. *Risk and Freedom: The Record of Road Safety Regulation*. Transport Publishing Projects, 1985.
2. Adams J. G. U. *Risk*. London: UCL Press, 1995.
3. Chouinard, Hayley, and Jeffrey M. Perloff. "Incidence of Federal and State Gasoline Taxes." *Economics Letters* 83 (2004), pp. 55–60.
4. Crandall, Robert W., Howard K. Gruenspecht, Theodore E. Keeler, and Lester B. Lave. *Regulating the Automobile*. Washington DC: Brookings Institution, 1986.
5. DiRenzo, J. B. Cima and E. Barber. "Parking Management Tactics." Vol. 3, *Reference Guide*. Washington, DC: U.S. Department of Transportation, 1981.
6. Harvey A.C., and J. Durbin. "The Effects of Seat Belt Legislation on British Road Casualties: A Case Study in Structural Time Series Modeling." *Journal of the Royal Statistical Society* 149 (1986), pp. 187–227.
7. Keeler, Theodore E., and Kenneth A. Small. "Optimal Peak-Load Pricing, Investment and Service Levels on Urban Expressways." *Journal of Political Economy* 85 (1977), pp. 1–25.
8. Kraft, Gerald, and Thomas Domencich. "Free Transit." In *Readings in Urban Economics,* ed. Matthew Edel and Jerome Rothenberg. New York: Macmillan, 1972, pp. 459–80.
9. Meyer, John R., and Jose A. Gomez-Ibanez. *Autos, Transit and Cities*. Cambridge, MA: Harvard University Press, 1981. Chapter 11, pp. 185–229.
10. Miller, T. R. "Costs and Functional Consequences of U.S. Roadway Crashes." *Accident Analysis and Prevention* 25 (1993), pp. 593–607.
11. Mohring, Herbert. "Congestion." Chapter 6 in *Essays in Transportation Economics and Policy,* eds. Jose A. Gomez-Ibanez, William B. Tye, and Clifford Winston. Washington DC: Brookings, 1999.
12. O'Regan, Katherine M., and John M. Quigley. "Cars for the Poor." *Access* 12 (Spring 1998), pp. 20–25.
13. Parry, Ian, and Antonio Bento. "Estimating the Welfare Effect of Congestion Taxes: The Critical Importance of Other Distortions within the Transport System." *Journal of Urban Economics* 51 (2002), pp. 339–65.
14. Parry, Ian, and Kenneth Small. "Should Urban Transit Subsidies Be Reduced?" *American Economic Review* 99 (2009), pp. 700–24.
15. Parry, Ian W. H. "Comparing Alternative Policies to Reduce Traffic Accidents." *Journal of Urban Economics* 56 (2004), pp. 346–68.
16. Peltzman, Sam. *Regulation of Automobile Safety*. Washington DC: American Enterprise Institute, 1975.
17. Shoup, Donald C. "Congress Okays Cash Out." *Access* (Fall 1998), pp. 2–8.
18. Small, Kenneth A. *Urban Transportation Economics*. Philadelphia: Harwood Academic Publishers, 1992.
19. Small, Kenneth A. "Urban Economics and Urban Transportation Policy in the United States." *Regional Science and Urban Economics* 27 (1997), pp. 671–91.
20. Small, Kenneth A., and Jose A. Gomez-Ibanez. "Road Pricing for Congestion Management: The Transition from Theory to Policy." In *Road Pricing, Traffic*

Congestion, and the Environment, eds. Kenneth J. Button and Erik T. Verhoef. Cheltenham, UK: Edward Elfar, 1998.

21. Small, Kenneth A., and Camilla Kazimi. "On the Costs of Air Pollution from Motor Vehicles." *Journal of Transport Economics and Policy* (1995).
22. Small, Kenneth and Erik Verhoef, *The Economics of Urban Transportation* (New York: Routledge, 2007).
23. Small, Kenneth, Clifford Winston, and Jua Yan. "Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability." *Econometrica* 73 (2005), pp. 1367–82.
24. Small, Kenneth, Clifford Winston, and Jua Yan. "Differentiated Road Pricing, Express Lanes, and Carpools: Exploiting Heterogeneous Preferences in Policy Design." *Brookings-Wharton Papers on Urban Affairs: 2006*, pp. 53–86.
25. Texas Transportation Institute. *2002 Urban Mobility Study*. mobility.tamu.edu/ums.
26. U.S. Department of Transportation. *Summary of Travel Trends, 2001 National Household Travel Survey.* Washington DC: U.S. Department of Transportation, 2004.

CHAPTER 11

Urban Transit

> *While real trolleys in Newark, Philadelphia, Pittsburgh, and Boston languish for lack of patronage and government support, millions of people flock to Disneyland to ride fake trains that don't go anywhere.*
>
> —Kenneth T. Jackson

In this second chapter on urban transportation, we explore the economics of urban mass transit—buses, light rail, and heavy rail. We will look at the choice of a travel mode from two perspectives: individual travelers and transportation planners. In the United States, less than 5 percent of commuters use mass transit, but transit ridership is higher in many large cities and among low-income commuters. Urban mass transit is heavily subsidized: the average subsidy is over 50 percent of operating cost. We'll discuss the rationale for transit subsidies and explore alternatives to traditional subsidies.

COMMUTING AND TRANSIT RIDERSHIP

Table 11–1 shows the means of transportation to work for the U.S. workforce. Overall, 4.7 percent of commuters use mass transit. In declining order of ridership are buses (53 percent of transit riders), subways or elevated trains (31 percent), and trains (11 percent). Looking back a few decades, transit ridership in 1950 was roughly twice the ridership in 2000. The share of workers using transit is 11 percent for central-city residents, compared to 2 percent for suburban residents. The transit shares for central-city workers are 47 percent in New York, 26 percent in Chicago, and 25 percent in Philadelphia.

Transit ridership varies substantially across metropolitan areas. In the New York metropolitan area, about 25 percent of workers use public transit. Three metropolitan areas have transit shares between 10 and 14 percent: Chicago, Washington DC, and Philadelphia. Eight metropolitan areas are in the trillion-mile club—areas where the annual passenger mileage is at least 1 trillion miles. At the top of the list is New York (18.4 trillion), followed by Chicago (3.7), Los Angeles (2.8), Washington DC (2.2), San Francisco (2.1), Boston (1.9), Philadelphia (1.5), and Seattle (1.0).

TABLE 11–1 Means of Transportation to Work, 2000

Travel Mode	Number of Commuters	Percent
Workers 16 years and over	128,279,228	100
Car, truck, or van	112,736,101	87.9
Drove alone	97,102,050	75.7
Carpooled	15,634,051	12.2
Public transportation	6,067,703	4.7
Bus or trolley bus	3,206,682	2.5
Streetcar or trolley car	72,713	0.1
Subway or elevated train	1,885,961	1.5
Railroad	658,097	0.5
Ferryboat	44,106	
Taxicab	200,144	0.2
Motorcycle	142,424	0.1
Bicycle	488,497	0.4
Walked	3,758,982	2.9
Other means	901,298	0.7
Worked at home	4,184,223	3.3

Source: U.S. Bureau of the Census. *Journey to Work*: 2000. Washington DC: U.S. Government Printing Office, 2004.

Together these eight metropolitan areas are responsible for four-fifths of the transit passenger miles among the nation's 38 largest metropolitan areas.

Transit ridership is relatively high among low-income households. In relatively small metropolitan areas (population less than 1 million), just over half of transit riders come from households with incomes of less than \$15,000. In these small metropolitan areas, only about 7 percent of transit riders come from households with income greater than \$50,000. In relatively large metropolitan areas (population greater than 1 million), about 25 percent of transit riders come from low-income households, while about 18 percent come from high-income households.

THE COST OF TRAVEL AND MODAL CHOICE

In this part of the chapter we look at modal choice from the perspective of the individual traveler. We imagine a city where a traveler can choose between three modes of travel: driving alone, riding a bus, or riding a train (heavy rail or light rail). The traveler's objective is to choose the mode of travel that minimizes the full cost of a trip, when the full cost includes both the monetary and time costs of travel. Studies of travel behavior tell us quite a bit about the magnitude of the various components of the cost of travel.

The full cost of a trip is the sum of the monetary and time costs of travel. The time costs can be divided into two parts, the time cost associated with accessing a vehicle, and the time cost of travel in the vehicle. The full cost is

$$\text{Trip cost} = m + T_a \cdot d_a + T_v \cdot d_v$$

where m is the monetary cost, T_a is access time, d_a is the marginal disutility of access time, T_v is in-vehicle time, and d_v is the marginal disutility of in-vehicle time. The first component of trip cost is the monetary cost.

- The monetary cost of travel (m) is either the transit fare or the cost of operating an automobile (the cost of gas, oil, and wear and tear).

The next component of trip cost is the cost of getting to and from the vehicle used in travel, and is equal to the product of the access time and the cost per minute of access time.

- The access time (T_a) is the time required to travel (*i*) between the home and the bus stop or rail station, and (*ii*) between the bus stop or rail station and the workplace.
- The marginal disutility of access time (d_a) is the dollar amount a traveler would be willing to pay to avoid one minute of access time. Another label for the marginal disutility of access time is the opportunity cost of access time. Studies of travel behavior suggest that on average, d_a is 80 percent of a traveler's wage. For example, if a traveler's wage is $18.00 per hour or $0.30 per minute, d_a = $0.24 per minute.

The next component of trip cost is the cost of time spent in the vehicle, and is equal to the product of the in-vehicle time and the cost per minute of in-vehicle time.

- The in-vehicle time (T_v) is the time spent in a car or a transit vehicle (bus, train, boat).
- The marginal disutility of in-vehicle time (d_v) is the dollar amount a traveler would be willing to pay to avoid one minute of in-vehicle time. Another label for the marginal disutility of in-vehicle time is the opportunity cost of in-vehicle time. Studies of travel behavior suggest that on average, d_v is 50 percent of a traveler's wage. For example, if a traveler's wage is $18.00 per hour or $0.30 per minute, d_v = $0.15 per minute.

The full cost of a trip is the sum of the three components: monetary cost, access time cost, and in-vehicle time cost.

It is worth emphasizing that marginal disutility of access time is larger than the marginal disutility of in-vehicle time. Travelers are willing to pay 80 percent of the wage to avoid an hour of access time, compared to 50 percent of the wage to avoid an hour of in-vehicle time. This has important implications for the design of transit systems. Consider a one-minute *decrease* in access time and a one-minute *increase* in in-vehicle time. Since the disutility of access time exceeds the disutility of in-vehicle time, the net effect is a decrease in the trip cost. To decrease access time, a transit authority could decrease the distance between bus stops or train stations (decrease walk time) or decrease the time between buses and trains (decrease wait time).

An Example of Modal Choice

To illustrate a traveler's modal choice, consider Carla, a commuter who travels 10 miles each way between her suburban home and her job in the central city (20 miles for the round-trip). She has three travel options: the automobile, the bus, and a fixed-rail transit system such as BART (San Francisco), Metro (Washington DC), or MARTA (Atlanta).

Table 11–2 shows the monetary and time costs of Carla's daily commute. We assume that Carla's wage is $0.30 per minute ($18.00 per hour). The marginal disutility of vehicle time is half her wage ($0.15 per minute) and the marginal disutility of access time is 80 percent of her wage ($0.24 per minute). We can divide the time cost of the commuting trip into two parts.

1. **Access Time.** Carla parks her car in her garage and in a parking lot close to her workplace, so the auto trip does not involve any access time. The bus trip involves 24 minutes of walking and waiting, while the rail trip involves twice as much access time. The access time for rail is higher because we assume that rail stations are more widely spaced than bus stops. To compute the access time cost, we multiply the access time by the marginal disutility (opportunity cost) of access time.
2. **In-Vehicle Time.** At 60 minutes for the round-trip, rail is the fastest mode because it operates on an exclusive right-of-way with infrequent stops. Next at 80 minutes is the automobile, which travels on congested roads. The bus travels on congested streets and stops to pick up passengers, so a bus trip is 10 minutes longer than an automobile trip. To compute the in-vehicle time cost, we multiply the in-vehicle time by the marginal disutility (opportunity cost) of in-vehicle time.

Adding up the time costs, we get $12.00 for the auto, $19.26 for the bus, and $20.52 for rail. The time costs for the transit options are higher because the marginal disutility of access time is so high. For example, although the rail option is faster and thus has lower in-vehicle cost ($9.00 versus $12.00 for the auto), this advantage is more than offset by a higher access cost of $11.52.

The eighth row of Table 11–2 shows the monetary costs of the three travel modes. The monetary cost of the auto trip is $4.00, based on an assumed operating cost of $0.20 per mile (including fuel, wear and tear, and mileage-related loss in market value). For most commuters in the United States, parking is provided free of charge, so this example does not include any parking costs. The monetary cost of a transit option equals its fare, which is assumed to be $3.00 for both the bus and rail.

In this simple example, the least costly commuting mode is the automobile. The cost advantage of the automobile is $6.26 over the bus and $7.52 over rail. Although the auto has a higher monetary cost, this disadvantage is more than offset by its lower access (walk and wait) costs. This simple example uses plausible numbers to show the large cost advantage of auto commuting, illustrating why, for the vast majority of commuters (over 75 percent in the U.S.), solo driving is the rational commuting choice.

For a minority of U.S. commuters, it is rational to commute by bus or rail. We can modify the simple example to see how the changes in the numbers could tip the balance in favor of the bus or rail.

1. **Lower income**. A decrease in the wage decreases the marginal disutility of travel time (the opportunity cost of travel time), reducing the advantage of the automobile.

Table 11–2 Example of Modal Choice

	Automobile	Bus	Rail
Access cost (walk and wait) ($)	0.00	5.76	11.52
T_a: Access time (minutes)	0	24	48
d_a: Marginal disutility per minute ($)	0.24	0.24	0.24
In-Vehicle cost ($)	12.00	13.50	9.00
T_v: In-vehicle time (minutes)	80	90	60
d_v: Marginal disutility per minute ($)	0.15	0.15	0.15
Total time cost ($)	12.00	19.26	20.52
Monetary cost ($)	4.00	3.00	3.00
TOTAL COST	16.00	22.26	23.52
Assumptions			
Wage per minute ($)	0.30		
Auto operating cost per mile ($)	0.20		
Length of trip (miles)	20		

2. **Improved transit service**. A decrease in bus travel time will decrease the cost of the bus relative to the automobile. For example, if the in-vehicle time of the bus decreased to match rail (60 minutes) and the bus access time dropped by a third (to 16 minutes), the bus will be less costly than the auto. Similarly, rail will be less costly than the automobile if the rail access time dropped to 16 minutes.
3. **Free transit?** The cost advantage of solo driving exceeds the transit fare ($3.00), so free bus service would not change the modal choice of our hypothetical commuter.
4. **Parking cost.** As we saw in the previous chapter, the full cost of parking for an urban workplace is about $15.00 per day. If our hypothetical commuter paid this cost, she would switch to the bus.
5. **Internalize auto externalities?** As we saw in the previous chapter, driving generates external costs from congestion, environmental damage, and accidents. To internalize the externalities in the typical city in the peak travel period, the driving tax would be $0.145 per mile (congestion tax = $0.085, environmental damage = $0.02, accident cost = $0.04). In our simple example, the driving tax would be $2.90 per day (20 miles times $0.145 per mile), which is almost half the gap between the cost of driving and the cost of the bus. Although our hypothetical commuter won't switch, a commuter with a cost gap less than $2.90 will switch.

Elasticities of Demand for Transit

A number of studies have measured the responsiveness of transit ridership to changes in prices and service. As documented by Small and Verhoef (2009), the overall price elasticity of demand for transit is −0.40. In other words, a 10 percent increase in fares will decrease the number of transit riders by about 4 percent. As shown in Table 11–3, the price elasticity of demand for transit varies across modes

TABLE 11–3 Price Elasticities of Demand for Bus and Rail Transit

Time of day	Bus	Rail
Peak demand	−0.40	−0.24
Off-Peak	−0.80	−0.48
Overall	−0.50	−0.30

Source: Adapted from Parry and Small (2009).

and the time of day. In general, the demand for bus travel is more price elastic than the demand for rail transit, and peak demand (during the periods of high volume of travel) is less price elastic than off-peak demand.

The overall inelasticity of demand has important implications for the transit operators. Since the elasticity is less than 1.0 in absolute value, there is a positive relationship between price and total fare revenue. Although an increase in the fare will decrease ridership, the percentage decrease in ridership will be smaller than the percentage increase in the fare, so total revenue will increase. In other words, fares and total fare revenue change in the same direction.

There is evidence that transit ridership is more responsive to changes in the quality of service than to changes in fares. A study of Boston commuters concluded that the elasticity of ridership with respect to travel times is −0.80, while the elasticity of ridership with respect to monetary cost (fares) is −0.50. One implication of these elasticities is that a simultaneous improvement in service and an increase in fares may increase ridership. For example, suppose a transit operator increased the speed and frequency of buses, decreasing the average trip time by 10 percent and increasing the fare by 10 percent. The improvement in service will increase ridership by 8 percent (elasticity with respect to time = −0.80), while the increase in the fare will decrease ridership by 5 percent (elasticity with respect to price = −0.50), leaving a net increase in ridership of 3 percent.

There is also evidence that transit ridership is more responsive to changes in access time than to changes in in-vehicle time. This is sensible because the marginal disutility of walking and waiting time is roughly 1.6 times the marginal disutility of in-vehicle time. Consider the effect of a one-minute decrease in access time and a one-minute *increase* in in-vehicle time. The time cost of transit will decrease because the marginal disutility of access time is greater. The decrease in the cost of transit relative to the cost of driving alone will increase transit ridership.

THE EFFICIENT VOLUME OF RIDERSHIP

In this part of the chapter we take the social perspective, and show how to determine the socially efficient volume of transit ridership. We'll start by looking at the general cost structure of mass transit—applicable to heavy rail, light rail, and buses. All three systems have capital costs (the cost of laying the rails and buying the

TABLE 11–4 Construction Cost for Urban Transit Systems

	Construction Cost (2005 U.S. dollars)	
	Per Route Mile ($ in millions)	Per Daily Trip ($ in thousands)
Heavy rail	202	35.73
Light rail	63	34.66
Exclusive busway		
Ottawa	53	3.41
Pittsburgh	28	6.44
Shared carpool lane		
El Monte Busway, Los Angeles	17	4.31
Shirley Highway northern Virginia	18	3.39
I-66, northern Virginia	32	9.73
Houston transit-ways (average)	8	6.50

Source: Small and Verhoef (2007).

vehicles), operating cost (labor, fuel, and maintenance), and rider cost (the disutility of travel or the opportunity cost of travel time).

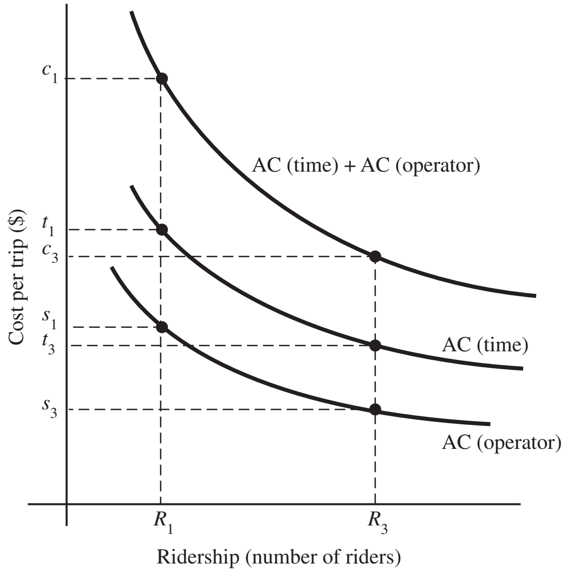
System and Rider Costs

Table 11–4 shows the construction cost for rail and bus systems. The cost per route mile is greatest for heavy rail, followed by light rail, busways (exclusive rights-of-way for buses), and shared carpool lanes. The cost per daily trip is computed by dividing the construction cost by the number of riders per day. Heavy rail and light rail have roughly the same cost per daily trip because although light rail has a lower cost per route mile, light rail also has lower ridership.

Figure 11–1 shows average-cost curves for a transit system. The lowest curve, AC (operator), shows the average cost of the transit operator for different levels of ridership. The curve is negatively sloped because of conventional scale economies: as volume increases, the fixed capital cost is spread over more riders. In Figure 11–1, a tripling of rider volume decreases the average operator cost from s_1 to s_3. The greater the fixed capital cost, the steeper the average-cost curve. A fixed-rail system has relatively high capital cost and thus a relatively steep average-cost curve. At the other extreme, the capital cost of a bus system is relatively low, so the average-cost curve is relatively flat.

The middle curve in Figure 11–1, AC (time), incorporates the disutility of riders' travel time. The curve is negatively sloped because of Mohring economies, named after the economist Herbert Mohring. Consider the effect of an increase in transit ridership along a particular bus route. Suppose the transit operator maintains the initial load factor (riders per bus) and accommodates the increased volume by running buses more frequently. An increase in service frequency decreases the access time cost, decreasing the rider trip cost. In Figure 11–1, a tripling of ridership decreases the average time cost from t_1 to t_3.

FIGURE 11–1 Scale Economies and Mohring Economies in Transit



The average operator cost curve is negatively sloped because of scale economies (spreading fixed capital cost), and the average time cost curve is negatively sloped because of Mohring economies (more frequent service with more riders).

We can illustrate Mohring economies with a simple example. Imagine a bus operator that provides just-in-time service. The operator waits to send a bus to a bus stop until the number of people waiting for a bus reaches some target quantity, say 24 riders. If commuters appear at a rate of one per minute, the bus headway (the time between buses) will be 24 minutes and the average access time will be 12 minutes (24/2). If ridership triples to three commuters per minute, the headway drops to 8 minutes and the average access time drops to 4 minutes. In other words, the increase in ridership decreases the average access time, decreasing the average time cost of riders.

The upper curve in Figure 11–1 shows the sum of average operator cost and average time cost. It is steeper than the other curves because an increase in volume decreases both the average operator cost and the average time cost. For example, a tripling of rider volume decreases the average cost from c_1 to c_3.

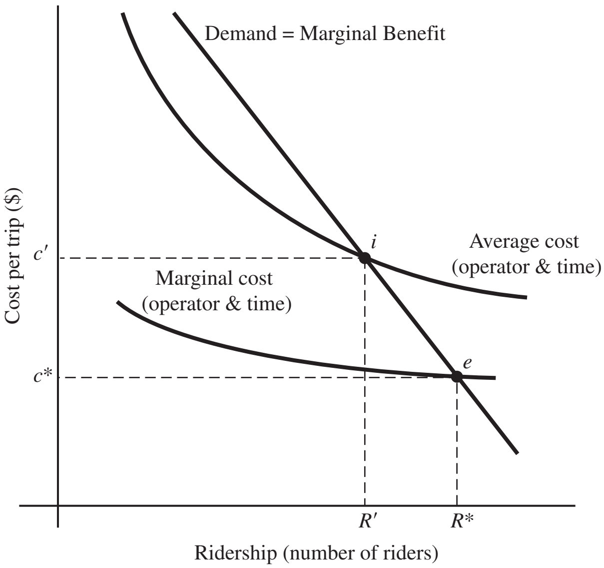
Although we have explained Mohring economies in terms of reduced time cost for riders, there is an alternative cost savings from increased ridership. Consider a route where buses initially have excess capacity—a relatively low load factor, for example, half-full buses. In this case, a doubling of ridership doesn't change the total operator cost (the operator runs its buses with the same frequency), but decreases the average operator cost because there are more riders. In real transit systems, an increase in ridership is likely to decrease both the average time cost of riders and the average operator cost.

Optimum Ridership and Price

Figure 11–2 shows the average-cost curve (including both operator and time costs) of an urban transit system, along with the marginal-cost curve. When an average-cost curve is negatively sloped, the marginal-cost curve lies below the average-cost curve. This is a matter of simple arithmetic: if average cost is decreasing as quantity increases, the marginal cost must be less than the average cost. In other words, if the marginal cost is less than the average cost, the marginal cost pulls down the average cost. In the case of transit, marginal cost is less than average cost for two reasons.

1. The average cost includes some capital costs that are independent of volume (and thus not part of marginal cost).
2. Because of Mohring economies, the marginal time cost (the change in aggregate time cost from one additional rider) is less than the average time cost. An additional rider incurs the average time cost but also triggers an increase in service frequency that decreases the time cost of other riders. In other words, an additional rider speeds up travel for the transit system.

FIGURE 11–2 Rationale for Transit Subsidy



The average-cost curve for transit is negatively sloped, and marginal cost is less than average cost. To balance the transit system budget, the fare equals the average operator cost, generating trip cost (rider time cost plus fare) c' and ridership R'. The socially efficient outcome, where marginal benefit = marginal cost, is shown by point *e*. To reach the efficient point, the transit operator cuts the fare to decrease the trip cost to c^*, generating a transit deficit.

Note that the Mohring economies in transit ridership are the opposite of the congestion externalities from driving: an additional driver slows down other drivers, while an additional transit rider speeds up other riders.

We can use Figure 11–2 to show the budget-balancing ridership. The linear curve is the demand curve for transit. Like other demand curves, it is also a marginal-benefit curve, showing the marginal benefit (willingness to pay) for different volumes. From the perspective of a transit rider, the full cost of transit equals the rider time cost plus the monetary cost (the transit fare). Suppose the operator sets a fare equal to the average operator cost. The average-cost curve in Figure 11–2 includes both the rider time cost and the operator cost (the fare), so it shows, for each volume, the trip cost borne by a rider. The equilibrium is shown by the intersection of the average cost curve and the demand curve. At point i, ridership is R', the trip cost is c', and the operator's budget is balanced.

As explained in "Tools of Microeconomics," the appendix at the end of this book, we can use the marginal principle to determine the socially efficient ridership. In Figure 11–2, the marginal benefit (shown by the demand curve) equals the marginal social cost at point e, so the socially efficient trip cost is c^* and the socially efficient ridership is R^*. To go from point i to point e (to increase ridership from R' to R^*), the transit authority cuts the fare below the budget-balancing fare. The lesson from Figure 11–2 is that a transit deficit is socially efficient.

Transit Subsidies

Taxpayers provide large subsidies for mass transit. In 2002, the total value of the subsidies from federal, state, and local governments was $23 billion, with $14 billion in subsidies for operating costs and another $9 billion in capital subsidies. The federal government provides about 30 percent of the subsidies, while state governments provide 36 percent and local governments provide the remaining 34 percent.

Table 11–5 shows passenger fare subsidies for the 20 largest U.S. transit operators. The numbers show the subsidy as a percent of transit operating cost. For example, the average subsidy of 54 percent (shown at the bottom of the fourth column of numbers) means that a transit authority with an average operating cost of $1 per rider receives a subsidy of $0.54 per rider. The average subsidy for rail is 44 percent, and the subsidy ranges from 29 percent for New York City Transit to 85 percent in Miami. Bus subsidies are generally higher, with an average of 69 percent and a range of 57 percent for New Jersey Transit to 89 percent for Portland.

We've seen that subsidies for mass transit are justified on the grounds of efficiency because of scale economies and Mohring economies. To illustrate the effects of subsidies, consider a bus system that is not subject to conventional scale economies: the average operator cost is constant. Although this assumption is implausible for a rail system because of the inherent indivisibilities in providing rail service (the rail and the right of way), it is plausible for a bus system. A bus operator can vary the number of buses with bus ridership, adding buses when ridership increases and shedding buses when ridership decreases.

TABLE 11–5 Passenger Fare Subsidies for the 20 Largest US Transit Operators in 2003

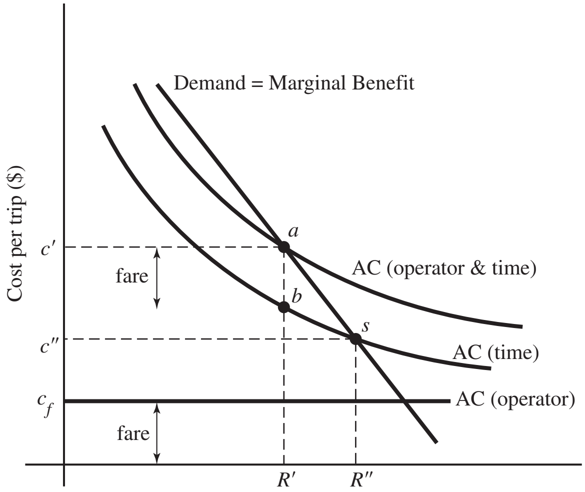
	Fare subsidy as percent of operating cost		
	Rail	Bus	Combined
MTA New York City Transit	29	59	41
New Jersey Transit Corporation	50	57	53
MTA Long Island Rail Road/Bus	53	61	53
Metro-North Commuter Railroad Co. (New York)	40	na	40
Washington Metropolitan Area Transportation Authority (Washington DC)	40	75	55
Massachusetts Bay Transportation Authority (Boston)	57	79	64
Los Angeles County Metropolitan Transportation	78	72	73
Chicago Transit Authority	59	64	62
Northeast Illinois Regional Commuter Railroad Corp (Chicago)	56	na	56
Southeastern Pennsylvania Transportation Authority (Philadelphia)	50	62	57
San Francisco Bay Area Rapid Transit District	42	na	42
Metropolitan Atlanta Rapid Transit Authority	67	71	69
Maryland Transit Administration (Baltimore)	72	74	73
King County Metro Transit (Seattle)	na	82	82
Metropolitan TA of Harris County (Houston)	na	82	82
Tri-County Metropolitan Transportation (Portland OR)	35	89	76
Miami-Dade Transit	85	75	77
Dallas Area Rapid Transit	89	87	88
Denver Regional Transportation District	63	80	79
Port Authority of Allegheny County (Pittsburgh)	81	73	75
Average (weighted by passenger miles)	**44**	**69**	**54**

Source: Adapted from Ian Parry and Kenneth Small, "Should Urban Transit Subsidies Be Reduced?" *American Economic Review* 99 (2009), p. 700–724.

Figure 11–3 shows the average operator cost of the bus system and the average time cost of riders. The average-cost curve for the operator is horizontal, reflecting the absence of conventional scale economies. As usual, the average-cost curve for travel time is negatively sloped, reflecting Mohring economies. If the transit operator sets a fare equal to the average operator cost (c_f) the trip cost is the sum of the rider time cost (shown by point *b*) and the bus fare, or c'. At this trip cost, ridership is R' riders (shown by point *a*), and fare revenue equals operator cost.

To illustrate the effects of a subsidy in the simplest way, consider a 100 percent subsidy: the bus fare decreases to zero. In that case, the rider's trip cost is simply the rider's time cost, and ridership increases from R' to R'' (shown by point *s*). The average trip cost decreases from c' to c''. The decrease in the trip cost exceeds the fare subsidy (the decrease in the monetary cost) because of Mohring economies. The subsidy increases ridership, which increases the frequency of bus service, which in turn decreases the time cost of travel and encourages more people to ride the bus. The result is a relatively large increase in ridership because the normal price effect (an increase in ridership from a decrease in monetary cost) is reinforced by a Mohring effect (an increase in ridership from a decrease in time cost).

FIGURE 11–3 Effects of a Bus Subsidy



A transit fare equal to the average operator cost balances the transit budget and generates trip cost c' and ridership R'. A 100 percent subsidy decreases the trip cost to c'' and increases ridership to R''. The decrease in the trip cost exceeds the fare subsidy because of Mohring economies (a move from point b to point s).

Parry and Small (2009) show that scale economies and Mohring economies are substantial. For Los Angeles bus service, the scale and Mohring economies are large enough to justify an operating subsidy of about 47 percent for peak travel and 81 percent for off-peak travel. For Washington DC rail service, the appropriate subsidies are 48 percent for peak travel and 84 percent for off-peak travel. For London rail service, the appropriate subsidies are 28 percent for peak travel and 60 percent for off-peak travel.

Rationale for Transit Subsidies: Congestion, Environment, and Accidents

A second rationale for transit subsidies is that automobile use generates externalities. As we saw in the previous chapter, drivers pay less than the full social cost of travel, so driving is underpriced relative to mass transit, a substitute good. Underpricing occurs because of congestion externalities, environmental externalities, and collision externalities. The underpricing of driving distorts modal choice in favor of driving. A transit subsidy decreases the relative cost of transit and increases transit ridership, in part by getting travelers to switch from driving to riding transit. In other words, a transit subsidy increases the relative cost of driving, decreasing the equilibrium number of drivers.

Parry and Small (2009) explore the magnitude of the benefits of transit subsidies from congestion relief, a cleaner environment, and reduced accidents. For Los Angeles, the external benefits are large enough to justify an operating subsidy of about 27 percent for peak bus service and 45 percent for peak rail service. For Washington DC, the appropriate subsidies are 9 percent for peak bus service and 37 percent for peak rail service. For London, the subsidies are 36 percent for peak bus service and 50 percent for peak rail service.

Incentive Effects of Transit Subsidies

Although subsidies for mass transit are justified on efficiency grounds, the current structure of transit subsidies does not provide incentives for cost minimization by the monopoly providers of transit service. There is evidence that the subsidies lead to relatively high operator costs because of excessive labor compensation, misallocation of labor (high-skilled workers performing low-skilled tasks), and inefficient mixes of capital and labor. As explained by Parry and Small (2009), a solution to the incentive problem is to switch from operator-based subsidies to user-side subsidies, under which the government pays a transit operator a fixed subsidy per passenger trip or per passenger mile. For example, if the subsidy is $0.10 per passenger mile, a transit operator that provides 200 million passenger miles per year would receive an annual subsidy of $20 million.

DESIGNING A TRANSIT SYSTEM

In this part of the chapter we take the perspective of a transportation planner and explore the features and costs of alternative urban transportation systems. We compare the costs of three types of mass transit (heavy rail, light rail, buses) to the cost of an automobile/highway system. To illustrate the cost of heavy rail, we use data from San Francisco's BART (Bay Area Rapid Transit). We can divide the cost of an urban transportation system into three types:

1. **Capital cost:** the cost of laying rails and buying the transit vehicles.
2. **Operating cost:** the cost of labor, fuel, and maintenance of vehicles, roads, and rails.
3. **Time cost:** the disutility (opportunity cost) of travel time.

Design Features for Mass Transit

The designer of an urban transportation system makes several types of decisions. The following trade-offs occur for different types of transit systems.

1. **Mainline versus integrated system.** A rail system is a mainline system: it relies on other modes to collect its riders from residential neighborhoods and

distribute riders to workplaces and other destinations: Riders walk, drive, or ride a bus to and from rail stations. In contrast, an integrated bus system keeps riders in the same vehicle for the entire trip.

2. **Distance between bus stops and rail stations.** The longer the distance between rail stations, the fewer the stops en route, so the faster the line-haul portion of a trip and the lower the in-vehicle time cost. On the other hand, travelers have longer distances to walk or ride other vehicles, generating relatively high access time costs. The disutility of access time is roughly 1.6 times the disutility of in-vehicle time, so the trade-off in time cost is not one-for-one.
3. **Frequency of service.** The more frequent the service, the higher the capital cost (more vehicles) and the higher the operating cost (more labor, fuel, and other inputs). On the other hand, more frequent service decreases riders' access time cost.

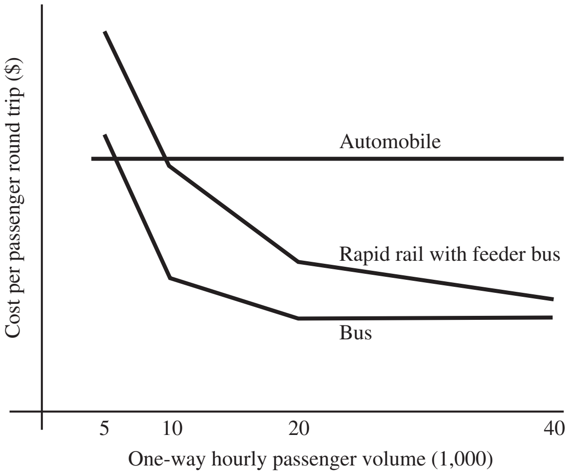
Choosing a System: Automobile versus Rail versus Bus

The pioneering study of system choice is by Meyer, Kain, and Wohl (1965). They consider several systems, including an auto-based highway system, an integrated bus system, and a rapid-rail system like San Francisco's BART. They incorporate all the costs of the alternative systems, including the capital cost (laying down tracks, building highways, and buying vehicles), operating costs, and the disutility of travel time.

Small and Verhoef (2007) adapted some of the key results of the pioneer study. Figure 11–4 shows the average costs of commuting along a corridor that connects residential areas to a high-density employment area. The commuting trip is a 10-mile line-haul trip and a 2-mile downtown distribution route.

1. **Cost of the automobile system.** The cost of the automobile system is the sum of the cost of acquiring and operating automobiles, the disutility of travel time, the cost of building the optimum road system, and external costs of automobile travel (congestion and pollution). The average cost curve is horizontal because the authors assume (*a*) the average operating cost and environmental cost do not depend on traffic volume and (*b*) in the long run, the highway is widened to accommodate any increase in traffic without any reduction of travel speeds.
2. **Cost of the bus system.** The cost of the bus system is the sum of the cost of acquiring and operating buses, the disutility of travel time, the cost of building roads suitable for buses, and the external costs of bus travel (pollution). The average-cost curve for the bus system is negatively sloped because of conventional scale economies (spreading fixed cost) and Mohring economies.
3. **Cost of the rapid-rail system.** The cost of the rapid-rail system is the sum of the cost of building and maintaining the rails and the exclusive right-of-way, the cost of acquiring and operating rail vehicles and feeder buses, the disutility of travel time, and the external costs of bus travel. The average-cost curve for the

FIGURE 11–4 Average Cost of Alternative Transportation Systems



The average cost of an automobile system is constant because roads are expanded to accommodate increased traffic volume. The transit curves are negatively sloped because of scale economies (spreading fixed cost over more riders) and Mohring economies. A bus system is less costly than an automobile system for volume exceeding 6,000 passengers, and less costly than a rail system for all volumes shown.

rail system is negatively sloped because of scale economies (spreading fixed cost) and Mohring economies (more frequent service). For low volumes, the average cost of rail exceeds the average cost of the bus system because of the higher cost of indivisible inputs for rail.

Figure 11–4 represents transportation costs for a city with a residential density that is typical of U.S. cities. The cost data in Figure 11–4 support two principal conclusions about the market roles of different transportation systems. First, for relatively low traffic volumes, the automobile is the least costly. In the example presented in the figure, the bus becomes competitive with the automobile at a volume of about 6,000 passengers per hour, and rail becomes competitive at a volume of about 10,000 passengers per hour. Second, the bus is less costly than a heavy-rail system for all volumes studied.

The authors also estimate the system costs for cities with relatively high residential densities. In those cases, rapid rail is less costly than a bus system for a volume of 30,000 passengers or more. In other words, rail transit is more efficient in metropolitan areas such as Chicago and New York, where corridor volumes are large enough to support heavy-rail systems. Recent experience with new heavy-rail systems in other metropolitan areas confirms this conclusion: Ridership on the new systems (in Washington DC, Atlanta, Miami, and Baltimore) has fallen well short of levels required to make heavy rail less costly than bus systems.

A Closer Look at Light Rail

Although both the bus and the rail systems are subject to scale economies that decrease average cost as volume increases, they differ in their fixed system costs. As shown in Table 11–4, in terms of cost per route mile, heavy rail is roughly three times as costly as light rail. The cost per route mile is much lower for exclusive busways and shared carpool lanes. In terms of the cost per daily trip, the two rail options have roughly the same cost, and the bus options have much lower costs. Although light rail has a lower construction cost per route mile, it also has lower ridership, negating its capital cost advantage relative to heavy rail.

In recent years, many medium-size cities have built light-rail transit systems. Light rail is the modern version of the trolley and streetcar systems that were built in the late 1800s and early 1900s. The first modern light-rail system opened in Edmonton in 1978, and since then light-rail systems have been build in Baltimore, Buffalo, Dallas, Denver, Los Angeles, Pittsburgh, Portland, Sacramento, San Diego, San Jose, and St. Louis. A key design feature for light rail is how to link the rail line with the bus system. In most cases, feeder buses are necessary complements to the rail line, imposing costs on travelers in the form of extra waiting time and connection hassle.

Richmond (1998) examines light-rail systems in 11 cities and compares their performance to bus systems. The general conclusion is that light rail is inferior to buses.

1. Light rail has higher capital costs. For example, the capital cost of the Long Beach light-rail system was $881 million, compared to the $168 million in capital costs that would have been required for an equivalent bus system.
2. Light rail has higher operating costs. Most data reported on the operating costs of light rail omit the costs of the feeder buses that bring riders to light-rail stations. Ignoring these costs, the average operating cost for light rail is somewhat higher or perhaps slightly lower than that for equivalent bus lines. For example, the cost per passenger mile for Portland's MAX is $0.38, compared to $0.32 for one equivalent bus line, and $0.39 for another. Once the cost of feeder buses is included, light rail is more costly than equivalent bus lines.
3. Light rail diverts passengers from buses. For the Blue Line in Los Angeles, 63 percent of riders were previously bus riders. In Portland, about 55 percent of MAX's riders switched from buses to light rail.

A light-rail system requires feeder buses to collect passengers, and this is expensive for transit operators and bothersome for potential riders. In the modern metropolitan area with dispersed employment and retail activities, it is difficult to attract enough riders to make light rail less costly than a well-designed bus system. In many cases, busways and other HOV systems would be less expensive—and more effective—in increasing transit ridership (Richmond, 1998). In other cases, simple and inexpensive changes in regular bus service (adding buses, changing routes or schedules, or decreasing fares) may be more efficient than big projects.

TABLE 11–6 Minimum Densities to Support Mass Transit

	Built-up Density: People Per Hectare of Land in Urban Use	Residential Density: People Per Hectare of Residential Land
One bus per hour	21	30
Two buses per hour	31	44
Light rail	37	53
Heavy rail	50	71

Notes: Hectare = 2.5 acres; intermediate service = 40 buses per day; high service = 120 buses per day.

Source: J. Holtzclaw, *Using Residential Patterns and Transit to Decrease Auto Dependence and Costs.* Washington DC: Natural Resources Defense Council, June 1994.

The Role of Density

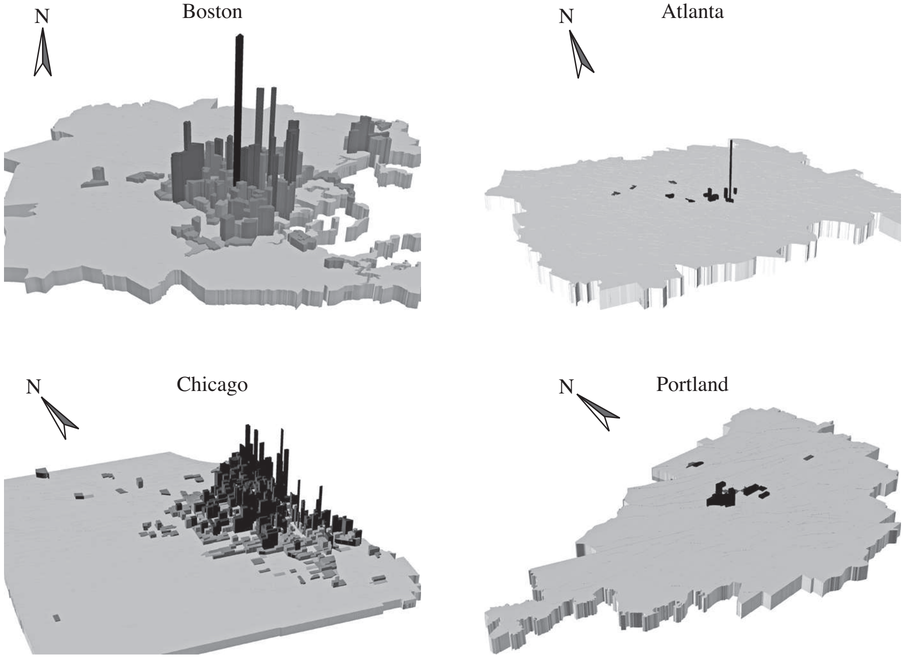
In an earlier chapter on land-use patterns, we briefly discussed the role of population density in supporting mass transit. Mass transit is feasible only if density around bus stops or train stations is high enough to attract enough riders to exploit conventional scale economies and Mohring economies. For most people, the maximum walking time to a transit stop is about 10 minutes, so a transit stop can serve households within an 800-meter radius.

Table 11–6 shows the minimum densities required to support various types of transit. The built-up density of a metropolitan area is defined as total population divided by the amount of land in urban use, including residential areas, industrial districts, commercial areas, roads, schools, and city parks. In contrast, residential density is defined as population divided by the residential area. The minimum density for hourly bus service is 21 people per built-up hectare and 30 people per residential hectare. As we move down the table, the minimum densities increase as we move to more frequent bus service, light rail, and heavy rail.

As we saw in an earlier chapter, few U.S. metropolitan areas meet these minimum density thresholds. New York (40 people per hectare) meets the threshold for light rail and bus service, and Honolulu (31) meets the threshold for intermediate bus service (two buses per hour). The 10 most dense metropolitan areas, with densities of 18 or more people per hectare, come close to the low-level threshold of bus service. Of course, these are average densities, and parts of some of these metropolitan areas have population densities high enough to support transit. For example, the density in New York City is 80 people per hectare. In contrast, European cities such as Barcelona (with a density of 171 people per hectare) and Paris (88 per hectare) have high enough densities to support the highest level of transit service.

Map 11–1 shows how the population densities in four metropolitan areas compare to the threshold density required for viable mass transit. Each census tract is extruded to the density required to support light rail (37 per hectare or 9,472 per square mile), generating a plateau at a height of 37. Each tract is then extruded again to show the actual population density. If the actual density exceeds the threshold density, the tract is extruded above the transit plateau. In Panels A and B, in Boston and Chicago, the actual population density exceeds the light-rail threshold for light rail for a large part of the city. As shown in Panels C and D, Atlanta and Portland have much lower population density, and a small fraction of their residents live in tracts that rise above the transit plateau.

MAP 11–1 Population Density and Viable Mass Transit



The plateaus are at a height of 37, indicating the minimum density required to support light rail. Each tract is extruded to show the actual population density, indicating the tracts with density above the threshold density.

DEREGULATION: CONTRACTING AND PARATRANSIT

In most cities in the United States, urban transportation is heavily regulated. Firms are prohibited from providing services that compete with the local transit operator, and the result is a transit monopoly. In addition to outlawing competition in operating fleets of transit vehicles, local regulations prohibit taxis from serving as a common carrier: a taxi cannot pick up additional passengers, either along a route chosen by a customer or a route chosen by the driver.

One reason for restricting entry into urban transit is to prevent cream-skimming by private firms. A public transit operator covers a large number of routes, even the ones with relatively low ridership and revenue. The revenue from more lucrative routes cover some of the losses on less lucrative ones. If firms were free to enter the transit market, they would take business away from public transit on profitable routes (skimming the cream) and hamper the city's cross-subsidization efforts. The result could be higher fares or reduced service on the low-volume routes. The obvious response to this problem is to allow the private sector to compete on profitable routes, and to continue subsidizing low-volume routes, making the subsidies explicit and transparent.

Contracting for Transit Services

One option for deregulation is to write contracts to have private firms provide specific transit services. The local government can specify the service characteristics of the transit system (e.g., headways, travel times, location of bus stops, fares) and then accept bids from private firms for the transit service. The Federal Transit Administration estimates that this sort of contracting would generate savings in operating costs between 25 and 30 percent. When the city of Tidewater, Virginia, contracted for bus service to low-density areas, a private transit company provided the same transit service at a cost per passenger of about $3 lower. Other cities using transit contracts have experienced similar cost savings. Cities also use contracting for subsidized dial-a-ride services.

Private firms provide transit services for a lower cost for three reasons. First, they pay lower wages. When BART accepted bids for feeder-bus service to serve transit stations, the average hourly wage among bidders was $2 below the wage of public transit workers. Second, private companies have more flexible work rules. They use split shifts and part-time workers, so they don't pay idle workers during the off-peak periods. Third, private companies use minibuses on low-density routes, saving on operating and capital costs.

Paratransit

A second option for deregulation is to allow private firms to compete for transit consumers. Deregulation would change the mix of transportation services available in most cities. The current system has two extremes—solo-rider taxis and large transit vehicles (large buses and rail cars)—and would be replaced by a system that provides travelers with a wider variety of services in a variety of vehicles. The term paratransit was adopted in the 1970s to describe a wide variety of services that fall between the private automobile and the conventional bus.

- **Shared-ride taxis (three to four passengers).** During World War II, shared-ride taxis thrived in Washington DC. Cab drivers displayed destination signs, allowing people along the route to hail cabs going their way. Currently a number of cities allow taxi sharing.
- **Jitneys (6 to 15 passengers).** Compared to a large bus, a jitney has a lower cost per passenger and can provide more frequent service. Jitneys also have the flexibility to change routes, pickup points, and schedules. Several cities allow jitney service, resulting in lower fares and profits for the provider. Atlantic City restricts the number of jitneys, and the market value of a jitney license was $160,000 in 1995.
- **Subscription commuter vans and buses (10 to 60 passengers).** Riders pay in advance for commuter bus service. In San Francisco, Golden Gate Transit established 22 commuter routes between suburban communities and the downtown financial district. In New York City, private bus lines carry about 60,000 workers per day to jobs in Manhattan. The private operators have lower costs than public buses in part because they pay lower wages.

Paratransit can fill the gap in the urban transportation system between solo-rider taxis and large public buses. In contrast to subsidized public buses, paratransit operations actually earn profits. In summarizing the prospects for paratransit, Cervero (1996) notes that:

> Given the fiscal cutbacks facing America's public transit industry today, the expansion of more entrepreneurial, commercial transportation services seems unavoidable. While critics charge that the poor will suffer as a result, other remedies—like user-side subsidies—are available for redressing inequities. Moreover the history of commercial paratransit is certainly not one of ignoring poor neighborhoods. For jitneys and neighborhood car services, low-income areas have traditionally been their market base.

The British Experience with Deregulation

In Britain, the transit industry was deregulated under the British Transport Act of 1985. The act relaxed controls on entry into the transit industry, reorganized most public transit operators as for-profit organizations, and introduced competitive bidding for some transit services. In addition, transit subsidies were cut significantly, leading to higher fares and service cuts. Deregulation led to an increased use of minivans, lower production costs as wages fell and work rules were relaxed, and the elimination of service on some low-volume routes.

Gomez-Ibanez and Meyer (1990) discuss three lessons from the British deregulation experience. First, it is possible to have both competition in the local bus industry and subsidies for unprofitable services. Cities can use competitive bidding to pick low-cost private providers. Second, deregulation causes service innovation and lower costs. Third, because most of the benefits from deregulation come from competition among transit firms, the public sector must develop policies to ensure competition.

TRANSIT AND LAND-USE PATTERNS

As we saw in Chapter 7, urban transportation technology affects urban form. The monocentric city of the early 20th century resulted from the combination of (*a*) streetcars for the transportation of workers, (*b*) primitive technology for the transportation of goods (the horse-drawn wagon) and (*c*) face-to-face contact for the transmission of information. The development of the truck and the interstate highway system allowed the decentralization of manufacturing jobs, and the development of electronic transmission of information allowed the decentralization of information (office) jobs. The automobile freed workers from their dependence on walking and streetcars, causing suburbanization and lower residential density.

Given the effects of changes in transportation and communication technology on urban form, a natural question concerns the influence of urban transit systems on land-use patterns. Do systems like BART and MARTA increase the density of employment and residence near transit stations?

TABLE 11–7 Transit Accessibility of Additional Residents and Jobs in Atlanta

	Residence	Jobs
Percent within 800 meters of MARTA station	2	1
Percent within 800 meters of bus line	13	22
Percent inaccessible to mass transit	85	77

Source: Alain Bertaud. "Clearing the Air in Atlanta: Transit and Smart Growth or Conventional Economics?" *Journal of Urban Economics* 54 (2003), pp. 379–400.

The experience of Atlanta shows that for mass transit, supply does not create its own demand. In the last two decades, the city has invested heavily in mass transit, building a rapid-rail system with 74 kilometers of tracks. Although the transport network runs smoothly and safely, it has failed to attract jobs and people. Between 1990 and 1999, nearly 700,000 people were added to the Atlanta metropolitan area, but only a small fraction of the new residents lived or worked at locations accessible to the transit system. As shown in the first column of Table 11–7, only 2 percent of the additional residents chose locations that are accessible to a MARTA station, and only 13 percent chose locations accessible to a bus line. As shown in the second column of Table 11–7, only 1 percent of new jobs were accessible to MARTA and 22 percent were accessible to a bus line. The fact that so few people chose locations accessible to mass transit suggests a weak connection between transit design and urban form, at least in Atlanta.

One of the objectives of San Francisco's BART was to increase employment opportunities in the areas near transit stations. Studies of BART suggest that the transit system had a moderately positive effect on employment near stations in downtown San Francisco, but not much of an effect elsewhere (Cervero and Landis, 1995). This is consistent with studies of other rail systems, which support two conclusions (Altshuler, 1979):

1. In a growing economy, rail transit contributes to the clustering of activities near downtown stations. These clustering effects are usually negligible outside the central business district.
2. Investment in rail transit is sensible only if it is used in concert with more powerful land-use instruments such as zoning and property taxation. If the government uses its zoning and tax policies to generate high-density development, rail transit provides an efficient means of delivering a large number of workers to the dense central area.

SUMMARY

Although transit ridership is relatively low in most U.S. metropolitan areas (5 percent of trips), ridership is relatively high in some large metropolitan areas. A large share of transit riders comes from low-income households. Here are the main points of the chapter.

1. The elasticity of demand for transit with respect to price is –0.40, and the elasticity of demand with respect to service is generally larger in absolute value.
2. On average, the marginal disutility of access time (walking and waiting) is about 80 percent of the market wage, while the marginal disutility of in-vehicle time is about 50 percent of the wage.
3. In most U.S. metropolitan areas, a bus system is more efficient than a rail system (either heavy rail or light rail). The exceptions are high-density metropolitan areas such as New York and Chicago.
4. Subsidies for mass transit are justified on efficiency grounds because (*a*) transit is subject to scale economies from indivisible inputs and Mohring economies from lower access costs, and (*b*) automobiles generate external costs from congestion, environmental damage, and accidents.
5. A switch from general subsidies to user-side subsidies would give transit operators a greater incentive to control cost.
6. Transit systems have modest effects on land-use patterns.

APPLYING THE CONCEPTS

For exercises that have blanks (_____), fill each blank with a single word or number. For exercises with ellipses (…), complete the statement with as many words as necessary. For exercises with words in square brackets ([increase, decrease]), circle one of the words.

1. **Distance between Bus Stops**
   Consider a city that decreases the distances between bus stops, which decreases the walking time of bus riders by 20 percent, increases the in-vehicle (line-haul) time by 10 percent, and increases operating cost by 10 percent. On the typical bus line, the initial ridership is 1,000 riders per hour. Suppose the elasticity of transit ridership with respect to line-haul time is –0.39 and the elasticity of transit ridership with respect to access time is –0.71.
   a. Ignoring the change in operating costs, ridership will [increase, decrease] by _____ percent (_____ riders), computed as …
   b. If the transit operator passes on the higher operating cost in the form of an increased fare, the combination of shorter distances between stops and a higher fare will [increase, decrease] ridership by _____ percent (_____ riders), computed as …
2. **Transit Improvement and Auto Volume**
   Consider Table 11–1 and the elasticities of demand for transit. Assume that the total number of commuters is fixed. Suppose we decrease the line-haul time of public transportation (otherwise known as transit) by 20 percent and the elasticity of transit ridership with respect to line-haul time is −0.39.
   a. The number of transit commuters will [increase, decrease] by _____ riders, computed as …
   b. The number of nontransit commuters will [increase, decrease] by _____ percent, computed as …

- **c.** If the change in nontransit commuting is distributed proportionately over the five nontransit options, the number of car, truck, and van commuters will [increase, decrease] by _____.

3. **Disney Transit**
   Consider the chapter-opening quote. Suppose your city's transportation manager hires the team of workers that designed the most popular Disneyland rides and asks them to redesign the city's transit system. The city's objective is to increase ridership while eliminating the transit subsidy. In other words, transit fares will cover the full cost of the system.
   - **a.** Predict the features of the Disney-inspired transit system.
   - **b.** Would you expect the new system to meet the city's objectives?
4. **Average Cost for Light Rail**
   Use a graph like Figure 11–4 to show the long-run average cost of a light-rail system. Draw the light-rail curve along with the curves for BART and a bus system. Explain your placement of the light-rail curve.
5. **Disutility of Bus versus Car**
   In drawing the cost curves in Figure 11–4, the authors assumed that the disutility of time spent riding a bus is the same as the disutility of time driving a car. Based on your own preferences, is this a realistic assumption? Draw a new set of cost curves consistent with your preferences. The range over which buses are more efficient than cars will be [wider, narrower].
6. **Response to Diamond Lanes**
   Zirconium City just converted one of the four lanes on its freeways to a diamond or HOV (high-occupancy vehicle) lane, used primarily by buses. The conversion shortened the line-haul time of buses by 10 minutes and increased the line-haul time of autos by three minutes. As a result, Buster switched from driving to taking the bus, but Otto continued to drive.
   - **a.** Buster switched to the bus because his disutility of _____ time is large relative to his disutility of _____ time.
   - **b.** Otto didn't switch because his disutility of _____ time is large relative to his disutility of _____ time.
7. **Free Transit?**
   Consider a city with an initial fare for public transit of $1.50 and an initial ridership of 120,000. Suppose the city cuts its transit fare to zero.
   - **a.** The total ridership of public transit will [increase, decrease] by _____ riders, computed as …
   - **b.** Illustrate with a graph.
8. **Segway and the Transit Deficit**
   Suppose a large city with exclusive busways gives each of its citizens a Segway, a self-balancing personal transportation vehicle. As the latest marvel of miniaturization, the transporter can be collapsed into a package the size of a briefcase and carried easily by a commuter onto a bus. The Segway allows travel at four times the speed of walking, and with a force-field accessory, can be used for comfortable travel even in nasty weather. The bus

fare equals the average system cost. Use a graph like Figure 11–2 to show the effects of the Segway on the equilibrium cost per trip (full price) and equilibrium volume.

9. **Internalizing Automobile Externalities**
   Consider a city where the fixed cost of a transit system is $140.00 per hour. The long-run marginal cost is constant at $1.00 per rider. The demand curve is linear, with a vertical intercept of $11.00 and a slope of $0.10 per rider.
   a. Use a graph to show the demand curve, the marginal-cost curve, and the average-cost curve.
   b. Under marginal-cost pricing, the price is _____, ridership is _____ riders, and the deficit per rider is _____.
   c. Suppose the city internalizes the externalities from automobiles, and the willingness to pay for transit increases by $4.00 at each ridership level. Under marginal-cost pricing, the price is $_____, ridership is _____ riders, and the deficit per rider is $_____.
10. **A Planner's Bomb in Atlanta**
   Consider the effects of dropping a planner's bomb on the Atlanta metropolitan area. A planner's bomb doesn't hurt any people, but destroys everything except the MARTA infrastructure (tracks, vehicles, and stations). Most important, it destroys all buildings, so the metropolitan area must be completely rebuilt. To simplify the geography, imagine that the MARTA tracks are radial, with two lines intersecting at the center.
   a. The planner's objective is to generate transit ridership equal to the levels observed in Barcelona while accommodating Atlanta's prebomb population. Describe the features of the plan for rebuilding the city.
   b. If the plan is implemented, would you expect the population of the metropolitan area to change?

REFERENCES AND ADDITIONAL READING

1. Altshuler, Alan A. *The Urban Transportation System*. Cambridge, MA: Joint Center for Urban Studies of MIT and Harvard, 1979.
2. American Public Transit Association. *Transit Fact Book 2010*. Washington DC, 2010.
3. Bertaud, Alain. "Clearing the Air in Atlanta: Transit and Smart Growth or Conventional Economics?" Journal of Urban Economics 54 (2003), pp. 379–400.
4. Cervero, Robert. "Commercial Paratransit in the United States: Service Options, Markets, and Performance." Working Paper No. 299, University of California Transportation Center, 1996.
5. Cervero, Robert, and John Landis. "The Transportation-Land Use Connection Still Matters." *Access* 7 (Fall 1995), pp. 2–10.
6. Gomez-Ibanez, Jose A. "A Dark Side to Light Rail?" *Journal of the American Planning Association* 51 (Summer 1985), pp. 337–51.

7. Gomez-Ibanez, Jose A., and John R. Meyer. "Privatizing and Deregulating Local Public Services: Lessons from Britain's Buses." *Journal of the American Planning Association* 56 (Winter 1990), pp. 9–21.
8. Holtzclaw. J. Using Residential Patterns and Transit to Decrease Auto Dependence and Costs. Natural Resources Defense Council, June 1994.
9. Meyer, John R., John F. Kain, and Martin Wohl. *The Urban Transportation Problem*. Cambridge MA: Harvard University Press, 1965.
10. Parry, Ian, and Kenneth Small. "Should Urban Transit Subsidies Be Reduced?" *American Economic Review* 99 (2009), p. 700–24.
11. Richmond, Jonathan E. "*New Rail Transit Investments—A Review*." Taubman Center for State and Local Government, 1998.
12. Small, Kenneth A. *Urban Transportation Economics*. Philadelphia: Harwood, 1992.
13. Small, Kenneth, and Erik Verhoef. *The Economics of Urban Transportation.* New York: Routledge, 2007.

PART FOUR

Education and Crime in Cities

This part of the book explores the two largest spending programs of local governments, education and crime control. Local governments in the United States spend about 48 percent of their budgets on education, and about 8 percent of their budgets on the criminal justice system. The quality of education varies significantly across space, both within and across cities. As a result, education plays an important role in households' location decisions. Similarly, spatial variation in crime rates affects residential location choice. In Chapter 12 we introduce the concept of the education production function and explore the contributions of the key inputs to the education process, including the home environment, peer groups, and teachers. In Chapter 13 we introduce the model of the rational criminal and explore the implications of rational criminal behavior for public policy.

CHAPTER 12

Education

Human history becomes more and more a race between education and catastrophe.

—H.G. Wells

This chapter explores the economics of education. Roughly half of the total expenditures of local governments in the United States, go to local school. The spending on K–12 education (kindergarten through high school) is roughly seven times the spending on police, the second largest expenditure category for local governments. As we've seen in earlier chapters, educational achievement varies significantly across space, both within and across metropolitan areas. As a result, the quality of local schools is an important factor in the location decisions of households and firms. And since a better educated workforce is more productive and innovative, local education influences urban economic growth.

The economic analysis of education is based on the education production function, which represents the relationship between the inputs and outputs of the education process. The inputs include variables controlled by schools (teachers, class size, curriculum) as well as variables beyond the control of schools, including the home environment. Until recently, education output has been defined in terms of scores on cognitive tests. Recent advances in theory and empirical work allow us to measure educational output as the change in lifetime earnings attributable to education. This approach provides a framework for benefit–cost analysis of changes in the educational input mix. For example, a decrease in class size (more teacher time per student) increases student achievement, but does the additional benefit in terms of increased earnings exceed the extra cost?

SPENDING AND EDUCATIONAL ACHIEVEMENT

In the United States in 2008, the average expenditure per K–12 student was $10,981. As shown in Table 12–1, spending per pupil varies significantly across the nation, from a low of $6,841 in Utah to a high of $20,807 in the District of Columbia. Between 1960 and 2010, real spending per pupil more than quadrupled. Over the last

TABLE 12–1 Per Pupil Spending in Selected States, 2007

State	Spending per Pupil ($)
Utah	6,841
Oklahoma	8,270
Mississippi	8,448
Tennessee	8,459
Arizona	8,630
Connecticut	15,063
New Jersey	18,174
New York	18,423
District of Columbia	20,807

Source: U.S. Department of Education (2010), Table 194.

TABLE 12–2 International Student Test Scores

Country	Average Score Mathematics	Average Score Science
Chinese Taipei	598	561
Korea, Republic	597	553
Singapore	593	567
Hong Kong SAR	572	530
Japan	570	554
Hungary	517	539
England	513	542
Russian Federation	512	530
United States	508	520
Lithuania	506	519
Czech Republic	504	539
Slovenia	501	538
Australia	496	515
Sweden	491	511

Source: Adapted from M. O. Martin, I. V. S. Mullis, & P. Foy (2007).

few decades, student performance as measured on cognitive tests has not changed much for mathematics and reading, but has decreased significantly for science (Hanushek, 2003).

Table 12–2 provides an international perspective on educational achievement. The table shows the average test scores (for mathematics and science) of the top ranked countries. On both tests, the worldwide average score among the 60 participating countries is 500. The United States ranked ninth for mathematics achievement, with an average score of 508. For science achievement, the U.S. ranked eleventh, with an average score of 520.

Table 12–3 shows student achievement data for the United States as a whole, for large cities as a group, and for selected cities. The National Assessment of Educational Progress (NAEP) measures achievement levels in reading and mathematics.

TABLE 12–3 Student Achievement in Selected Cities, 2009

Jurisdiction	Percent Below Basic Level	Jurisdiction	Percent Below Basic Level
United States	29	Chicago	49
Large cities	40	Atlanta	54
San Diego	32	Los Angeles	54
Boston	33	Baltimore	57
New York City	40	Cleveland	58
		Detroit	77

Source: U.S. Department of Education (2010), Table 145.

The table shows the percentage of eighth graders who scored below the basic level on the NAEP mathematics exam. Nationwide, 29 percent of students were subbasic, compared to 40 percent for all large cities. Among the large cities listed in the table, student performance was above average in San Diego, Boston, and New York, and below average in Detroit, Cleveland, and Baltimore.

In an earlier chapter we explored the variation in student performance across high schools in a central city. Table 8–4 provides data on student performance and socioeconomic characteristics for the eight high schools in the Portland school district. The table reveals substantial variation in student achievement across the high schools: the percentage of students who meet state standards for mathematics ranges from 44 percent to 80 percent. The percentage of students who are economically disadvantaged ranges from 12 percent to 70 percent.

The United States has a persistent and widespread racial achievement gap, a fact demonstrated by student performance on grade-four NAEP tests. In reading, 43 percent of white students are proficient, compared to 12 percent of black students. In math, 51 percent of white students are proficient, compared to 14 percent of black students. On every subject at every grade level, there are large achievement gaps between white and black students that continue to grow as children move through the school system (Dobbie and Fryer, 2009). Although part of the achievement gap can be explained by differences in socioeconomic background (including the income and education level of parents), two-thirds of the gap remains after controlling for differences in background.

THE EDUCATION PRODUCTION FUNCTION

The purpose of education is to develop cognitive, social, and physical skills. The basic cognitive skills (reading, writing, mathematics, logic) are necessary for employment and participation in a democracy. These skills also increase the enjoyment of leisure activities: they allow people to read books, understand jokes, and compute bowling scores. Schools also develop social skills: they teach children how to exchange ideas and make group decisions. Finally, schools develop physical skills: they teach children how to exercise and play games.

The production function shows the relationship between the inputs and outputs of the educational process. Suppose that achievement is defined as the change in the test score of a particular child over a one-year period. The production function is

$$Achievement = f(H, P, C, E, T)$$

Achievement depends the student's home environment (*H*), the classroom peer group (*P*), the curriculum (*C*), educational equipment (*E*, which includes books, computers, and lab equipment), and the quality of the teacher (*T*).

The Home Environment

The first input in the production function, the home environment (*H*), affects student achievement in three ways. First, parents set the rules of the household, establishing an environment that is either favorable or unfavorable to education. For example, an unfavorable environment is one in which children watch television instead of reading books or doing their homework. Second, parents can motivate their children by encouraging reading and studying, helping with homework, and rewarding success. Third, parents can provide instructional materials such as books and home computers, encouraging independent learning.

The quality of the home environment depends in part on the income and education level of the parents. In general, the children of wealthy and educated parents learn more because they receive more encouragement and assistance at home, and also pick up verbal and quantitative skills in everyday interactions with their parents. In contrast, children from poor and less educated families often learn less in school. In addition, children living in poverty are often malnourished, inhibiting their ability to learn. Based on dozens of empirical studies, it is clear that the home environment has a large effect on educational achievement.

Peer Effects

The second input in the production function is the student's classroom peer group (P). A child learns more if he or she is surrounded by smart and motivated children. Smart peers promote achievement because of cooperation (children learn from one another) and competition (children compete with one another). Motivated peers promote achievement because the teacher can spend less time disciplining and motivating students, and more time teaching. In addition, an unmotivated student provides an undesirable role model for other students.

Some recent studies estimate peer effects in secondary schools (high schools). Chinese students take two admissions tests, one to get into high school and a second to get into college. To measure the quality of a student's peers, we can use the average score on the high-school exam for the student's classmates. To measure achievement of the high-school experience, we can use the student's score on the college exam. Ding and Lehrer (2007) estimate that a 1 percent increase in peer quality increases achievement by 0.088 percent. Sund (2009) measures peer effects in Swedish high schools: if the quality of a student's classmates increases from the median quality (50th percentile) to a quality level at the 84th percentile, the student's achievement increases from the median achievement level to the 54th percentile.

There are important trade-offs associated with peer effects. A key policy issue concerns mixing students of different abilities in a single class, versus sorting students into classes of differing abilities (ability grouping or tracking). Suppose we place Doc (a high-achiever) in one of three classes: (1) a low-achiever class with Dopey; (2) a middle-achiever class with Midge; or (3) a high-achiever class with Heidi. For Doc, the best placement is in the high-achiever class with Heidi. But students in the middle-achiever class or low-achiever class would benefit from Doc's presence. If Doc is placed in the middle-achiever class, Midge would gain at Doc's expense. Similarly if Doc is placed in the low-achiever class, Dopey will gain at Doc's expense.

The evidence concerning the relative magnitudes of the peer effects is mixed. Using data from China, Ding and Lehrer (2007) suggest that middle achievers get the largest benefit from high-achieving peers, while low achievers get the smallest benefit. Using data from the United States, Burke and Sass (2008) suggest that if we put a high achiever in a middle-achiever class, aggregate achievement will increase: the gains of middle achievers exceed the loss of the high achiever. In contrast, placing a high achiever in a low-achiever class decreases aggregate achievement because the relatively small gains of low achievers are dominated by the loss of the high achiever. Using data from Sweden, Sund (2009) comes to a different conclusion: Low achievers have the most to gain from high-achieving peers, so placing a high achiever in a class with low achievers increases aggregate achievement. These conflicting results indicate the continued uncertainty about the magnitude of peer effects across students of different achievement levels.

SCHOOL INPUTS: THE IMPORTANCE OF TEACHERS

There is substantial variation across schools in productivity, as measured by student performance on standardized tests. The differences in productivity are not related to curriculum (C), school equipment (E), or the way a school is managed. Instead, differences in productivity result from differences in teacher productivity. In other words, the most productive schools are the ones with the most effective teachers.

Differences in Teacher Productivity

Teacher productivity can be measured by students' scores on cognitive tests. A study of inner-city schools found that during a single academic year, a student with a high-productivity teacher outperforms a child with a low-productivity teacher by up to one full grade level (Hanushek, 1992). For example, consider a child who starts third grade at grade level 2.0, right on schedule. With a high-quality third-grade teacher, the child will progress to grade level 3.5 by the end of the year, a half grade ahead of schedule. In contrast, with a low-quality teacher, the child will progress to grade level 2.5, a half year behind schedule.

To measure the productivity differences across teachers, we can compare an above-average teacher to an average teacher (Chetty et. al. [2010], Hanushek [2010]). Let's define a "superior" teacher as one whose average student test score places the teacher in the 84th percentile among teachers: in terms of student test scores, the superior teacher does better than 84 percent of teachers. Consider the following thought experiment. Suppose we replace an average teacher (at the 50th percentile) with a superior teacher (84th percentile), and then measure the resulting change in student test scores. The test score of the typical student will increase, moving the student from the 50th percentile of students to the 58th percentile.

We can translate changes in student test scores into changes in lifetime earnings (Chetty [2007], Hanushek [2010]). Recall that if we replace the average teacher with a superior teacher for one year, the typical student will move from the 50th percentile to the 58th percentile. The lifetime earnings of a student at the 58th percentile is roughly $21,311 greater than the earnings of the student at the 50th percentile, giving us a measure of the per-pupil benefit of our teacher substitution. For a class of 20 students, the economic value of a superior teacher relative to an average teacher is over $426,000. It is important to note that this is the annual value of a superior teacher: a superior teacher generates these gains in earnings each year.

The same logic applies to other sorts of teacher substitutions. For example, suppose we replace an average teacher (50th percentile) with a teacher at the 69th percentile. In this case, the typical student will move from the 50th percentile to the 54th percentile. This change in educational achievement will increase the lifetime earnings per student by $10,607, meaning that the lifetime earnings of a 20-student class increases by over $210,000. On the opposite side of the distribution of teacher productivity, suppose we replace an average teacher with a below-average teacher, for example a teacher at the 31st percentile. The typical student will move from the 50th percentile to the 46th percentile, and the lifetime earnings of students in a 20-student class will decrease by over $210,000.

The evidence on the effects of teacher quality on lifetime earnings has important implications for teacher personnel decisions. There would be a large payoff from taking low-productivity teachers out of the classroom, a process known as "deselecting" teachers. Hanushek and Rivkin (2010) estimate that if the United States were to replace the bottom 8 percent of teachers with average teachers, student test scores would increase by roughly 45 percent. This achievement gain is large enough to eliminate the performance gap between students in the United States and most other countries. For the national economy, deselecting the bottom 8 percent of teachers would increase earnings and annual GDP by roughly $112 trillion.

Characteristics of High-Productivity Teachers

Although it is clear that teachers differ in productivity, the list of characteristics that identify a productive teacher has proven elusive. Teaching requires subtle skills that cannot be easily measured, so it is difficult to predict which teachers will be the most productive. In looking for teacher characteristics that explain productivity differences, researchers have focused on education level (years of graduate coursework), experience (years of teaching), and communication skills (verbal ability).

- *Education level.* There is no evidence that teachers who complete graduate courses in education are more productive than teachers with only a bachelor's degree. In other words, graduate coursework in education does not increase teacher productivity.
- *Experience.* The consensus among researchers is that teaching experience increases productivity, but just for the first few years of teaching.
- *Verbal skills.* The most effective teachers have superior communication skills. Students learn more from teachers who score high on standard tests of verbal ability.

The Effect of Class Size

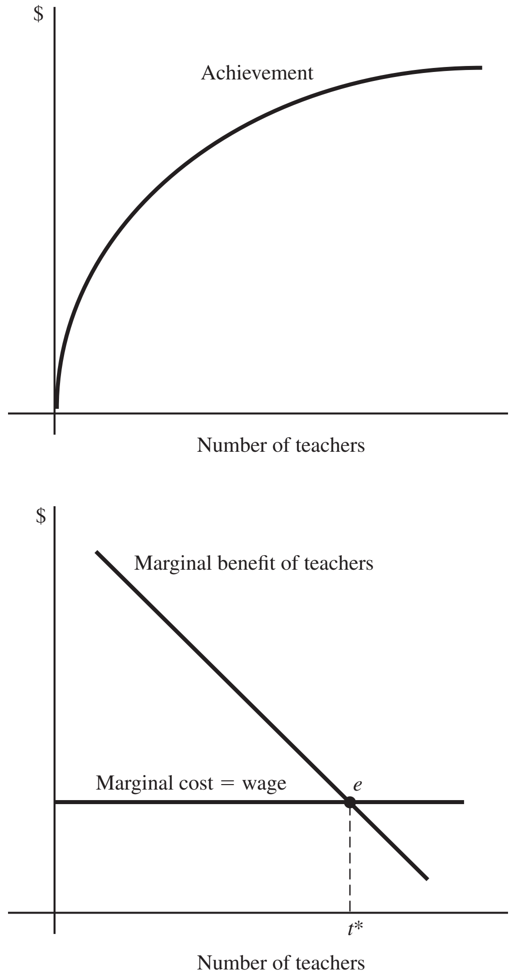
There is convincing evidence that class size affects student achievement. This is sensible because the smaller the class, the greater the teacher time per student. There is some evidence that the achievement gains from small classes are relatively large for low-income and low-achieving students. The fact that students learn more in smaller classes doesn't necessarily mean that small classes are better. Reducing class size requires an increase in the number of teachers, and the question is whether the benefit exceeds the cost.

Figure 12–1 shows how to determine the efficient number of teachers and thus the efficient class size. Consider a school with a fixed population of 600 students that will decide how many teachers to hire. If the school hires t teachers, the class size will be $(600/t)$. For example, if the school hires 40 teachers, the class size will be 15 students. The upper panel of the figure shows student achievement (measured as gains in lifetime earnings) as a function of the number of teachers. The curve is concave, meaning that achievement increases with the number of teachers, but at an decreasing rate. In other words, the benefit of a smaller class diminishes as the class shrinks. The bottom panel shows the marginal benefit of a teacher, equal to the change in achievement from one additional teacher (and the resulting decrease in class size). The marginal-benefit curve is negatively sloped, consistent with the concave total-benefit curve in the upper panel (and diminishing returns to smaller classes). The horizontal line is the marginal cost of a teacher, equal to the wage paid to each teacher.

We can use the marginal principle (explained in the appendix to the book) to show the efficient number of teachers and the efficient class size. Point e shows the efficient choice, where the marginal benefit of a teacher equals the marginal cost. For the first t^* teachers, the gain in student achievement exceeds the extra cost of the teacher, so hiring these teachers is efficient from the social perspective. It is efficient to stop at t^* teachers and a class size of $s^* = (600/t^*)$ because beyond this point, the additional achievement gains are dominated by the extra costs.

The practical policy question is, How does the benefit of additional teachers compare to the cost? Krueger (1999) provides an example of the effects of a

FIGURE 12–1 Efficient Number of Teachers and Class Size



If schoolwide achievement increases with the number of teachers at a decreasing rate, the marginal-benefit curve for teachers is negatively sloped. The efficient number of teachers (t^*) is shown by the intersection of the marginal-benefit curve and the marginal-cost curve at point *e*. The efficient class size is the number of students in the school divided by t^*.

one-third reduction in class size. For a 4-year implementation of smaller classes, the extra cost per student is roughly $7,400. The estimated increase in lifetime earnings is $9,603 per student for men and $7,851 per student for women. In other words, the estimated benefit exceeds the cost by about 30 percent for men and about 6 percent for women. As Krueger notes, there is a great deal of uncertainty about the numbers behind these computations, and he concludes, "[T]hese calculations suggest that the benefit of reducing class size in terms of future earnings is in the same ballpark as the costs."

Teacher Compensation

As we've seen, teachers are the key input to the education production process, and teachers vary in their productivity. It would be reasonable to expect that differences in teacher wages would match the differences in productivity, with superior teachers getting higher wages than mediocre teachers. The market for teachers doesn't work that way. Instead, the wage of the typical teacher is largely determined by two factors: teaching experience (years teaching) and the teacher's own educational attainment.

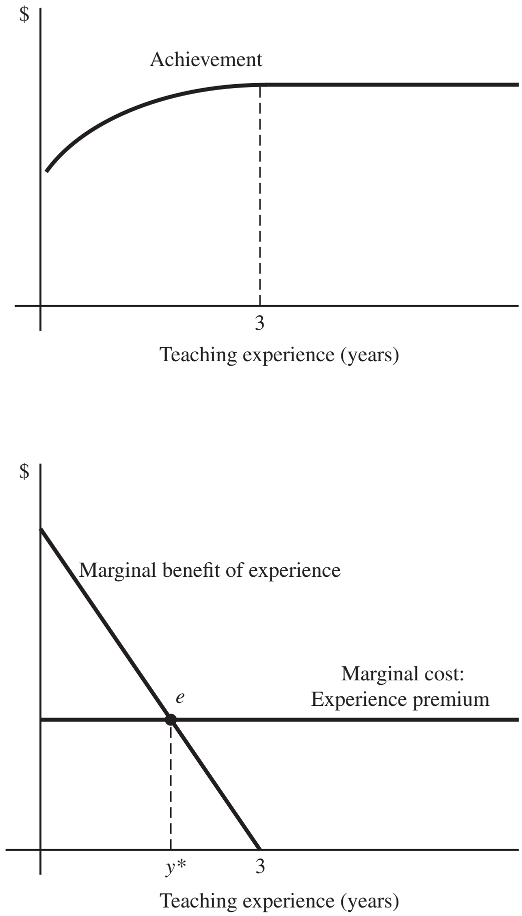
Consider first the role of teaching experience. For the nation as a whole, the average teacher salary increases by roughly $1,000 per year for the first 20 years of teaching. In 2008, the average teacher salary was $49,600, with a range of $38,200 for a beginner, to $40,100 for a teacher with 3 years experience, to $57,800 for a teacher with 20 years experience. In other words, a 20-year teacher earns 1.44 times as much as a 3-year teacher. As we saw earlier in the chapter, studies of the education production function suggest that teacher productivity increases for the first few years of teaching, and then levels off. Therefore, the observed wage gap of 44 percent between a 20-year teacher and a 3-year teacher is puzzling: why pay more for more experienced teachers when on average they are no more productive?

Figure 12–2 illustrates the puzzle. The upper panel shows teacher productivity, measured as student achievement, as a function of the years of teaching experience. Productivity increases for the first three years, and then levels off. In the lower panel, the negatively sloped curve is the marginal benefit of teacher experience, the change in achievement for a one-year increase in experience. The marginal benefit is positive for the first three years, and then is zero for additional years. The horizontal curve is the marginal cost of experience, the increase in the teacher wage per year of experience. Using the marginal principle, the rational choice is shown by point *e*: an efficient school will hire teachers with y^* years of experience, which is clearly less than three years. In this simple model, hiring a teacher with more than 3 years experience is irrational because a school bears a cost (a higher wage) without any benefit.

This simplified model abstracts from some important features of schools. Specifically, much of the on-the-job learning in the first few years of teaching comes from interactions with more experienced teachers. As a result, it may be efficient to pay an experience premium to teachers who boost the productivity of their fellow teachers. This is the logic behind Master-Teacher programs, under which experienced teachers who serve as mentors to new teachers are paid more.

Consider next the role of teacher education. In the United States, the wage premium for a master's degree in education is roughly 26 percent: On average, a teacher with a master's degrees earns roughly 26 percent more than a teacher without one. As we saw earlier in the chapter, there is no evidence that graduate coursework increases

FIGURE 12–2 The Puzzle of the Experience Premium



The productivity of teachers increases with experience for the first few years, then levels off. The marginal-benefit curve for experience is negatively sloped, and reaches a value of zero after a few years (3 in this example). The efficient level of experience (y^*) is shown by the intersection of the marginal-benefit curve and the marginal-cost curve at point e.

teacher productivity. The results of studies of the education production function have been confirmed by recent experience with Teach For America (TFA), the program that puts college graduates in the classroom without providing traditional teacher training. Students assigned to TFA teachers did better on mathematics exams and about the same on reading exams (Staiger and Rockoff, 2010).

This presents another puzzle about teacher wages. If graduate education doesn't increase productivity, the marginal benefit of graduate coursework is zero.

In contrast, a teacher's salary increases with the years of graduate education, so the marginal cost of graduate education is positive. Given the zero marginal benefit and positive marginal cost, the rational response of a school is to hire teachers with zero graduate education. The puzzle is that schools pay for something—graduate course work—that doesn't increase student achievement.

INNOVATION: CHARTER SCHOOLS

Some recent innovations in K–12 education provide important insights into the education production process. A charter school is more flexible than a traditional school in terms of its curriculum and management. Several cities have experimented with "no-excuse" charter schools, which have an extended school day, emphasize discipline, establish high expectations for student achievement, and monitor student progress with frequent testing. Recent studies suggest that no-excuse schools produce significant achievement gains for students.

Promise Academy

A good example of a no-excuse charter school system is Promise Academy in Harlem, New York City. The schools have an extended school day and year, and also provide after-school tutoring for children in need of remedial work in reading and mathematics. On average, an academy student spends about twice as much time on school work as a student in a traditional school. The schools emphasize the culture of achievement and hard work. A key objective of the Promise Academy is to recruit and retain high-productivity teachers, and the schools use student scores on standardized tests to measure teacher performance and reward superior teachers. In the early years of Promise Academy, the turnover rate of teachers was relatively high (almost 50 percent in the first year) as the schools searched for the most effective teachers.

A recent study shows that the Promise Academy schools generate large achievement gains (Dobbie and Fryer, 2009).

- The typical student entering an academy school in sixth grade scored in the 39th percentile among New York students in both mathematics and reading. By the eighth grade, the percentile ranking of the typical academy student rose to 74th in mathematics and 53rd in reading.
- For middle schoolers, the typical black student entered the charter school at the 20th percentile of the white achievement distribution in mathematics (80 percent of white students had higher scores). After three years, the typical black academy student reached the 55th percentile (45 percent of white students had higher scores).

The reported achievement gains are remarkably large, and further study is necessary to confirm the magnitude of the gains and determine their causes. Dobbie and Fryer (2009) provide two tentative conclusions. First, it is plausible that the primary source of the achievement gains is high teacher productivity, a result of deselecting mediocre teachers and hiring superior teachers. Second, it is possible that the large gains resulted from the combined effects of several features of the schools, including higher teacher productivity and the focused learning environment.

Boarding Schools

A recent study of students in boarding schools provides some insights into the effect of the home environment (Curto and Fryer, 2011) on educational achievement. The idea behind a boarding school is to take a student out of an unfavorable home environment. Like other no-excuse charter schools, SEED schools in the District of Columbia and Baltimore have an extended school day, provide extensive after-school tutoring, monitor progress with frequent testing, and have high expectations for student achievement.

The reported achievement effects of SEED schools are large. To measure the effect of a SEED boarding school, suppose we move a student who attends a traditional public school and lives at home into a SEED boarding school. Curto and Fryer (2011) report that each year spent at the new school generates a gain of 9 percentile points in mathematics and 8 percentile points in reading. For example, a student who enters at the 20th percentile in math will move to the 29th percentile after year one, to the 37th percentile after year two, and so on to the 56th percentile after year four. The achievement gains are a bit larger than the gains generated in regular no-excuse charter schools, indicating that there are benefits associated with boarding schools. A careful accounting of all the costs of boarding suggests that the achievement gains are not large enough to offset the substantial cost associated with housing students five days a week.

SPENDING INEQUALITIES AND PUBLIC POLICY

The traditional funding source for K–12 education is the local property tax. Starting in the 1970s, citizens in many states challenged the constitutionality of property-tax funding, citing the substantial inequalities across school districts in spending per pupil and student achievement. In most state constitutions, education is identified as a fundamental right for all citizens. In contrast, education is not mentioned in the U.S. Constitution, so inequalities in spending are not proscribed by the U.S. Constitution.

As a result of court cases challenging the constitutionality of education finance, states have developed several alternative notions of equity in K–12 education (Yinger, 2004).

1. *Adequacy.* Each local school district provides an education that meets or exceeds some minimum statewide standard.
2. *Access equality.* Voters in each school district have access to the same effective tax base. This means that a given property tax rate will generate the same revenue per pupil in every school district.

3. *Equality.* Each school district provides the same level of education. Although several court cases have adopted a standard of educational equality, no court has indicated how a state is to measure equality.

Given these alternative notions of equity, there is substantial variation across states in systems of education finance. Most states that confront the issue of equity focus on reducing spending inequalities.

Intergovernmental Grants: Foundation Plans

The states use several types of intergovernmental grants to address inequalities across school districts. Under a foundation grant, a state gives larger grants to school districts with relatively low property tax bases. The foundation grant per pupil is

$$Grant = Foundation\ level - Foundation\ tax\ rate \cdot Local\ property\ value\ per\ pupil$$

To illustrate, consider a state with a foundation level of \$8,000 and a foundation tax rate of 3 percent (0.03). For a school district with \$200,000 of property value per pupil, the foundation grant is \$2,000:

$$\text{Foundation grant} = \$8{,}000 - 0.03 \cdot \$200{,}000 = \$2{,}000$$

The foundation grant equals the difference between the foundation level and the local revenue that *could* be generated if the district were to impose the foundation tax rate.

Note that the foundation grant is independent of the local tax rate. Continuing our example, Table 12–4 shows spending options with different local tax rates. Given the local tax base of \$200,000 per pupil, for every percentage point of the tax, local tax revenue and spending increases by \$2,000. If the school district chooses a tax rate of 2 percent, it generates \$4,000 of local revenue per pupil, and combined with a foundation grant \$2,000, the district can spend \$6,000 per pupil. A tax rate of 2.5 percent generates an additional \$1,000 in local property taxes, but doesn't change the grant. In other words, if the school district spends more than \$2,000 on education (the foundation grant), the local cost per dollar on education is one dollar.

The idea behind a foundation grant is to provide more money to school districts with lower tax bases. The foundation tax rate determines the rate at which the grant varies with the local property tax base. For example, if the foundation tax rate is 0.03, the grant decreases by \$0.03 per dollar of the local tax base. For a district with only \$100,000 of property value per pupil, the foundation grant is \$5,000:

$$\text{Foundation grant} = \$8{,}000 - 0.03 \cdot \$100{,}000 = \$5{,}000$$

This district gets an additional \$3,000 in grant money per pupil because its tax base is \$100,000 lower ($0.03 \times \$100{,}000 = \$3{,}000$).

TABLE 12–4 Tax and Spending Options with a Foundation Grant

Local Tax Rate	Local Tax Revenue	Foundation Grant	Education Spending
2%	\$4,000	\$2,000	\$6,000
2.50%	\$5,000	\$2,000	\$7,000

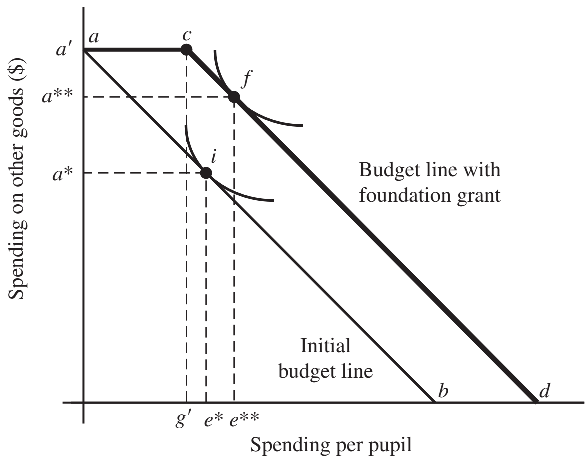
Response to a Foundation Grant

How does a foundation grant affect spending in the school district? As we'll see later in the book, the median-voter model predicts that the outcome of voting on a local issue will be the choice of the median voter. On the issue of education spending, the median voter divides the voting population into two groups, one with a lower level of desired spending, and a second with a higher level. Therefore, to predict the effect of a foundation grant on the educational spending of a particular school district, we predict the effect of the grant on the choices of the median voter.

Figure 12–3 shows the budget lines and indifference curves of the median voter. The horizontal axis measures per pupil spending, and the vertical axis measures spending on other goods. The budget line *ab* shows the initial trade-off between education and other goods: the median voter has a total of a' per pupil to spend on education and other goods, and every dollar spent on education decreases spending on other goods by one dollar. Suppose that before the foundation grant, the voter maximizes utility at point *i* (education spending = e^* and spending on other goods = a^*), and this is the point chosen by the local school district.

Consider the effects of a foundation grant equal to g'. The grant shifts the budget line outward from *ab* to *acd*. Point *c* is in the new budget set because the voter can use the grant to get g' worth of education while spending all of his or her own money (a') on other goods. The points along the line connecting *c* and *d* are in the new budget set. Starting from point *c*, the voter faces a dollar-for-dollar trade-off between education and other goods. The shift of the budget line moves the utility-maximizing

FIGURE 12–3 The Effects of a Foundation Grant



A foundation grant shifts the budget line of the median voter from *ab* to *acd*, and the utility maximization point moves from point *i* to point *f*. The grant increases the spending on education by $e^{**} - e^*$ and increases spending on other goods by $a^{**} - a^*$.

point from i to f, with education spending e^{**} and spending on other goods a^{**}. In other words, the grant increases spending on both education (from e^* *to* e^{**}) and other goods (from a^* *to* a^{**}).

How does the local community use an education grant to increase spending on other goods? The district can fund its education program of e^{**} with the grant g' and only $(e^{**} - g')$ of local tax revenue. The grant decreases the local contribution to education from e^* to $(e^{**} - g')$, so the school district cuts local taxes and allows its citizens to spend more money on other goods. From the perspective of local citizens, the foundation grant is equivalent to an increase in income, which increases the demand for all "normal" goods, including education and other goods. By cutting taxes to free up money to spend on other goods, the school district is simply responding to the choice of its citizens, as represented by the median voter.

The median voter model predicts that a foundation grant will have the same effects on education spending as an increase in income. As a result, the model predicts that only a small fraction of a foundation grant will be spent on education. Empirical studies suggest that for intergovernmental grants in general, the fraction of a grant spent on the target local public good is actually much higher, around 40 percent (Oates, 1999). This is known as the flypaper effect: A relatively large fraction of grant money sticks where it first hits (the school district), so less is available to spend on other goods. In the case of education grants, the fraction of a grant spent on education is between 30 percent and 65 percent (Card and Payne, 2002).

How do foundation grants affect the degree of inequality in education spending across school districts? The largest grants are given to the school districts with the lowest property values, so the greatest stimulus for education will occur in low-wealth districts. Empirical studies suggest that systems of foundation grants do in fact reduce education spending gaps (Card and Payne [2002], Chaudhary [2009]).

Matching Grants: Guaranteed Tax Base

The foundation grant is a nonmatching grant in the sense that the dollar value of the grant is independent of the local tax rate and local tax revenue. In contrast, a matching grant increases with the amount of local revenue generated, so the local cost per dollar of additional spending on education will be less than one dollar. Consider a matching grant with a match rate of 0.25. An additional \$0.80 of local tax effort increases the state contribution by one-fourth of \$0.80, or \$0.20, bringing total spending to \$1.00. In this case, the local cost of a dollar of additional education spending is only \$0.80. In general, the local cost per dollar of education spending is $(1/(1 + m))$, where m is the match rate. In our example, the local cost per dollar of education is $(1/1.25) = \$0.80$

Compared to a foundation grant, a matching grant has a larger stimulative effect because it has both an income effect and a substitution effect. Like the foundation grant, the matching grant increases the median voter's real income, increasing the desired spending on all "normal" goods, including education (the income effect). A matching grant also decreases the opportunity cost of education spending, generating a substitution effect that increases education spending as the median

TABLE 12–5 Options under a Guaranteed Tax Base Plan

Tax Rate	Local Tax Revenue	Grant	Total Spending
2%	$4,000	$1,000	$5,000
2.50%	$5,000	$1,250	$6,250

voter substitutes education for other goods. In our example, a matching rate of 0.25 decreases the opportunity cost (the local cost) of education spending from $1.00 to $0.80.

The traditional matching-grant program for education is the guaranteed tax base plan (GTB), also known as district power equalizing. The state specifies a guaranteed tax base per pupil, meaning that each school district has access to the same *effective* tax base. The grant per pupil is

$$\begin{aligned} \textit{GTB grant} = \textit{Local tax rate} \cdot (&\textit{Guaranteed tax base per pupil} \\ &- \textit{Local tax base per pupil}) \end{aligned}$$

The state picks the guaranteed tax base and each school district picks a tax rate. The higher the tax rate, the greater the local revenue and the larger the grant.

To illustrate, suppose the guaranteed tax base is $250,000 and the local tax base is $200,000. In other words, the gap between the two tax bases is $50,000. As shown in Table 12–5, with a tax rate of 2 percent, the district raises $4,000 locally and gets a grant of $1,000, for a total of $5,000. With the higher tax rate of 2.5 percent, local tax revenue is $5,000 and the grant is $1,250, for a total of $6,250. Education spending increases by $1,250, but local tax revenue increases by only $1,000, so the local cost per dollar of education is $0.80. In general, the local cost per dollar of education under a GTP plan equals the ratio of the local tax base to the guaranteed tax base. In our example, the local cost per dollar is $200,000/$250,000 = $0.80.

Effects of Equalization Plans on Spending and Achievement Inequalities

A number of studies have explored the effects of equalization plans on spending inequalities across school districts. For the states whose reforms were triggered by the courts, spending inequalities decreased significantly (Evans, Murray, Schwab, 1997; Card and Payne, 2002). Spending per pupil increased in low-spending school districts, but didn't change much in high-spending districts, so reductions in inequalities resulted largely from "leveling up." To illustrate the leveling up, the changes in school districts with different starting points were as follows.

- *Low spending.* For a school district in the 25th percentile (25 percent of districts spent less), education spending increased by 27 percent.
- *Medium spending.* For the median school district (50th percentile), spending increased by only 15 percent
- *High spending.* For a school in the 75th percentile, spending did not change significantly.

What about the states that reformed their education finance systems without the pressure of court orders? In these states, reform plans did not have significant effects on either spending per pupil or spending inequalities across school districts. The general lesson is that court mandates are necessary for real reform.

Michigan is among a handful of states that has taken complete control of K–12 education finance (Courant and Loeb [1997], Chaudhary [2009]). In 1993, the state eliminated the local property tax as a source of education funding, and increased the state sales tax to make up for lost revenue. The state now determines educational spending for all but the wealthiest school districts. The reform package increased spending per pupil in small rural districts and decreased spending per pupil in poor urban areas and rich suburban areas. The plan reduced spending inequalities significantly.

The Michigan reform package also affected inequality in student achievement. In school districts where funding increased, the additional money was spent on reducing class size and increasing teacher salaries. As a result, the mathematics test scores of fourth graders increased, but the scores of seventh graders did not change. Chaudhary (2009) concludes that in general, education finance reforms will reduce achievement inequalities if increases in spending are used to increase teacher salaries in a way that increases teacher productivity.

How do central-city schools fare under the reform of the education finance system? As explained earlier in the book, central-city schools have relatively high costs because a large fraction of their students come from poor families. These schools devote more time and resources to security measures, dealing with family and health crises, and teaching children with weak educational preparation and English skills. Central-city schools have relatively high costs and had above-average spending levels before reforms were implemented. As a result, they either receive relatively small benefits from equalization programs, or actually get less money (Courant and Loeb (1997); Duncombe and Yinger (1997)). If the formulas were modified to incorporate cost differences, some central-city school districts would receive two or three times as much grant money.

EDUCATION IN CENTRAL CITIES

The production-function approach provides important insights into why the educational achievement in central cities is relatively low.

1. *Household sorting*. As we saw earlier in the book, households tend to sort with respect to the demand for local public goods and the demand for housing. As a result, households are segregated with respect to income and educational attainment. Low-income central-city neighborhoods have less favorable home environments and peer groups, generating relatively low educational achievement.
2. *Higher production cost.* As we've seen, it costs more to produce education in central cities, and so a given education budget generates less in terms of student achievement.

Given the challenges associated with providing education in low-income, high-cost central cities, what is the appropriate policy? The key input to the education production process is teachers, and it would be sensible to focus extra resources on changing the quantity and quality of teachers. On the quality side, there is a large payoff from increasing teacher productivity, and the best way to increase productivity is to replace mediocre teachers with better teachers. On the quantity side, a decrease in class size increases student achievement. The policy question is whether the benefits of increased achievement justify the cost of smaller classes. While it is too early to draw any conclusions from the experiences of Promise Academy and other experimental charter schools, changes in the learning environment might be an important part of a strategy to increase achievement.

SUMMARY

1. The production function summarizes the relationship between education inputs and achievement. The most important inputs are the home environment, the peer group, and teachers.
2. Teacher wages increase with experience and graduate education, despite the fact that productivity is unaffected by graduate coursework, and productivity increases with experience for only the first few years of teaching.
3. Teachers vary in productivity, and substituting an above-average teacher for an average teacher generates large increases in the lifetime earnings of students. Similarly, deselecting mediocre teachers generates large gains in earnings.
4. A decrease in class size increases achievement, along with the cost per pupil.
5. Some types of charter schools generate large achievement gains.
6. States use matching and nonmatching grants to reduce spending inequalities across school districts. A matching grant has a larger stimulative effect because it has both an income effect and a substitution effect.

APPLYING THE CONCEPTS

1. **Trade-off from Mixing.** Hiram is a high-achieving student who will be placed in one of three classes: a high-achiever class, a medium-achiever class, or a low-achiever class. The following table shows some of the gains and losses to Hiram and his new classmates. Suppose the gain of the low achiever class exceeds the gain to the medium achiever class.
   - **a.** The most efficient placement is the low class if the low-achiever gain is at least _____ and the gap between the low-achiever gain and the medium-achiever gain is at least _____.
   - **b.** The most efficient placement is the medium class if the medium-achiever gain is at least _____ and the gap between the low-achiever gain and the medium-achiever gain is no more than _____.

c. Provide an example for (*b*). In this case, the aggregate gain is _____ and the loss of Hiram relative to being in the high-achiever class is _____.

Gains	Low Achiever Class	Medium Achiever Class	High Achiever Class
Gains of new classmates			0
Gains of Hiram	−6	−2	0
Aggregate gain			0

2. **Verbal Ability versus Experience**. Consider two inputs to teacher productivity, verbal ability, and teaching experience.
   a. Draw a production isoquant for education, with years of experience on the horizontal axis and the teacher's verbal ability (SAT score) on the vertical axis.
   b. Given the experience premium (about $1,000 per year), use the input choice model (isoquant and isocost curves) to show the cost-minimizing combination of experience and verbal ability.
   c. Under what circumstances will the cost-minimizing combination involve 10 years of experience? Illustrate with your graph.
3. **Verbal Ability versus Graduate Coursework**. Consider two inputs to teacher productivity, verbal ability, and graduate coursework.
   a. Draw a production isoquant for education, with graduate courses on the horizontal axis and the teacher's verbal ability (SAT score) on the vertical axis.
   b. Given the coursework premium, use the input choice model (isoquant and isocost curves) to show the cost-minimizing combination of graduate coursework and verbal ability.
4. **Efficient Number of Teachers.** In Figure 12–1, suppose the total achievement curve can be described as Achievement = $120 \cdot t^{1/2}$, so the marginal-benefit curve for teachers can be described as marginal benefit = $60/t^{1/2}$.
   a. Compute the efficient number of teachers for the following teacher wages: {6, 10}.
   b. Illustrate with a graph like the one shown in Figure 12–1.
   c. If you know calculus, show how we go from the expression for achievement to the expression for marginal benefit.
5. **Benefit-Cost of Graduate Coursework**. Consider the empirical evidence concerning the productivity effects of graduate coursework in education and the coursework premium. Take the school perspective and draw a figure like Figure 12–2 to show the efficient level of graduate coursework.
6. **Foundation Grant.** Consider a state with a foundation level of $11,000 and a foundation tax rate of 3 percent (0.03). The school district has $200,000 of property value per pupil. Fill the blanks.

Local Tax Rate	Local Tax Revenue	Foundation Grant	Education Spending
2%			
3%			

7. **The Flypaper Effect.** Consider a school district with a median income of \$50,000 and initial spending on education of \$5,000 per pupil. Suppose the state provides an intergovernmental grant of \$1,000 per pupil. Assume that the income elasticity of demand for education is 1.0.
   a. For the median voter, the desired spending on education increases to ________ per pupil. In other words, ________ percent of the grant is spent on education, leaving ________ percent to be spent on other goods. Illustrate with a graph like Figure 12–3.
   b. The flypaper effect suggests that the actual spending on education will increase to about ________ per pupil. Use your graph to illustrate the flypaper effect.
   c. How would your answer to (*a*) change if the income elasticity of demand for education were 2.0 rather than 1.0?
8. **Income and Substitution Effects of Matching Grant.** Using Figure 12–3 as a starting point, suppose the state provides a matching grant, and the budget line for the matching grant goes through point *f* (the choice of the median voter under a foundation grant). Show the choice of the median voter under the matching grant. The matching grant generates a ________ (larger, smaller) spending per pupil because. . . .

REFERENCES AND ADDITIONAL READING

1. Burke, Mary, and Tim Sass. "Classroom Peer Effects and Student Achievement," National Center for Analysis of Longitudinal Data in Education Research Working Paper 18 (2008).
2. Card, David, and Alan B. Krueger. "School Resources and Student Outcomes: An Overview of the Literature and New Evidence from North and South Carolina." *Journal of Economic Perspectives* 10, no. 4 (Fall 1996), pp. 31–50.
3. Card, David, and Abigail Payne. "School Finance Reform, the Distribution of Test School Spending, and the Distribution of Student Test Scores," *Journal of Public Economics* 83 (2002), pp. 49–82.
4. Chaudhary, Latika, "Education Inputs, Student Performance, and School Finance Reform in Michigan," *Economics of Education Review* 28 (2009), pp. 90–98.
5. Chetty, Raj, John Friedmand, Nathaniel Hilger, Emmanueal Saez, Diane Witmore Schanzenbach, and Danny Yagan, "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," NBER Working Paper (2010).
6. Courant, Paul N., and Susanna Loeb. "Centralization of School Finance in Michigan." *Journal of Policy Analysis and Management* 16, no. 1 (1997), pp. 114–36.
7. Curto, Vilsa and Roland Fryer, "Estimating the Returns to Urban Boarding Schools: Evidence from SEED," NBER Working Paper (2011).
8. Ding, Weili, and Steven Lehrer. "Do Peers Affect Student Achievement in China's Secondary Schools?" *Review of Economics and Statistics* 89 (2007), pp. 300–12.
9. Dobbie, Will and Roland Fryer. "Are High-Quality Schools Enough to Close the Achievement Gap? Evidence from a Bold Social Experiment in Harlem," Working Paper, Harvard University 2009.

10. Duncombe, William, and John Yinger. "Why Is It So Hard to Help Central-City Schools?" *Journal of Policy Analysis and Management* 16, no. 1 (1997), pp. 85–113.
11. Evans, William N., Sheila E. Murray, and Robert M. Schwab. "Schoolhouses, Courthouses, and Statehouses after *Serrano*." *Journal of Policy Analysis and Management* 16, no. 1 (1997), pp. 10–31.
12. Hanushek, Eric A. "Throwing Money at Schools." *Journal of Policy Analysis and Management* 1 (1981), pp. 19–24.
13. Hanushek, E.A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy, 100*, 84–117.
14. Hanushek, Eric. "The Failure of Input-Based Schooling Policies," *Economic Journal* 113 (2003), pp. 64–98.
15. Hanushek, Eric A. "The Economic Value of Higher Teacher Quality," NBER Working Paper (2010).
16. Hanushek, Eric A., and Steven G. Rivkin. "Generalizations about Using Value-Added Measures of Teacher Quality," *American Economic Review: Papers and Proceedings* 100 (2010), pp. 267–71.
17. Kreuger, Alan, "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics* (1999), pp. 497–532.
18. Martin, M.O., I.V.S. Mullis, & P. Foy. *TIMSS 2007 International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades* (2008). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
19. Martin, M.O., I.V.S. Mullis, & P. Foy. *TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades* (2008). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
20. Oates, Wallace E., An Essay on Fiscal Federalism," Journal of Economic Literature 37 (3), (1999), pp 1120–1149.
21. Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. "Teachers, Schools, and Academic Achievement." NBER Working Paper Number 6691 (1998).
22. Rouse, Cecilia. "Schools and Student Achievement: More Evidence from the Milwaukee Parental Choice Program." *Federal Reserve Bank of New York Economic Policy Review* (March 1998), pp. 61–76.
23. Staiger, Douglas O., and Jonah E. Rockoff. "Searching for Effective Teachers with Imperfect Information." *Journal of Economic Perspectives* 24, no. 3 (Summer 2010): 97–118.
24. Sund, Krister. "Estimating Peer Effects in Swedish High School Using School, Teacher and Student Fixed Effects." *Economics of Education Review* 28 (2009), pp. 329–36.
25. U.S. Department of Education, *Digest of Education Statistics 2010*. Washington DC: National Center for Education Statistics (2010).
26. Yinger, John (ed.) 2004. *Helping Children Left Behind: State Aid and the Pursuit of Educational Equity*. Cambridge: MIT Press.

CHAPTER 13

Crime

Erle Gardner, the writer of detective stories, was paid by the word, and his villains were always killed by the last bullet in the gun. When asked why his heroes were so careless with their first five shots, he responded, "Every time I say bang in the story, I get three cents. If you think I'm going to finish the gun battle while my hero has fifteen cents worth of unexploded ammunition in his gun, you're nuts."

—Bartlett's Book of Anecdotes (2000)

The economic approach to crime is based on the notion that criminals base their decisions on the costs and benefits of crime and respond to incentives. As a society, we can reduce crime by adding police officers, prosecutors, and prison cells to increase the certainty of severe penalties for crime. We can also reduce crime by adding teachers and other school resources that transform dropouts into high-school graduates who are less likely to commit crime because they have better lawful opportunities. As a society, we make the difficult choice of how much crime to allow. Although a crime-free environment sounds appealing, what would we sacrifice to get it? Some crimes are more costly to prevent than to experience, so the socially efficient level of crime is positive.

CRIME FACTS

The Federal Bureau of Investigation (FBI) collects data from local police departments on seven index crimes, divided into personal and property crimes:

- **Personal crime.** The victim of a personal crime is placed in physical danger. For some crimes, the objective is to injure the victim (homicide, rape, aggravated assault). For other crimes, the objective is to steal property, but the criminal uses a show of force to coerce the victim (robbery).
- **Property crime.** These are crimes of stealth rather than force and include burglary (illegal entry of a building), larceny (purse snatching, pocket picking, and bicycle theft), and auto theft.

TABLE 13–1 FBI Index Crimes, 1960–2003

	Number of Crimes per 100,000 People					
	1960	1970	1980	1990	1995	2003
Personal Crime						
Murder	5.0	7.8	10.2	9.4	8.2	5.7
Rape	9.5	18.6	36.8	41.2	37.1	32.1
Aggravated assault	85	177	299	424	418	295
Robbery	60	187	251	257	221	142
Property Crime						
Auto theft	182	457	502	658	561	433
Larceny	1,024	2,124	3,167	3,184	3,045	2,415
Burglary	504	1,152	1,684	1,236	988	741
Total Index Crimes	1,870	3,949	5,950	5,820	5,278	4,064

Source: U.S. Federal Bureau of Investigation. *Crime in the United States*, *Various Years*. Washington DC: U.S. Government Printing Office.

The FBI data provide only a partial picture of the crime scene. Among the crimes omitted in the *Uniform Crime Reports* are disorderly conduct, shoplifting, arson, employee theft, and drug-related offenses.

Table 13–1 lists the crime rates for the period 1960 to 2003, expressed as the number of crimes per 100,000 people. The total crime rate rose from 1960 to 1980, fell slowly between 1980 and 1995, and then dropped rapidly between 1995 and 2003. Later in the chapter, we'll explore the reasons for the dramatic drop in crime in the 1990s. The FBI data include only the crimes that are reported to the police—38 percent of all property crimes and 48 percent of personal crimes. A more complete picture comes from the victimization surveys of the Department of Justice. The surveys indicate that the overall level of crime has decreased since its peak in 1981.

The Victims of Crime

Who are the victims of crime? As Table 13–2 shows, victimization rates vary with income and place of residence. Another factor is race (not shown in the table).

- **Income.** Victimization rates for violent crime decrease as income increases. For example, a person in a household with an income less than $7,500 is nearly three times as likely to be victimized as a person in a household with an income above $75,000. Differences in victimization rates for property crime are not so clear-cut. Although the lowest income group has a relatively high victimization rate, the differences between other income groups are relatively small.
- **Place of residence.** Victimization rates are lowest in rural areas and highest in central cities. The suburbs fall between the two extremes.
- **Race.** For violent crime, the victimization rate is 29.1 for blacks and 21.5 for whites. Blacks are also more frequently the victims of property crime.

TABLE 13–2 Criminal Victimization Rates, 2003

		Violent (per 1,000 people)			Property (per 1,000 households)			
	Population (million)	Total	Robbery	Assault	Total	Burglary	Motor Vehicle Theft	Theft
Household Income								
Less than $7,500	8	49.9	9	39.3	204.6	58	6.3	140.3
$7,500–$14,999	16	30.8	4	25	167.7	42.2	7.3	118.3
$15,000–$24,999	25	26.3	4	21.5	179.2	38.4	8.9	131.9
$25,000–$34,999	24	24.9	2.2	21.8	180.7	35.3	12.3	133.1
$35,000–$49,999	32	21.4	2.1	18.3	177.1	27.6	9.5	140
$50,000–$74,999	35	22.9	2	20.4	168.1	24.9	8.4	134.7
$75,000 or more	48	17.5	1.7	15.4	176.4	20.8	11.9	143.7
Region								
Northeast	45	21	2.7	18.1	122.1	20.5	7.2	94.4
Midwest	56	23.6	2.7	19.4	160.2	32.5	6.9	120.9
South	86	21.1	2.5	17.8	160.5	32.2	7.8	120.4
West	52	25.2	2.1	22.5	207.4	30.6	15.2	161.6
Residence								
Central city	66	28.2	3.7	23.8	216.3	38.7	13	164.7
Suburban	116	21.3	2.3	18.1	144.8	24	9.3	111.6
Rural	57	18.6	1.6	16.4	136.6	30.5	4	102.1

Source: U.S. Bureau of Justice. Criminal Victimization in the United States, 2003. Washington DC, 2005.

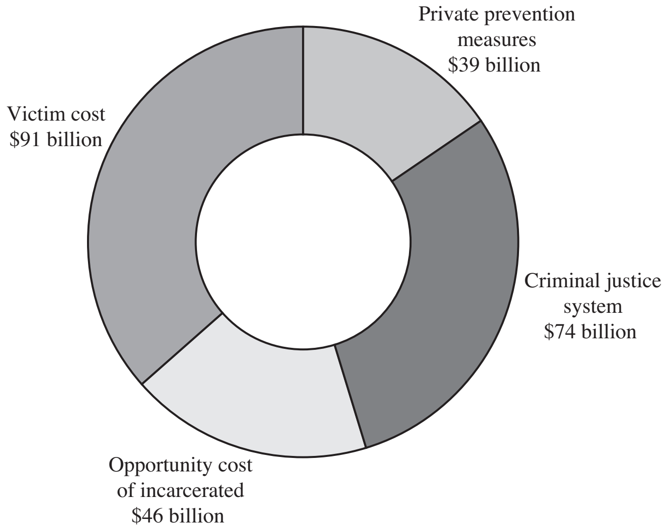
The Costs of Crime

Figure 13–1 (page 341) shows the estimated costs of crime in 1992. The costs incurred by victims include the value of lost property, medical expenses for injuries, the opportunity cost of lost work time, pain and suffering, and the value of lives cut short. The costs of the criminal justice system include the costs of police, the courts, and correction facilities. Citizens spent about $39 billion on their own prevention measures, including locks and hired guards. The opportunity cost of having 1.35 million people in jails and prisons instead of working was $46 billion. Altogether, the cost of crime was $250 billion per year, or 3.8 percent of GDP.

THE RATIONAL CRIMINAL

The economic approach to crime is based on the notion that criminals are rational like everyone else and commit a crime if the benefit exceeds the cost. Of course, crime is an uncertain enterprise, and potential criminals must consider the likelihood of different outcomes. We'll start our discussion of the rational criminal with a simple crime that most of us have at least considered, double parking.

FIGURE 13–1 The Costs of Crime



Source: Based on Richard Freeman. "Why Do So Many Young American Men Commit Crimes and What Might We Do About It?" *Journal of Economic Perspectives* 10, no. 1 (1996), pp. 25–42.

The Economics of Double Parking

Suppose you have an opportunity to buy the last ticket to a concert, but to get it, you must double park your car, violating the law. If your consumer surplus from getting the concert ticket is $44, that's your benefit from the crime of double parking. Suppose you have a 50 percent chance of getting caught and paying a fine of $36. Is the risk of double parking—a 50–50 chance of getting $44 or losing $36—worth taking?

People differ in their willingness to take the risks associated with crimes. Some people, given an equal chance of either getting $44 or losing $36, will take the risk. Other people won't, but they might take the risk if the benefits were higher or the costs were lower. For example, if the consumer surplus from the ticket were $200, more people would commit the crime. Similarly, if the fine and the probability of getting caught were low enough, more people would break the law. In other words, as the benefit of the crime increases relative to the cost, more people will commit the parking crime.

What if you believe that violating the law is simply wrong? Most people have an underlying aversion to engaging in antisocial behavior such as crime and experience an anguish cost when they commit a crime. We will incorporate anguish cost into the economic approach to crime and see what happens when people have different anguish costs.

Expected Utility and the Decision to Commit Crime

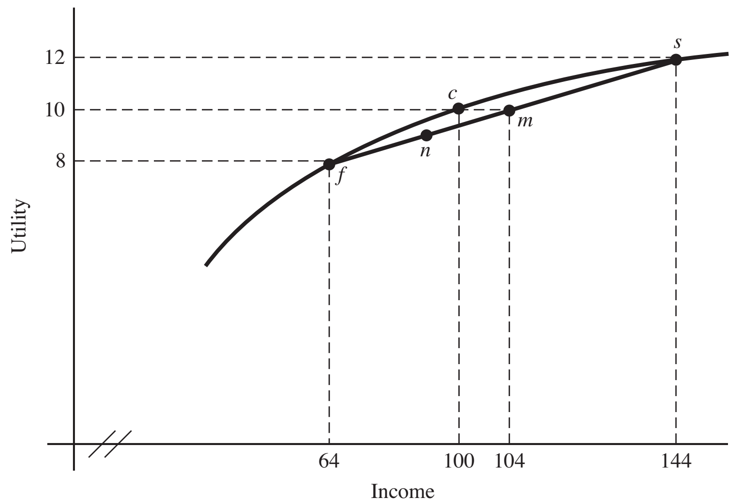
We'll use a numerical example to illustrate the decision to commit a burglary. To keep the example simple and transparent, we'll keep the numbers small. One way to interpret the numbers in our example is to imagine that they are in thousands of dollars and apply to a 10-year period. Suppose a person can earn \$100 in a lawful occupation and can supplement this income with money earned in weekly burglaries. His objective is to maximize his expected utility, and he will choose either a lawful or a criminal life, whichever generates the highest expected utility.

The crime decision is based on utility maximization, so we must translate monetary amounts into a measure of utility or satisfaction. In Figure 13–2, the utility curve shows the relationship between income (on the horizontal axis) and utility (on the vertical axis, measured in utils). The utility curve is concave, reflecting the assumption of diminishing marginal utility of income: As income increases, utility increases, but at a decreasing rate. This means that the first dollar of income is worth more in utility terms than the second, which is worth more than the third, and so on. We will use the following simple utility function:

$$Utility = (Income)^{1/2}$$

In words, utility is the square root of income.

FIGURE 13–2 Expected Utility from Crime



The utility curve is concave because of diminishing marginal utility of income. Point *c* shows the certain lawful outcome, with utility = 10. Point *s* shows the outcome with successful crime. Point *f* shows the outcome with a failed attempt at crime. With a probability of prison = 0.50, the expected utility from crime is the average of 12 utils (point *s*) and 8 utils (point *f*), or 10 utils, shown by point *m* (halfway between *s* and *f*).

TABLE 13–3 The Expected Utility of Crime

	Baseline	Higher Probability of Prison	Longer Prison Term	Less Loot	Higher Income	Lower Probability of Prison
Lawful income ($)	100	100	100	100	400	100
Loot ($)	44	44	44	21	44	44
Probability of prison	0.5	0.75	0.5	0.5	0.5	0.25
Prison time	0.36	0.36	0.51	0.36	0.36	0.36
Lawful utility = (lawful income)$^{1/2}$	**10**	**10**	**10**	**10**	**20**	**10**
Utility from successful crime						
Net income = Legal income + Loot	144	144	144	121	444	144
Utility = (Net income)$^{1/2}$	12	12	12	11	21	12
Utility from failed crime						
Prison cost = prison time · legal income	36	36	51	36	144	36
Net income = Legal income – prison cost	64	64	49	64	256	64
Utility = (Net income)$^{1/2}$	8	8	7	8	16	8
Expected utility from crime (utils)	**10**	**9**	**9.5**	**9.5**	**18.5**	**11**

The first column of numbers in Table 13–3 shows how to compute the utility associated with the lawful and criminal options. The first four numbers are the values of key parameters. The lawful income is $100, and a program of weekly burglaries generates $44 worth of loot over the time period considered. If a person commits crime, the probability of eventually being caught and sent to prison is 0.50. The prison term for a criminal who is caught is 0.36 units of time in prison (e.g., 3.6 years over a decade). A criminal who is caught also loses any loot stolen, and supervision after being released from prison prevents a return to crime.

We can compute the utility levels associated with three possible outcomes. First, the lawful option generates an income of $100 and thus a certain utility of 10 utils (the square root of $100). This is shown as point *c* in Figure 13–2. Second, a successful criminal earns a net income of $144 (equal to $100 + $44 in loot) and receives utility = 12 utils (point *s*). An unsuccessful criminal spends 0.36 units of time in prison, leaving only 0.64 units of time to earn lawful income. The net income for the failed criminal is $64 (0.64 times $100) and utility = 8 utils (point *f*).

A potential criminal doesn't know ahead of time whether he will succeed or fail at crime. But since we know the utilities of success and failure and the probability of each, we can compute the expected utility of crime, equal to a weighted average of the two values, with the probabilities as the weights:

$$EU\{U_1, U_2, p_1, p_2\} = p_1 \cdot U_1 + p_2 \cdot U_2$$

p_1 is the probability of an outcome that generates a utility of U_1 utils. In our example, the expected value of crime is 10 utils:

$$EU\{12, 8; 0.50, 0.50\} = 0.50 \cdot 12 + 0.50 \cdot 8 = 10 \text{ utils}$$

In graphical terms, the utility is shown by the midpoint of the line connecting the two utility points (point *m* is midway between *s* and *f*). It is the midpoint because each outcome is equally likely.

The person will choose crime if the expected utility of crime exceeds the certain utility of remaining lawful. In the example shown in the first column of Table 13–3, the lawful utility equals the expected utility of crime, so the person is indifferent between the two options. The person is just as well off with either a certain lawful utility of 10 utils or a risky crime career, with equal chances of either 12 or 8 utils.

If we look at the monetary payoffs from crime and lawful activity, there is a puzzle. A person can earn \$100 for certain in the lawful world or have a 50-50 chance at either \$144 or \$64 in the criminal world. The expected income in the criminal world is \$104 (0.50 times \$144 + 0.50 times \$64), or \$4 higher than the certain income in the lawful world. If a criminal life provides a higher expected income, why doesn't the person have a distinct preference for crime over lawful activity?

The reason for indifference rather than a distinct preference for crime is diminishing marginal utility of income. Using the lawful income as a starting point, a switch to crime gives a 50 percent chance of gaining \$44 (increase income from \$100 to \$144) and a 50 percent chance of losing \$36 (decrease income from \$100 to \$64). Because of diminishing marginal utility of income, the pleasure of gaining \$40 of income (moving upward along the utility curve) equals the pain of losing only \$36 of income (moving downward along the utility curve). As a result, the person is indifferent between a lawful life with certain income of \$100 and a criminal life with equal chances of \$144 or \$64. Because of diminishing marginal utility of income, the person is indifferent between lawful activity and crime, even though crime generates a higher expected income.

Preventing Crime

Our example shows that our potential criminal is indifferent between crime and a lawful life. We can tip the balance away from crime by changing the values of the key parameters. In the second column of Table 13–3, the probability of imprisonment increases to 0.75, meaning that a criminal is more likely to lose the loot and go to prison. This change doesn't affect the utility levels associated with criminal success and failure but simply changes the probability of each outcome—and the expected utility of crime:

$$EU\{12, 8; 0.25, 0.75\} = 0.25 \cdot 12 + 0.75 \cdot 8 = 9 \text{ utils}$$

The expected utility from crime is now less than the lawful utility, so the person will not commit crime. In other words, an increase in the certainty of punishment decreases crime.

In Figure 13–2, the increase in the probability of imprisonment moves the crime outcome from point *m* to point *n*. Point *m* is the midpoint between points *s* (success) and *f* (failure), showing what happens with a 50 percent chance of failure. As the probability of failure (prison) increases, we move closer to point *f* and farther from point *s*. Point *n* is three-fourths of the distance between *s* and *f*, showing what happens when the probability of failure is 0.75.

We can also tip the balance away from crime by increasing the penalty for crime. In the third column of Table 13–3, the prison time for a failed criminal increases to 0.51. This change affects only the failed criminal, increasing the prison cost to $51 and decreasing the net income to $49. As a result, the utility for a failed criminal decreases to 7 utils. The expected utility of crime decreases to 9.5 utils:

$$EU\{12, 7; 0.50, 0.50\} = 0.50 \cdot 12 + 0.50 \cdot 7 = 9.5 \text{ utils}$$

The utility from crime is now less than the lawful utility, so the person will not commit crime. An increase in the severity of punishment reduces crime.

We can also tip the balance away from crime by decreasing the value of the loot. In the fourth column of Table 13–3, the loot is $21 instead of $44. The decrease in loot affects only the successful criminal: The net income drops to $121 and the utility drops to 11 utils. As a result, the expected utility from crime drops to 9.5 utils (the average of 11 utils and 8 utils), which is less than the lawful utility. Less loot means less crime.

Would a person with higher income be more or less inclined to commit crime? The fifth column of Table 13–3 shows the calculations for a person with four times as much income. The higher-income person gets the same loot, but has four times the opportunity cost of prison time. As a result, the lawful utility (20 utils) exceeds the expected utility from crime (18.5 utils), and the high-income person will not commit crime. Because the opportunity cost of crime increases with income while the benefits do not, we expect less crime among high-income people.

Morality and Anguish Costs

So far we have assumed that people do not consider the moral consequences of crime. In fact, most people have an aversion to committing antisocial acts, and they won't commit crime even if the expected payoff is positive. Of course, some people are less troubled by committing antisocial acts, and they are more likely to commit crime. We can incorporate morality by introducing an anguish cost, defined as the cost of committing an antisocial act. For example, suppose a person's anguish cost for a life of crime is 2 utils. In Table 13–3, the numbers for the expected utility of crime would drop by 2 utils.

We can use the sixth column of Table 13–3 to illustrate how anguish cost prevents crime. The probability of prison is relatively low (0.25), so the expected utility of crime before considering anguish cost (11 utils) exceeds the lawful utility (10 utils). If the anguish cost is 2 utils, however, the expected utility from crime would drop to 9 utils, below the lawful utility. Of course, this person would commit crime if the payoff were high enough to offset his 2-util anguish cost. For example, if the certainty or severity of punishment were low enough, the person would commit crime.

Incorporating morality and anguish costs helps explain why most people don't commit crime even when it appears that the payoff from crime is positive. In the sixth column of Table 13–3, before we include anguish cost, the expected utility of crime exceeds the lawful utility by 1 util. A person with an anguish cost of 2 utils will not commit crime, but a person with an anguish cost of only 0.50 utils will. In other words, morality explains why two people who face the same benefits and costs of crime may make different choices.

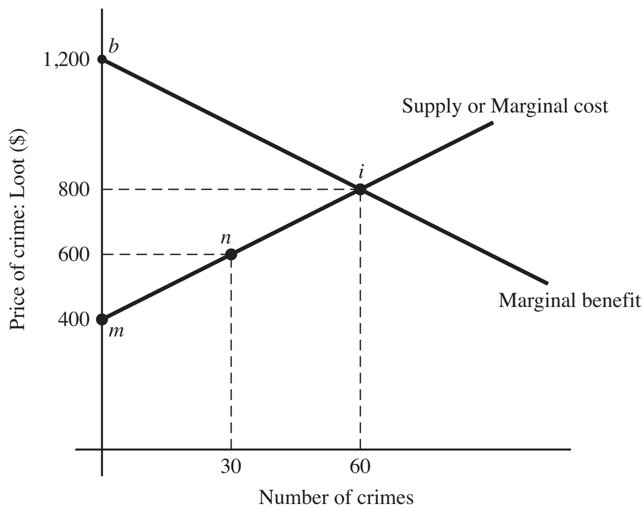
THE EQUILIBRIUM QUANTITY OF CRIME

We can use the insights from the model of the rational criminal to discuss the equilibrium quantity of crime. We'll take the perspective of criminals, using their benefit and cost curves to show how much crime rational criminals commit. As we'll see, we can use various crime policies to shift the benefit and cost curves and thus reduce crime. To simplify matters, we will switch from units of utility to dollars, allowing us to measure the costs and benefits of crime in monetary terms.

Drawing the Supply Curve

Like any other supply curve, the supply curve for crime shows the relationship between the price of crime and the number of crimes supplied (committed). The price is the benefit experienced by the criminal, equal to the loot or booty captured in the crime. The crime supply curve shows how the number of crimes committed increases with the loot. In Figure 13–3, the vertical intercept at point *m* indicates that the first crime is committed when the loot reaches $400. For a lower level of loot, say $399, no crimes would be committed because the benefit of a crime is less than the cost of committing it. As the value of loot increases, the quantity of crime supplied increases: If the loot is $600 per crime, 30 crimes will be committed; if the loot is $800, there will be 60 crimes.

FIGURE 13–3 Equilibrium Quantity of Crime



The supply curve shows the marginal cost of crime, which is positively sloped because potential criminals vary in their opportunity costs and anguish costs. The marginal benefit curve is negatively sloped because targets vary in their loot, and the most lucrative targets are victimized first. The equilibrium occurs at point *i*, where the marginal benefit equals the marginal cost.

A supply curve is also a marginal-cost curve, as explained in Section 2.2 in "Tools of Microeconomics," the appendix to the book. The crime supply curve shows, for each quantity of crime, the cost incurred for the marginal crime. For example, point *n* shows that with a loot of \$600, 30 burglaries are committed. If the value of loot were only \$599, the 30th burglary would not be committed, indicating that the benefit of the 30th burglary (\$599) is less than its cost. When the loot rises to \$600, the 30th crime is committed because now the benefit exceeds the cost. Therefore the cost of the 30th crime is just below \$600. Similarly, the 60th crime is committed when the loot rises to \$800, so the marginal cost of that crime is just below \$800.

We know from the model of the rational criminal that the cost of crime to the criminal is determined by four variables:

- The probability of being caught and imprisoned.
- The length of the prison term.
- The opportunity cost of time spent in prison, which varies with income.
- The anguish cost of crime, which varies across individuals.

Let's assume for the moment that all potential criminals face the same probability of prison and the same prison term.

The vertical intercept of the supply curve shows the cost of crime for the criminal with the lowest crime cost. Point *m* indicates that the lowest-cost criminal has a crime cost of \$400. This cost includes the opportunity cost of prison time and the anguish cost, both of which vary across potential criminals. If everyone faces the same probability of prison, the same prison term, and the same anguish cost, then the first crime will be committed by the person with the lowest opportunity cost—the lowest lawful income. Alternatively if everyone has the same lawful income, the first crime would be committed by the person with the lowest anguish cost. In general, criminals on the lower end of the supply curve are people with relatively low income and low anguish costs.

The supply curve is positively sloped because potential criminals vary in their opportunity costs and anguish costs. As we move upward along the supply curve, a larger loot induces people with higher opportunity and anguish costs to enter the crime market. For example, at point *n* a total of 30 crimes are committed by people whose cost of committing a crime is less than \$600. A bigger loot persuades people with higher opportunity cost and anguish cost to commit crime.

The Marginal-Benefit Curve and the Equilibrium Quantity of Crime

Figure 13–3 also shows the marginal benefit curve for crime from the criminal's perspective. The marginal benefit of a crime is the loot taken. The marginal-benefit curve is negatively sloped because crime targets vary in the amount of loot available and the difficulty in grabbing it. At the top of the marginal-benefit curve at point *b*, the most lucrative and easy target, with a loot of \$1,200, is targeted first, so the marginal benefit of the first crime is \$1,200. As we move downward along the marginal-benefit curve, criminals turn to progressively less lucrative targets, with less loot and greater difficulty in grabbing it.

Point *i* in Figure 13–3 shows the initial equilibrium in the market for crime. The equilibrium price (loot) is $800 per crime and 60 crimes are committed. For the first 60 crimes, the criminal's marginal benefit (the loot) exceeds his marginal cost, so the equilibrium quantity of crime is 60. Criminals stop at 60 crimes because for the 61st crime the marginal cost exceeds the loot, so additional crime does not pay.

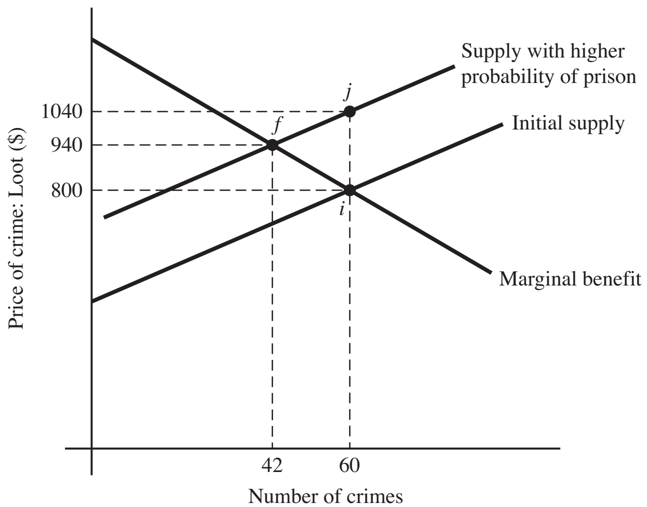
Increasing the Certainty of Punishment

We as a society can shift the crime supply curve by increasing the certainty of punishment. An increase in the probability of prison increases the cost of committing crime and shifts the supply curve upward. In Figure 13–4, the marginal-cost curve shifts upward by $240. For example, the cost of the 60th crime is now $1,040, up from $800. Of course, to increase the probability of imprisonment, we must use more resources (police and judges) to capture and convict criminals.

The upward shift of the supply curve decreases the equilibrium number of crimes. The new equilibrium is shown by point *f*: The equilibrium number of crimes drops from 60 to 42. This is a deterrent effect of the criminal-justice system: When the probability of prison increases, potential criminals respond to the higher cost by committing fewer crimes.

How responsive are criminals to increases in the certainty of punishment? The estimated elasticity of crime with respect to the probability of imprisonment is −0.30: A 10 percent increase in the probability decreases crime by about 3 percent. The elasticity of crime with respect to the arrest ratio (the number of arrests divided

FIGURE 13–4 Public Policy Shifts the Supply Curve and Decreases Crime



An increase in the probability of imprisonment increases the marginal cost of committing crime, shifting the supply curve upward by $240. The equilibrium quantity of crime decreases to 42 and the equilibrium price (loot) increases to $940.

by the number of crimes committed) is −0.30 as well. The elasticity of the crime rate with respect to the number of police officers is between −0.40 and −0.50. In general, there is convincing evidence that an increase in the certainty of punishment decreases crime.

Increasing the Severity of Punishment

We can also shift the supply curve by changing the length of the prison term. Like a higher probability of imprisonment, a longer prison term means a higher marginal cost of crime, resulting in an upward shift of the supply curve. If nothing else changed, we would expect the equilibrium crime rate to decrease.

Studies of criminal behavior suggest that longer prison sentences do not have much of an effect on crime rates. The estimated elasticity of crime with respect to the length of prison terms is close to zero. Longer prison terms cause other changes in the criminal environment that offset the higher costs associated with more time in prison:

1. **Hardening the criminal.** Criminals have a relatively low aversion to committing antisocial acts, and a longer prison term may reduce their aversion further, making crime more likely when they are released. A decrease in crime anguish costs shifts the supply curve downward, at least partly offsetting the deterrent effect of a longer prison term.
2. **Prison schooling.** If prison allows a criminal to learn the tricks of the trade from other criminals—or at least to learn from their mistakes—a longer prison term means a more skillful criminal. This leads to a lower probability of being caught, and thus a lower cost of committing crime. In other words, prison schooling shifts the supply curve downward, at least partly offsetting the deterrent effect of a longer prison term.

LEGAL OPPORTUNITIES AND EDUCATION

So far we have focused on the obvious strategies to reduce crime. Increasing the certainty or severity of punishment increases the cost of committing crime. In this part of the chapter, we'll look at more subtle strategies that reduce crime by increasing the value of lawful activities. As we saw in Table 13–3, an increase in wages increases the opportunity cost of crime. One way to increase wages is to increase educational attainment, especially the rate of high-school graduation.

Lawful Opportunities and Crime

An increase in the wage for lawful employment increases the opportunity cost of crime and decreases the supply of crime. In graphical terms, the effect of an increase in the wage is similar to the effect of increasing the probability of imprisonment, as shown in Figure 13–4: An increase in the wage from lawful activity shifts the crime supply curve upward, decreasing the equilibrium level of crime.

How responsive is the crime rate to the opportunities for lawful work? Consider first the connection between the unemployment rate and the crime rate. Although there is a positive relationship between unemployment and crime, the overall relationship is relatively weak. In contrast, the crime rates of first offenders are relatively sensitive to the unemployment rate. Specifically, teenage crime rates are lower in cities with more legal opportunities.

Consider next the connection between wages and crime. A recent study concludes that the elasticity of crime with respect to the wages of low-skilled workers is relatively large, between −1.0 and −2.0 (Gould, Weinberg, Mustard, 2002). In other words, a 10 percent increase in wages decreases crime by between 10 percent and 20 percent. Grogger (1991, 2000) shows that the wages of low-skilled workers and crime rates move in opposite directions. The recent trend of lower wages for low-skilled labor presents a policy challenge (Freeman, 1995):

> How to improve the job market for less skilled young American men, and reverse the huge decline in their earnings and employment opportunities is the problem of our times, with implications for both crime and many other social ills.

Education as Crime-Fighting Policy

Education reduces crime by increasing the opportunities for lawful work. College graduates earn almost twice as much as high-school graduates and high-school graduates earn almost 1.5 times as much as dropouts. So the link between education and crime is the graduation premium: Graduation increases wages, and higher wages decrease crime. Given the large graduation premium (50 percent) and the large elasticity of crime with respect to wages (−1.0 to −2.0), education policy has the potential to be a powerful anticrime policy.

A recent study suggests that investment in high-school education is an effective tool for reducing crime (Lochner and Moretti, 2004). The effects of high school education are measured in two ways:

1. **An additional year of high school.** Each additional year decreases the crime participation rate by about 0.10 percentage points for white males and by 0.40 percentage points for black males.
2. **Graduation.** High-school graduation decreases the crime participation rates of white males, with reductions of 9 percent for violent crime, 5 percent for drug crime, and 10 percent for property crime. The elasticity of arrest rates with respect to high-school graduation rates is −2.0 for violent crime and −1.3 for motor-vehicle theft.

The authors compute the benefits and costs of a small increase in the high-school graduation rate. Each year of schooling has a per-pupil cost of $6,000, so if getting a student from dropout status to graduation takes one more year of schooling, the additional cost per graduate is $6,000. Given the 50 percent graduation premium, the typical graduate benefits by earning $8,400 more per year for the rest of his working life. In addition, the resulting reduction in crime generates external benefits for the rest of society, about $1,600 per year for the rest of the graduate's working life. For a one-time expense of $6,000, society gets a crime-reduction benefit of $1,600 per year for 30 or 40 years.

APPLICATIONS: BIG-CITY CRIME AND THE CRIME DROP

We can use the insights from the model of the rational criminal to explain two observations. First, big cities experience higher crime rates than small cities. Second, during the 1990s, crime rates for both violent and property crime decreased by about a third.

Why Are Crime Rates Higher in Big Cities?

Crime rates increase with city size. Large cities (population at least 250,000) have twice as much violent crime per capita as small cities (population less than 10,000). For property crime, the big-city crime rate is about 30 percent higher. Overall, the elasticity of crime with respect to city size is 0.15: A 10 percent increase in population increases the crime rate by about 1.5 percent (Glaeser and Sacerdote, 1996).

Why are crime rates so much higher in big cities? Glaeser and Sacerdote provide three reasons:

1. **More loot (25 percent of difference).** Big cities have more lucrative targets: The average value per crime is about $900 in a city of 1 million, compared to a value of about $550 in a small city.
2. **Lower probability of arrest (15 percent of difference).** As shown in Table 13–4, bigger cities have lower arrest rates. Arrest rates are lower in big cities because (a) the pool of suspects is larger and (b) lawful citizens in impersonal big cities are less inclined to help their neighbors and the police in crime-control efforts.
3. **More female-headed households (50 percent of difference).** It's not clear why crime rates are higher in cities where a relatively large fraction of households are female headed. The authors speculate that children raised in single-parent families may have fewer job skills and less powerful ethical restraints on criminal behavior.

Since 1970, the correlation between city size and crime has weakened, and in recent years crime rates in big cities have dropped.

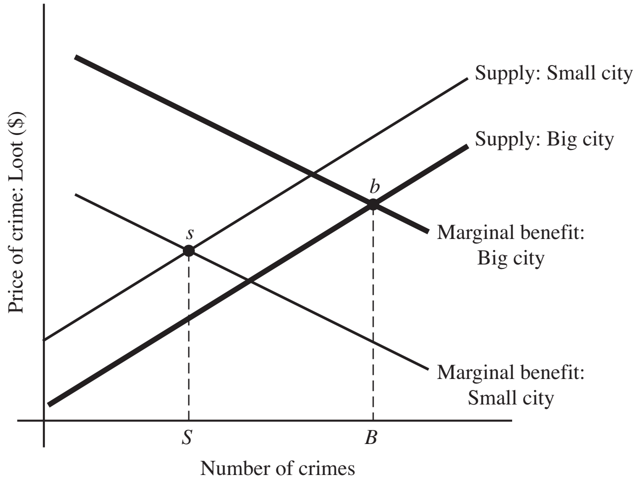
Figure 13–5 illustrates the reasons for higher crime rates in big cities. A big city has a lower arrest rate, so the cost of committing crime is lower. The lower cost generates a lower supply (marginal cost) curve. In addition, there is more loot in

TABLE 13–4 Arrest Rates and City Size

Population (1,000)	25–50	50–100	100–250	250–500	500–1 Million	More than 1 Million
Arrest rate (%)	12	11	11	10	8	7

Source: Edward L. Glaeser, and B. Sacerdote. "Why Is There More Crime in Cities?" NBER Working Paper #5430, 1996.

FIGURE 13–5 More Crime in Big Cities



A big city has more loot, so the marginal benefit curve is higher. A big city has a lower probability of arrest, so the marginal cost (supply curve) is lower. The equilibrium in the big city (point *b*, with *B* crimes) generates a higher crime rate than a small city (point *s*, with *S* crimes).

big cities, so the marginal benefit curve is higher. The equilibrium in the big city is shown by point *b*, and the equilibrium in the small city is shown by point *s*. The combination of lower marginal cost and higher marginal benefit leads to a higher crime rate in the big city.

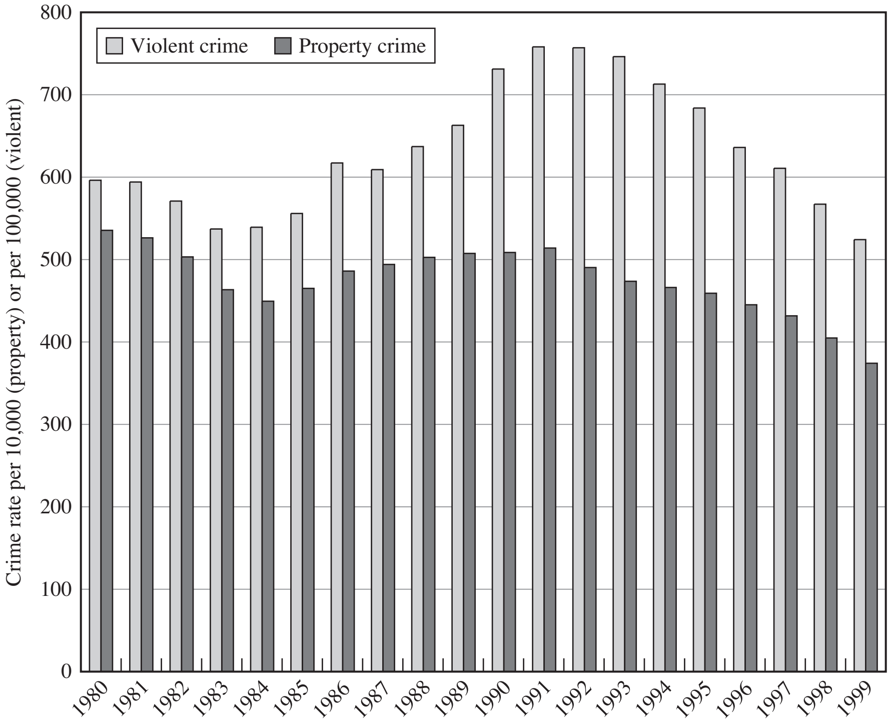
Why Did Crime Rates Decrease in the 1990s?

During the 1990s, crime rates for both violent and property crime decreased by about a third. As shown in Figure 13–6 (page 353), both violent crime and property crime peaked in 1991 and decreased steadily through the rest of the decade.

A recent study explores the factors that reduced crime during the 1990s (Levitt, 2004), and Figure 13–7 (page 354) summarizes the conclusions:

- **Strong economy.** There were more jobs and higher wages, causing a 2 percent reduction in property crime.
- **Demographics.** A decrease in the share of the population in the crime-prone years of 16–24 decreased violent crime by 2 percent and property crime by 5 percent.
- **Police techniques.** A number of innovative police policies, including community policing and more aggressive control of public nuisances, reduced crime by a relatively small amount.
- **Increase in police.** Over the decade, the number of police officers per capita increased by 14 percent (at a cost of $8.4 billion per year), decreasing crime by 5.5 percent.

FIGURE 13–6 Crime Rates, 1980–1999

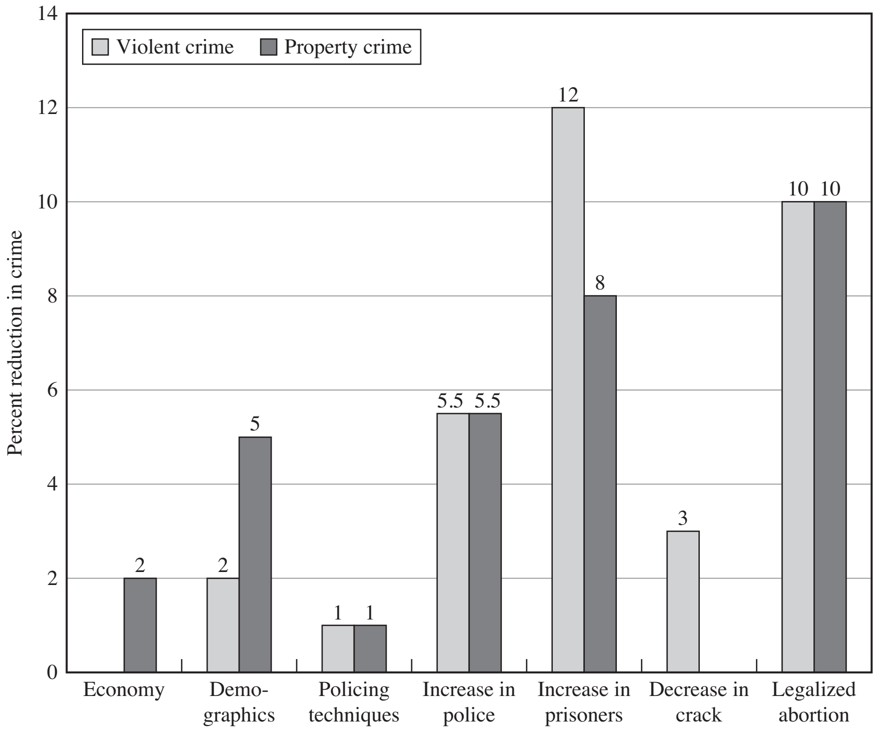


Source: FBI Uniform Crime Reports.

- **Increase in prisoners.** The national prison population doubled over this period, decreasing violent crime by 12 percent and property crime by 8 percent.
- **Decrease in crack cocaine sales.** In the 1980s, there was lucrative trade in crack cocaine, and rival drug sellers battled over market areas in central cities. These turf battles generated a lot of violent crime in cities. As crack cocaine sales dropped over the 1990s, so did the violent turf battles, reducing violent crime by about 3 percent.

A subtle and surprising factor in lower crime is legalized abortion. The legalization of abortion in 1974 decreased the number of unwanted births. There is evidence that crime rates are higher among children born to reluctant parents. The wider availability of abortion starting in the 1970s reduced the number of children born in such circumstances, and thus decreased the number of crime-prone people maturing in the 1990s. As shown in Figure 13–7, Levitt concludes that the legalization of abortion cut crime rates by 10 percent. A number of other studies have considered the connection between abortion and crime, with mixed results (Sen, 2007). Some provide evidence that support the conclusions of Levitt, while others suggest that the connection is much weaker.

FIGURE 13–7 Why Did Crime Drop in the 1990s?



Source: Data from Steven Levitt, "Understanding Why Crime Fell in the 1990s: Four Factors That Explain the Decline, and Six That Do Not." *Journal of Economic Perspectives* 18 (2004), pp. 163–190.

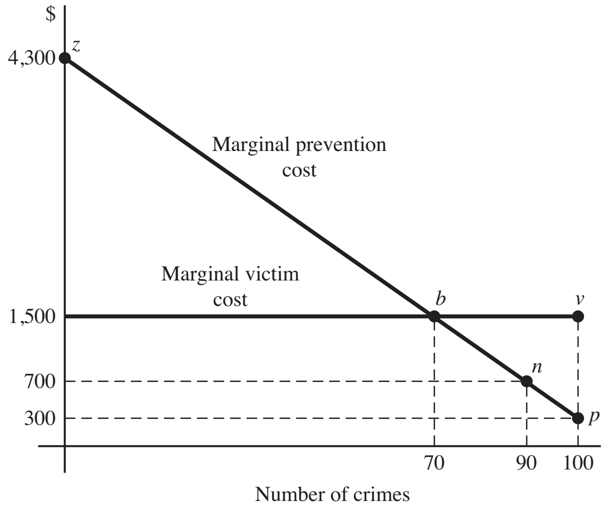
HOW MUCH CRIME?

A society can use its resources—labor, capital, and land—in different ways to promote economic well-being. The problem of crime presents society with some stark choices about how much crime to experience. It would be possible, in principle, to cut the crime rate to one-tenth or one-hundredth of its current level. The question is whether such a dramatic reduction in crime would be socially efficient. As we'll see, the reason we tolerate so much crime is that some crimes are more costly to prevent than to experience.

The Optimal Amount of Crime

Consider a society's choice of how much burglary to allow. As we saw earlier in the chapter, the public sector can decrease crime by using resources to increase the certainty and severity of punishment. In addition, potential victims can deter crime by investing in security measures such as locks, guards, and alarms. We

FIGURE 13–8 The Socially Efficient Amount of Crime



The socially efficient quantity of crime is where the marginal victim cost equals the marginal prevention cost, shown by point *b* with 70 crimes. Going beyond that point, crimes are more costly to prevent than to experience.

can combine these two sorts of prevention efforts into a single measure of crime-prevention costs.

We will use the marginal principle to determine the optimum level of crime. The marginal principle is reviewed in Section 1.1 of "Tools of Microeconomics," the appendix at the end of the book. In Figure 13–8, the horizontal axis measures the number of crimes committed, which ranges from zero to 100. The negatively sloped curve is the marginal cost of crime prevention. Starting at point *p*, if we don't allocate any resources to crime prevention, there will be 100 crimes. The marginal-cost curve indicates that the cost of preventing a single crime (reducing crime from 100 to 99) is $300. As we prevent more and more crimes, the marginal cost of prevention increases, and we move upward along the marginal prevention cost curve. For example, the marginal cost of preventing crime number 90 is $700 (shown by point *n*), and the marginal cost continues to increase, reaching $4,300 to prevent the last crime (at point *z*). In other words, it is relatively easy to go from 100 to 90 crimes, but the lower the crime rate, the more costly it is to prevent another crime.

The other cost of crime is experienced by the victim. As we saw earlier in the book, victim costs include the opportunity cost of lost work time, monetary losses, and the costs of injuries. Miller, Cohen, and Wiersema (1996) estimate the following costs for different crimes: $370 for larceny, $1,500 for burglary, $4,000 for auto theft, $13,000 for armed robbery, and $15,000 for assault. In Figure 13–8, we assume that each burglary imposes a cost on society of $1,500, so the marginal victim cost is constant at $1,500.

The socially efficient quantity of crime minimizes the sum of prevention and victim costs. The total cost is minimized where the two marginal-cost curves intersect, which happens at point *b*, with 70 burglaries. If we start with 100 burglaries (point *p*), the 100th burglary has a victim cost of $1,500 but costs only $300 to prevent. We can spend $300 to save $1,500, so the total cost of crime decreases by $1,200. Moving upward along the prevention-cost curve, we see that the marginal prevention cost is less than the marginal victim cost down to a crime rate of 70, so that's the place to stop. If we were to move beyond point *b* to fewer crimes, the cost of preventing each crime would exceed the cost of experiencing it, so we would be better off at point *b*.

Differences in victim costs generate differences in the socially efficient level of crime. As we saw earlier in the chapter, robbery has a higher victim cost. If it has the same marginal prevention cost as burglary, the socially efficient number of robberies will be smaller. This means that it is sensible for the government to use more resources in the prevention of more serious crime, in part by imposing longer prison terms. So there is some economic logic behind the notion of making the punishment fit the crime.

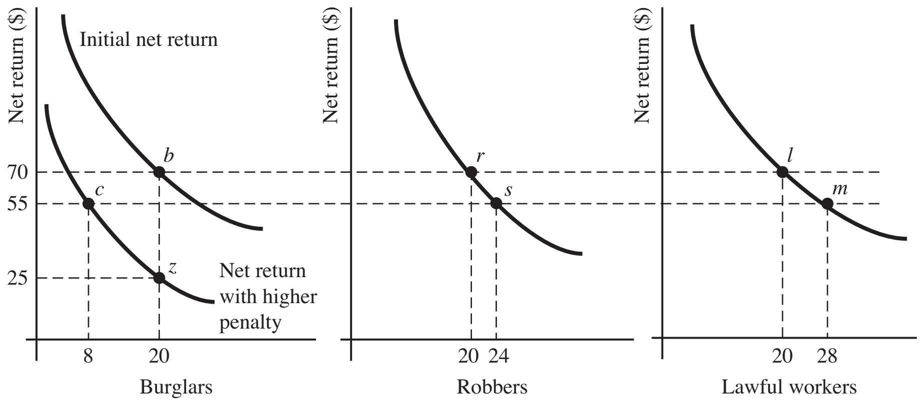
Crime Substitution and the Principle of Marginal Deterrence

Criminals have options too. We, as a society, pick a set of crime penalties, one for each type of crime (e.g., one year for burglary, three years for armed robbery), providing criminals with a menu of crime penalties. Criminals respond by picking the most lucrative crime, given the penalty menu and the payoffs from different crimes. This has important implication for policies that "get tough" on one crime or another. For example, if we triple the prison term for burglary to make it the same as the penalty for robbery, how would that affect the number of burglaries and armed robberies?

Figure 13–9 shows the implications of a longer prison term for burglary. A group of 60 people choose between burglary, armed robbery, and a lawful job. The net return from an activity equals the expected benefit (loot or wages) minus the expected cost, which includes the expected penalty for crime. In equilibrium, the net returns of the three alternative activities will be equal, making the marginal person indifferent among the three activities. For each of the three options, the net return curve is negatively sloped, indicating that the larger the number of people in an activity, the lower the net return. The initial equilibrium is shown by points *b* (burglary), *r* (robbery), and *l* (lawful), with 20 people in each activity. The net return for each activity is $70 per day.

Suppose we increase the penalty for burglary. In the left panel of Figure 13–9, the increase in the expected cost of crime shifts the net-return curve for burglary downward. If the number of burglars remained at 20, the net return would drop to $25 (shown by point *z*). The net return is now higher in the other activities,

FIGURE 13–9 Equalizing Penalties and Crime Substitution



In equilibrium, the net return of burglary, robbery, and lawful work are equal. An increase in the penalty for burglary shifts the burglary return curve downward. Equilibrium is restored with a lower common net return ($55, down from $70), fewer burglars (8 instead of 20), more robbers (24 instead of 20) and more lawful workers (28 instead of 20).

so people will switch to robbery and lawful work. As they do, we move upward along the new curve for burglary (from point *z* to point *c*), downward along the curve for robbery (from *r* to *s*) and downward along the curve for lawful work (from *l* to *m*).

In the new equilibrium, there are fewer burglars, but more robbers and lawful workers. The equilibrium is shown by points *c*, *s*, and *m*: The net return of each activity is $55, and the number of people in the three activities adds up to 60. Of the 12 people who stop committing burglary, four switch to robbery and eight choose lawful work.

How does the larger penalty for burglary affect the victim cost of crime? It depends on the change in the mix of crime and the victim cost of each crime. The victim cost of armed robbery is almost nine times the victim cost of burglary, so we would break even in terms of total victim cost if, for each additional robbery, the number of burglaries decreased by nine. In our simple example, there are four additional robberies and only 12 fewer burglaries, so the total victim cost increases. Of course, this is just an example; if the increase in robberies were smaller and the reduction in burglaries were larger, the total victim cost would decrease.

The principle of marginal deterrence is that penalties should increase with the victim cost of crime. For example, the penalty for burglary should be less than the penalty for armed robbery, with almost nine times the victim cost. Although equalizing penalties for the two crimes would decrease burglary, it would also cause some criminals to switch to more costly armed robberies. The challenge for crime policy is to develop a penalty menu that generates the mix of crimes that minimizes the social cost of crime.

THE ROLE OF PRISONS

Earlier in the chapter we saw that putting people in prison decreases crime rates. The elasticity of crime with respect to the prison population is about −0.25 for property crime and about −0.40 for violent crime. In this part of the chapter, we explore why incarceration reduces crime. Specifically, we look at three ways that prisons decrease crime:

- **Deterrence.** The threat of being locked up in prison persuades some people not to commit crime.
- **Incapacitation.** Isolating criminals from potential victims prevents crime.
- **Rehabilitation.** Prisons may improve the attitudes or skill level of convicts, making them less likely to commit crimes after they leave.

We discussed deterrence earlier in the chapter. Recall that an increase in the certainty of punishment is more effective than an increase in severity. According to a recent study (Levitt, 1998), each burglary arrest deters 2.3 burglaries and each arrest for car theft deters 0.50 car thefts.

Incapacitation

The second function of the prison system is to take criminals out of circulation. The benefit of keeping a criminal locked up equals the number of crimes prevented times the social cost per crime. Studies of the incapacitation effects generate mixed results. DiIluio (1996) suggests that incapacitating the typical criminal prevents between 17 and 21 crimes per year, while Levitt (1998) suggests that the number of crimes prevented is much smaller.

A recent study of the prison system in Texas quantifies the trade-offs with prisons (Spelman, 2005). The author computes the benefit of incapacitation as the victim cost avoided by holding an inmate in prison. In 2000, the marginal benefit of incapacitation was $15,000 per inmate. The authors estimate that the marginal cost of holding an inmate in prison is about $36,000, a figure that includes the cost of building and operating prisons as well as the opportunity cost of having a person in prison rather than working. These numbers indicate that the marginal cost of incarceration exceeds the marginal benefit, meaning that the number of inmates exceeds the socially efficient number.

The author of the Texas study notes that the marginal benefit of $15,000 excludes some potential benefits from incarceration. The figure excludes the benefits associated with a reduced fear of crime or the savings on protective measures such as locks and guard dogs. If these nonvictim costs are high enough, the marginal benefit of incarceration could be greater than or equal to the marginal cost. In that case, the current level of incarceration in Texas could be justified on efficiency grounds.

Rehabilitation

The third function of prisons is to rehabilitate criminals by providing them with the skills and attitudes required for success in lawful employment after their release. About one-third of inmates participate in education and vocational training programs, and about one-third participate in drug and alcohol programs. If we add participation in programs that provide training in life skills, about two-thirds of inmates participate in some type of rehabilitation program.

The simple facts on rehabilitation are not encouraging. Roughly two-thirds of former inmates are rearrested within three years of release, and roughly half return to prison within three years. Released inmates account for between 10 percent and 12 percent of property and violent crimes. Dozens of studies have measured the effects of rehabilitation programs for adults, and the consensus is that they are ineffective for three reasons.

1. It is difficult to change the antisocial attitudes that make certain people receptive to crime.
2. By the time the typical criminal reaches the prison rehabilitation stage, he or she has committed dozens of crimes, and is entrenched in the criminal world.
3. It is difficult to make legal opportunities more profitable than crime: Crime pays, and it is difficult to increase the job skills of an adult criminal.

There is evidence that anticrime programs targeted at youths pass the benefit-cost test. The high cost of crime means that even a relatively expensive program that reduces crime rates by a small amount will pass the test. The average juvenile-delinquent program reduces crime rates by a small amount (Lipsey, 1992). Some early-intervention programs help reduce crime among youths, again by modest amounts (Mendel, 1995).

SUMMARY

The model of the rational criminal suggests that like other people, criminals respond to incentives. Here are the main points of the chapter:

1. Crime is risky because there is a chance of being caught and paying a large penalty. A potential criminal compares the certain utility of lawful work to the expected utility of criminal activity.
2. An increase in the probability of punishment has a larger deterrent effect than an increase in the severity of punishment.
3. The optimum amount of crime is the level at which the marginal victim cost equals the marginal prevention cost.
4. Education reduces crime by increasing the opportunities for lawful work. High-school graduation decreases crime participation rates by 9–10 percent.
5. The dramatic reductions in crime during the 1990s resulted from more police and prisoners, more favorable demographics (caused in part by legalized abortion), a decline in crack cocaine, and a stronger economy.

APPLYING THE CONCEPTS

For exercises that have blanks (_____), fill each blank with a single word or number. For exercises with ellipses (. . .), complete the statement with as many words as necessary. For exercises with words in square brackets ([increase, decrease]), circle one of the words.

1. **Women and Crime**
   During the 1970s, the crime rate for women increased five times faster than the crime rate for men. Make a list of the possible reasons for the more rapid increase for women. Then check the conclusions of the article "Women and Crime" by Ann Bartel (see the list of references at the end of the chapter). Do the key factors identified by Bartel appear on your list?
2. **Expected Utility Numbers**
   Consider a potential criminal with lawful income of $100, potential loot of $125, a probability of imprisonment of 0.50, and a prison term of 0.51 units of time. Fill the blanks in the following table.
   a. The expected utility from crime (equal to _____) is [greater, less] than the lawful utility (equal to _____).
   b. When we move from Table 13–3 to this example, the loot [increases, decreases] while the prison time [increases, decreases], and the expected utility of crime [increases, decreases] because the change in _____ is large relative to the change in _____.

Lawful income ($)	100
Loot ($)	125
Probability of prison	0.50
Prison time	0.51
Lawful utility	_____
Utility from successful crime	
Net income	_____
Utility	_____
Utility from failed crime	
Prison cost	_____
Net income	_____
Utility	_____
Expected utility from crime (utils)	_____

3. **Does a More Severe Penalty Work?**
   Using the first column of Table 13–3 as a starting point, modify the numbers to make the loot $156 and the anguish cost from crime 1 util.
   a. The payoff from crime (the expected utility minus anguish cost) is _____ utils and the payoff (utility) from lawful activity is _____ utils.
   b. Suppose the length of the prison term increases to 0.91 units of time, and nothing else changes from part (*a*). The payoff from crime [increases, decreases] to _____ utils.

c. Suppose that extra schooling from the longer prison term increases criminal productivity, decreasing the probability of prison to 0.40. No other numbers change from part (*b*). The payoff to crime [increases, decreases] to _____ utils.

d. Suppose that a longer prison term reduces anguish cost to zero, and no other numbers change from part (*c*). The payoff from crime [increases, decreases] to _____ utils and is [less, greater] than the payoff from lawful activity.

4. **Budget-Balancing Prison Change**
Consider a state with a fixed prison capacity of 1,000. When a law is passed establishing a minimum term of 10 years, (twice the current average term), the state cuts the number of prisoners in half, cutting the probability of imprisonment for crime in half. Use a graph like Figure 13–3 to illustrate the effects of this law.
   a. Doubling the average prison term shifts the marginal [benefit, cost] curve [up, down] because . . .
   b. Cutting the probability of imprisonment shifts the marginal [benefit, cost] curve [up, down] because . . .
   c. The combination of doubling the term and cutting the probability shifts the marginal [benefit, cost] curve [up, down] because . . .
   d. The equilibrium number of crimes [increases, decreases, doesn't change] because . . .
5. **Big City Crime**
The chapter lists the three reasons for higher crime rates in large cities. For each reason below, redraw Figure 13–3, assuming all other conditions remain constant.
   a. More loot.
   b. Lower probability of arrest.
   c. More female-headed households.
6. **Unhappy Birthday?**
In general, participation in crimes such as robbery and burglary increases with age until about age 17, then declines. Consider two states, A and J, that have the same crime penalties for adults (18 years or older), but J has lighter penalties for juveniles (16–18 years). For each state, draw a curve that shows the crime participation rate as a function of age, for ages 16–24. Explain any differences between the two curves.
7. **Victim Cost and Optimum Crime Rate**
Using Figure 13–8 as a starting point, suppose the marginal victim cost increases to $1,900, while the marginal prevention cost curve is unchanged.
   a. Illustrate the change with a graph similar to Figure 13–8.
   b. The socially efficient crime rate decreases to _____ crimes because . . .
8. **More Crime in Low-Income Neighborhood**
Consider a city with two neighborhoods with different crime rates. The burglary rate in the low-income neighborhood is 15, compared to 5 in the high-income

neighborhood. The two neighborhoods have the same victim costs per burglary ($1,500). Suppose this outcome is socially efficient.

a. A higher crime rate in the low-income neighborhood is socially efficient if the low-income neighborhood . . .

b. Illustrate your answer to (*a*) with a graph.

c. One possible reason for the difference noted in (*a*) is that low-income households . . .

9. Marginal Deterrence: Unchanged Social Cost

Using the initial equilibrium in Figure 13–9 as a starting point (points *b*, *r*, and *l*), suppose an increase in the penalty for burglary decreases the equilibrium number of burglaries from 20 to 2. The total social costs of crime (for burglary and robbery) will be unchanged if the number of robberies [increases, decreases] from 20 to _____ because . . .

10. Decrease in the Lawful Return

Consider the example shown in Figure 13–9. The initial net return is $70 in each activity, and there are 20 people in each activity. Suppose that instead of increasing the burglary penalty, we decrease the net return to lawful activity by half. This change is represented by a downward shift by $35 of the net-return curve for lawful activity.

a. At the original allocation of 20 people in each sector, people will switch from _____ to _____ and _____ because . . .

b. The switching noted in part (*a*) [increases, decreases] the net return to lawful activity and [increases, decreases] the net return to crime.

c. In the new equilibrium, the net return is [greater, less] than $70, the number of lawful workers is [greater, less] than 20, and the number of burglars is [greater, less] than 20.

d. Draw a trio of curves like the ones shown in Figure 13–9 to illustrate the answers to (*a*), (*b*), and (*c*).

e. In the new equilibrium, the sum of burglars, robbers, and lawful workers is _____, and the net return to burglary is [greater than, less than, equal to] the net return to lawful workers.

REFERENCES AND ADDITIONAL READING

1. Bartel, Ann P. "Women and Crime: An Economic Analysis." *Economic Inquiry* 42 (1979), pp. 29–51.
2. Blumstein, Alfred, and Joel Wallman. *The Crime Drop in America*. New York: Cambridge University Press, 2000.
3. DiIulio, John J., Jr. "Help Wanted: Economists, Crime, and Public Policy." *Journal of Economic Perspectives* 10 (1996), pp. 3–24.
4. Freeman, Richard B. "The Labor Market." In *Crime,* eds. J. Q. Wilson and Joan Petersilia. San Francisco: Institute for Contemporary Policies, 1995.
5. Freeman, Richard B., and Harry J. Holzer. "The Black Youth Employment Crisis: Summary of Findings." In *The Black Youth Employment Crisis,* ed.

Richard B. Freeman and Harry J. Holzer. Chicago: University of Chicago Press, 1986.
6. Glaeser, Edward L., and B. Sacerdote. "Why Is There More Crime in Cities?" NBER Working Paper no. 5430, 1996.
7. Gould, Eric D, Bruce A. Weinberg, and David B. Mustard. "Crime Rates and Local Labor Market Opportunities in the United States: 1979–1997." *Review of Economics and Statistics* 84 (2002), pp. 45–61.
8. Grogger, Jeffrey. "An Economic Model of Recent Trends in Violence." Chapter 8 in *The Crime Drop in America,* eds. Alfred Blumstein and Joel Wallman. New York: Cambridge University Press, 2000.
9. Grogger, Jeffrey. "Certainty vs. Severity of Punishment." *Economic Inquiry* 29 (1991), pp. 297–309.
10. Levitt, Steven D. "Juvenile Crime and Punishment." *Journal of Political Economy* 106 (1998), pp. 1156–85.
11. Levitt, Steven D. "The Effect of Prison Population Size on Crime Rates: Evidence from Prison Overcrowding Litigation." *Quarterly Journal of Economics* 111 (1996), pp. 319–52.
12. Levitt, Steven D. "Understanding Why Crime Fell in the 1990s: Four Factors That Explain the Decline, and Six That Do Not." *Journal of Economic Perspectives* 18 (2004), pp. 163–190.
13. Levitt, Steven D. "Why Do Increased Arrest Rates Appear to Reduce Crime: Deterrence, Incapacitation, or Measurement Error?" *Economic Inquiry* 36 (1998), pp. 353–72.
14. Lipsey, Mark. "Juvenile Delinquency Treatment: A Meta-Analysis Inquiry into the Variability of Effects." In *Meta-Analysis for Explanation: A Casebook,* ed. Thomas D. Cook et al. Beverly Hills, CA: Sage Foundation, 1992.
15. Lochner, Lance, and Enrico Moretti. *American Economic Review* 94 (2004), pp. 155–89.
16. Marvel, Thomas B., and Carlisle E. Moody. "Prison Population Growth and Crime Reduction." *Journal of Quantitative Criminology* 10 (1994), pp. 109–40.
17. Mendel, Richard A. "Prevention or Pork? A Hard-Headed Look at Youth-Oriented Anti-Crime Programs." Washington DC: American Youth Policy Forum, 1995.
18. Miller, Ted, Mark Cohen, and Brian Wiersema, "Victim Costs and Consequences: A New Look" (Washington DC: National Institute of Justice, 1996)
19. Raphael, Steven, and Michael Stoll. "The Effects of Prison Releases on Regional Crime Rate." Brookings-Wharton Papers on Urban Affairs 2004 (2004), pp. 207–243.
20. Sen, Anindya. "Does Abortion Lead to Lower Crime? Evaluating the Relationship between Crime, Abortion, and Fertility." *The B.E. Journal of Economic Analysis and Policy* 7, no. 1, Article 48 (2007).
21. Spelman, William. "Jobs or Jails? The Crime Drop in Texas." *Journal of Policy Analysis and Management* 24 (2005), pp. 133–65.

22. Tauchen, Helen, Ann Dryden Witte, and Harriet Griesinger. "Criminal Deterrence: Revisiting the Issue with a Birth Cohort." *Review of Economics and Statistics* 76 (1994), pp. 399–412.
23. Thaler, Richard. "Econometric Analysis of Property Crime—Interaction between Police and Criminals." *Journal of Public Economics* 8 (1977), pp. 37–51.
24. Wilson, James Q. *Crime and Public Policy.* San Francisco: Institute for Contemporary Studies, 1983.
25. Wilson, James Q. *Thinking about Crime*. New York: Basic Books, 1975.
26. Witte, Ann D. "Estimating the Economic Model of Crime with Individual Data." *Quarterly Journal of Economics* 94 (1980), pp. 57–84.

PART FIVE

Housing

This part of the book explores the economics of the urban housing market and evaluates the merits of various housing policies. Chapter 14 explains why housing is different from other products: Housing is heterogeneous (dwellings differ in size, age, design, and location) and durable, and moving from one house to another is very costly. The filtering model of the housing market explains the economic forces that cause dwellings to move down the quality ladder to households with progressively lower incomes. As we'll see in Chapter 15, the federal government spends about $30 billion per year to assist low-income households, with the money spent on public housing, subsidized private housing, and vouchers issued to low-income households. In addition, the federal government sacrifices about $66 billion in tax revenue per year to subsidize mortgage interest, and most of the benefits go to high-income households.

CHAPTER 14

Why Is Housing Different?

The fellow that owns his own house is always just coming out of a hardware store.

—KEN HUBBARD

Last week I helped my friend stay put. It's a lot easier than helping someone move. I just went over to his house and made sure that he did not start to load his stuff into a truck.

—MITCH HEDBERG

Housing has three features that make it different from other products. First, the housing stock is heterogeneous, with dwellings that differ in size, age, style, interior features, utilities, and location. Second, housing is durable and can deteriorate over time at a fast rate or a slow one, depending on the maintenance and repair decisions of its owner. Third, moving is costly, so when income or housing preferences change, consumers don't instantly adjust their housing consumption. Instead, they wait until the gap between the ideal house and their actual house is large enough to justify the large cost of moving. In this chapter, we explore the implications of these unusual features of the housing market.

HETEROGENEITY AND HEDONICS

The housing stock is heterogeneous, with each dwelling offering a different bundle of housing services. Dwellings differ in size, layout, style, utilities (heating and electrical), and the quality of the interior and the exterior. As we saw in the chapter on neighborhood choice, when you choose an apartment or house, you also choose a neighborhood, with its own bundle of housing services. Neighborhoods differ in accessibility to jobs and social opportunities, local public goods and taxes, and environmental quality.

What determines the equilibrium price of a dwelling? Under the hedonic approach, we determine the price of each part of the housing bundle. A hedonic study of the market might generate the following information:

1. **Base price.** The average house, with a price of $200,000, has three bedrooms, is five miles from the city center, and its roof is six years old.
2. **Access price.** The price drops by $2,000 for every additional mile from the city center.
3. **Size.** The price increases by $30,000 for every additional bedroom.
4. **Roof age.** The price of housing decreases by $500 for every year of roof age.
5. **Air quality.** The price decreases by $1,000 for every additional unit of air pollution.
6. **Schools.** The price increases by $2,000 for every one-unit increase in the average test score of students in the local elementary school.

To predict the price of a particular dwelling, we add to the base price to reflect differences between the average dwelling and a particular dwelling. For example, a fourth bedroom adds $30,000 to the price, while a new roof adds another $3,000 and a four-unit difference in air pollution adds $4,000. If the average test scores in the local school are three points higher than the city average, that adds another $6,000. Adding up these adjustments, the predicted price of the dwelling is $243,000.

The classic hedonic study is by Kain and Quigley (1970). They used data from the St. Louis housing market in the 1960s to estimate the dollar values of different housing attributes. Figure 14–1 shows the percentage changes in monthly rent and house value for one-unit changes in each of various housing features. For example, a one-unit increase in interior quality increases monthly rent by 2.1 percent and market value by 5.6 percent. The exterior quality of nearby dwellings was measured on a scale of 1 (bad) to 5 (excellent). A one-unit increase in the quality of adjacent dwellings increased rent by 3 percent and market value by 5.3 percent. A one-unit increase in the quality of dwellings on the block increased rent by 6 percent and market value by 2.9 percent.

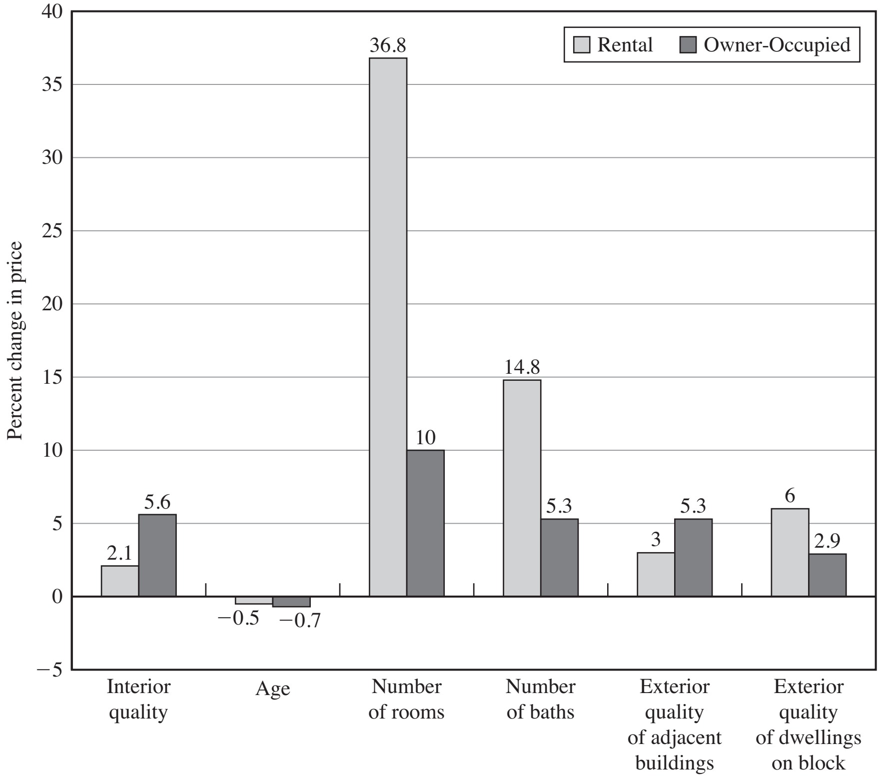
Other hedonic studies have explored the effects of amenities on housing prices. Among the neighborhood characteristics with positive effects on housing prices are proximity to jobs, high-performing schools, transit stations, and churches. In contrast, property values are lower in neighborhoods close to areas with high crime rates, toxic waste facilities, and noisy highways.

How does a household choose among alternative dwellings, each of which provides a different bundle of characteristics? Most consumers do not have access to a hedonic study of their housing market. As they shop, they gather their own information about the implicit prices of location, size, and design features. The household then chooses the best affordable bundle.

DURABILITY, DETERIORATION, AND MAINTENANCE

Housing is durable in the sense that with proper maintenance, a dwelling can provide housing services for 100 years or more. But in the absence of routine maintenance and repair, a dwelling deteriorates, and the quality of the dwelling decreases

FIGURE 14–1 Results of Hedonic Study



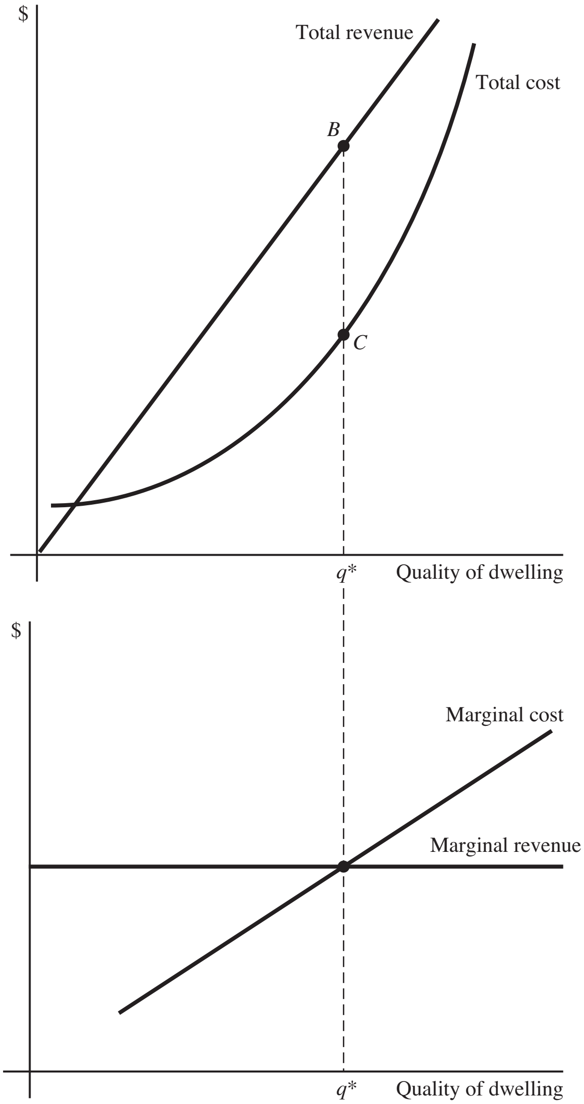
Source: JF Kain and JM Quigley, *Housing Markets and Racial Discrimination: A Microeconomic Analysis* (New York: NBER, 1975). Table 8.2, p. 200.

over time. Imagine a quality ladder for housing, with the highest quality dwellings at the top of the ladder and progressively lower quality as we move down the ladder. Each year, a property owner must decide where on the quality ladder to position his or her dwelling. If the owner does nothing, the dwelling will drop one or more rungs on the ladder. With a moderate expenditure on maintenance and repair, the owner can keep the dwelling at the same level. To raise the quality of the dwelling, the owner must spend a substantial amount of money to renovate or remodel.

Picking the Quality Level

We can use Figure 14–2 (page 370) to explore a property owner's decision about where on the quality ladder to position an existing dwelling. The horizontal axis measures the quality level, a general representation of the quality of housing services generated by a dwelling. As shown in the upper panel, the cost of managing and maintaining a dwelling increases with its quality. This is sensible because maintenance and repair costs are required to offset physical deterioration. The higher the

FIGURE 14–2 Picking a Position on the Quality Ladder



The owner's objective is to maximize profit, the gap between total revenue and total cost. Profit is maximized at the quality level where the marginal benefit (the price of quality) equals the marginal cost: q^*.

quality, the greater the expense required to keep the dwelling at that quality level. The total-cost curve is convex, reflecting diminishing returns to maintenance: As the quality increases, it becomes progressively more costly to maintain the dwelling at the given quality level.

Consumers are willing to pay more for higher quality dwellings. For a rental dwelling, the price of quality is the change in the monthly rent for a one-unit increase in quality. In the upper panel of Figure 14–2, the linear curve shows the total revenue (monthly rent) on a dwelling as a function of the quality level. The total-revenue curve is linear, reflecting the assumption that consumers' willingness to pay for a dwelling increases linearly with quality. For example, doubling the quality doubles a consumer's willingness to pay for a dwelling, allowing the owner to double the rent.

The owner's objective is to maximize profit, equal to the gap between total revenue and total cost. In the upper panel of Figure 14–2, the gap is maximized at a quality level q^*. The lower panel uses the marginal principle, reviewed in Section 1.1 of "Tools of Microeconomics," the appendix at the end of the book. The profit-maximizing quality is where the marginal benefit of quality equals the marginal cost. The marginal benefit is the marginal revenue, the change in monthly rent per unit change in quality (the slope of the total-revenue curve, the price of quality), and the marginal cost is the slope of the total-cost curve. To maximize profit (the gap between total revenue and total cost), the owner picks the quality where marginal revenue equals marginal cost.

Changes in Quality and Retirement

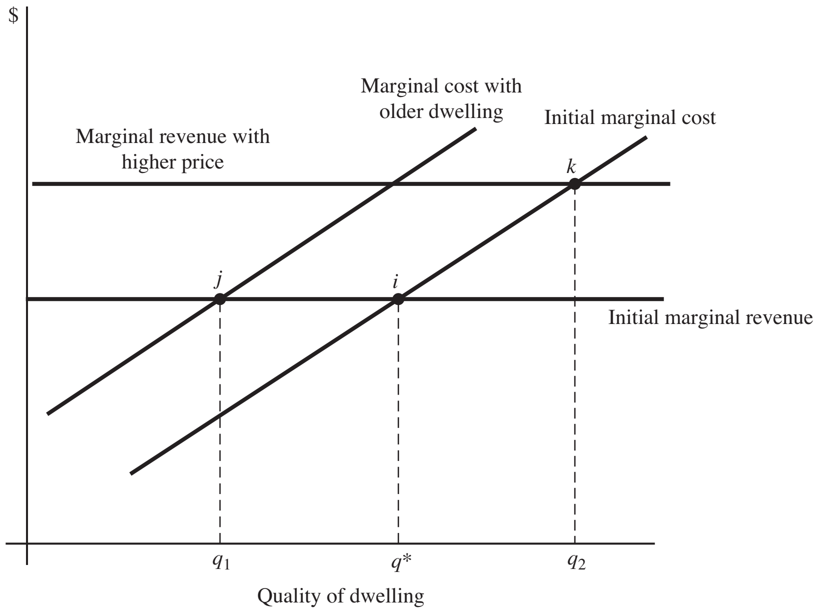
Figure 14–2 shows the quality choice with a particular set of revenue and cost curves. The profit-maximizing quality level is affected by changes in revenue or costs. As a dwelling ages, the cost of maintaining a given quality level increases, shifting the marginal-cost curve upward, as shown in Figure 14–3 (page 372). As a result, if the price hasn't changed (the marginal-revenue curve hasn't shifted), the profit-maximizing quality decreases. In Figure 14–3, this is shown as a move from point *i* to point *j*. If the marginal cost continues to increase as the dwelling ages, the owner's chosen quality level will continue to decrease. Eventually, the marginal-cost curve will lie entirely above the marginal-revenue curve. At that point, the cost of keeping the dwelling on the market exceeds the revenue that can be earned, so the property is retired—withdrawn from the market.

The profit-maximizing quality level is also affected by changes in price. An increase in price shifts the marginal-revenue curve upward, increasing the profit-maximizing quality, as shown by the move from point *i* to point *k*. In this case, the owner spends money to upgrade the dwelling because the extra revenue from a higher quality dwelling exceeds the extra cost of upgrading. In the opposite direction, a decrease in the price will shift the marginal-revenue curve downward, decreasing the quality. If the price reduction is large enough that the marginal-revenue curve lies entirely below the marginal-cost curve, the property will be retired.

When a dwelling is retired from the housing market, there are three possible scenarios:

1. **Boarding up.** A dwelling can be boarded up and taken off the market temporarily. This will be the best option if (*a*) the price is expected to increase sometime in the future and (*b*) the opportunity cost of holding wealth in housing

FIGURE 14–3 Dwelling Age and Quality Level



As a dwelling ages, the marginal cost of maintenance increases, decreasing the profit-maximizing quality level (point *i* to point *j*). An increase in the price of quality increases the profit-maximizing quality (point *i* to point *k*).

instead of another asset, such as a bank account, is relatively low. This sort of temporary retirement was common during the Great Depression.

2. **Conversion.** A dwelling can be converted to nonresidential use such as an office, a store, or a parking lot. Conversion is profitable if the alternative activity generates enough profit to offset the cost of conversion.
3. **Abandonment.** The owner will disown the dwelling if the best alternative use does not generate enough profit to cover the cost of conversion.

Abandonment and Public Policy

We've seen that a dwelling will be abandoned if it cannot be profitably used for housing or any alternative use. Local tax policy can contribute to the abandonment problem if the property tax is inflexible. As a dwelling moves down the quality ladder over time, profit decreases. For example, suppose the annual profit from a dwelling is initially $4,000 and the annual tax is $3,000. If the profit drops to $2,000 per year and the property tax is fixed at $3,000, the owner will abandon the property because the tax exceeds the profit. In contrast, if the property tax were flexible, it would be cut in half from $3,000 to $1,500, and the owner would have an incentive to keep the property because the profit ($2,000) still exceeds the tax.

White (1986) showed that during the 1980s the property tax was the most important factor in abandonment in New York City. The elasticity of abandonment with respect to the property tax was 1.65: a 10 percent increase in the property tax increased the frequency of abandonment by 16.5 percent. For example, if the average assessed value of properties in the Brownsville section of Brooklyn were cut by $1,000 (a 6 percent reduction) the resulting decreases in property taxes would lower the abandonment rate from 17 percent per year to 14.8 percent. Given this large elasticity, a tax cut would generate a fiscal surplus for the city. Although the tax liability per property would decrease, the direct revenue loss would be offset by (1) an increase in the number of properties on the tax rolls and (2) a decrease in the number of properties that the city must either take over or demolish.

There are external costs associated with abandonment. Recall the third axiom of urban economics:

Externalities cause inefficiency

Abandoned buildings provide targets for vandals and graffiti artists, and they quickly become eyesores. They often become the temporary homes and retail outlets for transients and drug dealers, so they contribute to crime. For these reasons, abandonment decreases the relative attractiveness of a neighborhood, decreasing the rent that other landlords can charge for their properties. The decrease in profit on other properties encourages more abandonment, so the process can be self-reinforcing.

Durability and Supply Elasticity

The durability of housing has important implications for the market supply curve and the price elasticity of supply. An increase in the price of housing increases the quantity supplied in three ways:

1. **Build more new dwellings.** An increase in price increases the profitability of new housing, so more will be built.
2. **Increase maintenance on used dwellings.** A higher price gives owners a greater incentive to spend money on maintenance and repair to slow the movement of dwellings down the quality ladder.
3. **Upgrade used dwellings.** A higher price provides owners an incentive to move their dwellings up the quality ladder by renovation and remodeling.

The bulk of housing on the market at any particular time is used housing. The general rule of thumb is that in a given year, new construction is between 2 percent and 3 percent of the total housing stock. So the supply response to a price hike is determined in large part by the response of used housing. The supply of used housing is relatively inelastic for relatively long periods of time for two reasons. First, the rate of deterioration is relatively low, so even if an increase in maintenance halts the movement of dwellings down the quality ladder, the response is relatively small. Second, remodeling and renovation are relatively expensive, so it takes a relatively large price hike to make upgrading worthwhile.

The same logic applies to a decrease in the market price. If lower prices halt new construction altogether, the quantity supplied decreases by only 2 to 3 percent per year. A lower price decreases maintenance spending and speeds the movement down the quality ladder, but even the fastest deterioration rate is relatively slow. In general, supply is relatively inelastic for long periods.

What is the price elasticity of the supply of housing? The existing studies of housing supply suffer from a number of statistical problems, so their results must be interpreted with caution (see Olsen, 1969 and Quigley, 1979). Ozanne and Struyk (1978) estimate that the 10-year supply elasticity of used housing is between 0.20 and 0.30. In other words, a 10 percent increase in the market price, sustained over a 10-year period, increases the quantity of used housing by between 2 and 3 percent. Over a 10-year period, new construction provides only about 30 percent of the housing stock, so their estimate applies to 70 percent of the housing stock. De Leeuw and Ekanem (1971) estimate that the long-run supply elasticity for rental housing is between 0.30 to 0.70.

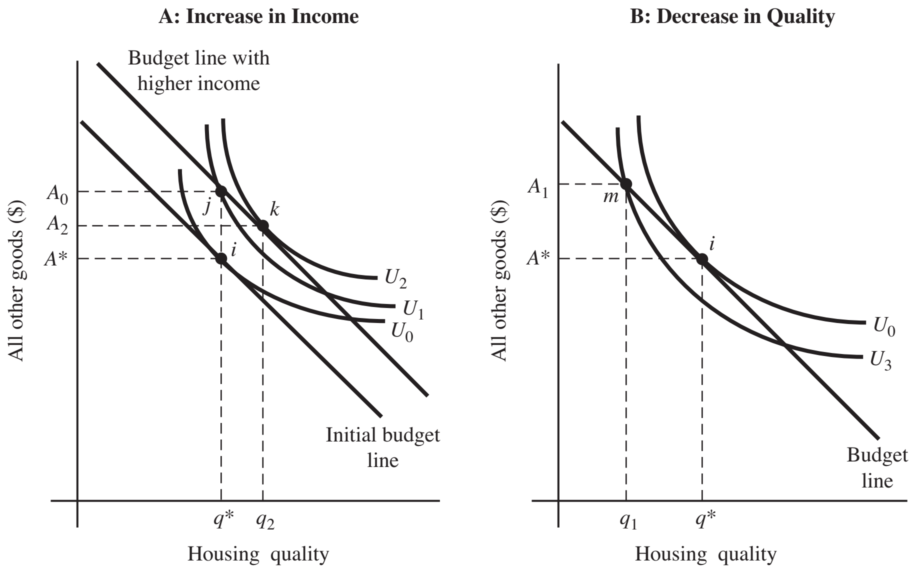
MOVING COSTS AND CONSUMER DISEQUILIBRIUM

For most households, a change in housing consumption requires a move to a different dwelling, and the cost of moving is substantial. In addition to the large cost of moving furniture and other possessions, there is also a large personal cost associated with leaving a neighborhood, with its familiar people, schools, and stores. The notion of neighborhood attachment captures the idea that a move to a new neighborhood disrupts social and consumption patterns, imposing a substantial cost on the household.

Consider first a household whose income is increasing over time. Like many other goods, the utility-maximizing housing consumption increases with income. Specifically, the income elasticity of demand for housing is about 0.75 (Ellwood and Polinski, 1979): A 10 percent increase in income increases housing consumption by about 7.5 percent. As income increases, the gap between the household's ideal dwelling and its actual dwelling will grow, and it eventually may become large enough to justify moving.

We can use the consumer choice model to represent the household's options. The choice model is reviewed in Section 4 of "Tools of Microeconomics," the appendix at the end of the book. In Panel A of Figure 14–4, the starting point is *i*: Given an initial income level represented by the lower budget line, utility is maximized at point *i*, with a housing quality level q^* and A^* of other goods. An increase in income shifts the budget line to the northeast. If the household remains in its original dwelling, housing consumption doesn't change and the household goes to point *j*. All the additional income is spent on other goods, and utility increases from U_0 to U_1. A move to a different dwelling with quality $= q_2$ (point *k*) would generate a higher utility level (U_2). If moving cost were zero, the household would instantly move from point *i* to point *k*. But with large moving costs, the increase in utility ($U_2 - U_1$) must be large enough to offset moving costs. If the household's income continues to increase, the utility gap will eventually be large enough to trigger a move.

FIGURE 14–4 Moving Costs and Consumer Decisions



An increase in income shifts the budget line outward. If the consumer doesn't move, she goes from point *i* to *j* and utility increases from U_0 to U_1. If she moves, she gets to point *k* with utility level U_2. Moving is sensible if the difference in utility ($U_2 - U_1$) is large enough to offset the cost of moving.

A decrease in the quality of the dwelling moves the consumer upward along the budget line from *i* to *m* and decreases utility from U_0 to U_3. The owner will move to a different dwelling with quality = q^* if the difference in utility ($U_0 - U_3$) is large enough to offset the cost of moving.

Consider next a household that occupies a dwelling that moves down the quality ladder over time. As shown in Panel B of Figure 14–4, as quality decreases, so does the monthly rent (the price per unit of quality times quality), so the household moves upward along the original budget line from point *i* to point *m*. As the household moves away from its utility-maximizing point, its utility decreases, and at point *m*, the utility level is shown by the indifference curve U_3. To restore the original utility level U_0, the household must move, but moving generates a moving cost. The household will continue to move upward along its budget line until the utility gap ($U_0 - U_3$) is large enough to offset the cost of moving.

There are two lessons from Figure 14–4. First, households do not instantly respond to a change in circumstances, such as an increase in income or a decrease in the quality of housing, but instead tolerate a gap between the ideal and actual consumption levels. Second, when a household moves, it eliminates the housing gap, so the change in housing consumption is likely to be large. In fact, if a household anticipates future changes in the utility-maximizing consumption level, it may overshoot its current ideal level to reduce the size of a future gap between the ideal and actual consumption.

THE FILTERING MODEL OF THE HOUSING MARKET

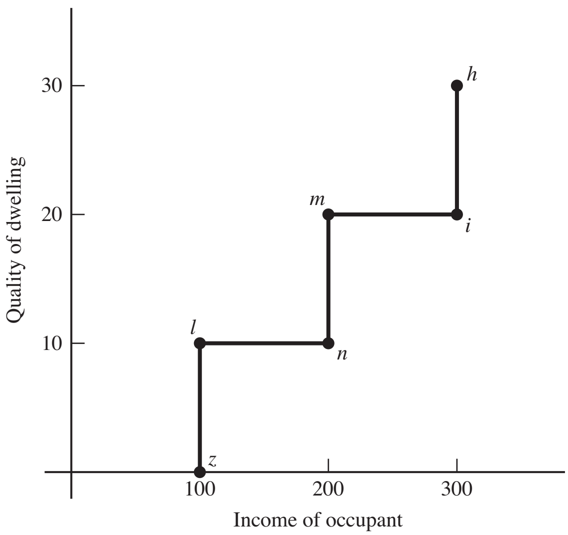
The filtering model of the housing market shows how a dwelling changes over time in quality and in the income of its occupants. The filtering process has two basic features:

1. **Decrease in quality.** The quantity of housing services generated by a dwelling (summarized in "quality") decreases over time because of physical deterioration, technological obsolescence, and changes in housing fashion.
2. **Decrease in occupant income.** As a dwelling moves down the quality ladder, it is occupied by households with progressively lower incomes.

Filtering and the Housing Stepladder

Figure 14–5 represents the essential features of the filtering model. The horizontal axis shows household income, with three types of households, low-income ($100), medium-income ($200), and high-income ($300). The vertical axis shows the quality

FIGURE 14–5 Filtering and the Quality Stepladder



Housing quality is assumed to be proportional to income, and quality decreases by 10 units over the decade. A dwelling occupied by a high-income household moves from point *h* to *i* (quality decreases), and then from point *i* to *m* (vacated by the high-income household after it moves to new housing and then occupied by the middle-income household). Similarly, a dwelling initially occupied by a middle-income household goes from *m* to *n* to *l*, and the dwelling occupied by a low-income household goes from *l* to *z* (retirement).

of housing, with three levels, low (10 units), medium (20), and high (30). Suppose the demand for housing is proportional to income, with each household occupying a dwelling with a quality level equal to 10 percent of its income. At the start of a decade, the chosen points are *h* for the high-income household, *m* for the medium-income household, and *l* for the low-income household.

We will make a few assumptions to simplify matters and make the filtering process transparent. First, suppose that regular maintenance activity by the property owner means that each dwelling loses one unit of quality per year. Second, this underlying movement down the quality ladder can be reversed, but only at a substantial cost for remodeling and renovation. Third, each household has a maximum gap between its ideal and actual housing quality. Once the gap reaches 10 units, the household closes the gap by either upgrading its existing dwelling or moving to a different dwelling.

These assumptions mean that all the action happens at the end of each decade. Each household experiences a growing mismatch between the ideal and the actual quality of housing. The high-income household starts at point *h*, but by the end of the decade is at point *i*, with a 10-unit gap between the ideal and the actual quality. Similarly, the middle-income household goes from point *m* to point *n*, and the low-income household goes from *l* to *z*. At the end of the decade, each household has to choose between restoring the house to the original quality, upgrading, or moving to a different house.

Consider first the choice of the high-income household. The decrease in quality could be reversed by replacing old pipes and leaky windows, repairing dinged woodwork and walls, and retrofitting the house with the latest communications technology. In most cases, this sort of upgrading is costly relative to the price of a new house, which is built from scratch with new materials and modern technology. Therefore, most high-income households move rather than upgrade. Moving to a new house gets them back to point *h*, with a quality level $q = 30$. They vacate their old house with $q = 20$, selling it to the highest bidder.

The middle-income household has a similar problem, with a twist. An alternative to upgrading the original house (moving it from *n* back to *m*) is to buy a used house from a high-income household. If the supply of used housing with quality level 20 is relatively large (if a large number of high-income households sell their old houses), the price will be low enough that moving is less costly than upgrading the old house. So the dwelling formerly occupied by the high-income household moves from point *i* to point *m*. To summarize, the filtering process is shown by the movement of a dwelling down the housing stepladder, from point *h* to *i* (decrease in quality) and then from point *i* to *m* (decrease in occupant income).

The final step involves low-income households that can either upgrade or move into the dwellings vacated by the middle-income households. If the supply of used housing of quality level 10 is relatively large, the price will be low enough that it is more efficient to move rather than upgrade. So the household moves from point *z* back to point *l* by moving to a dwelling that has moved down the stepladder from point *m* to *n* (decrease in quality) and then from point *n* to *l* (decrease in occupant income). The houses vacated by the low-income households with quality $= 0$ are then retired from the housing market.

The filtering process allows each household to restore its desired quality level and get back to its original position on the quality ladder. This occurs even though each dwelling moves down the ladder over the decade. The production of new housing offsets the decline in the quality level of old housing in two ways. First, for each dwelling retired at the low end, one new house is built at the high end. Second, high-income households are accommodated in new housing, freeing up used dwellings and thus allowing other households to reach their ideal quality level.

This little housing model generates tidy results, with a perfect matching of dwellings and households. Of course, things don't operate so smoothly in real markets, but the model captures the essential features of the filtering process. One realistic complication worth mentioning is that the housing occupied by the wealthiest households does not typically filter down to lower-income households. Much of the housing occupied by the wealthy has luxury features (e.g., 10-foot ceilings, fancy fixtures, spa-like bathrooms, and open space) that middle-income and low-income households don't demand. As a result, there is a greater incentive to maintain these dwellings to prevent their movement down the quality ladder.

Subsidies for New Housing

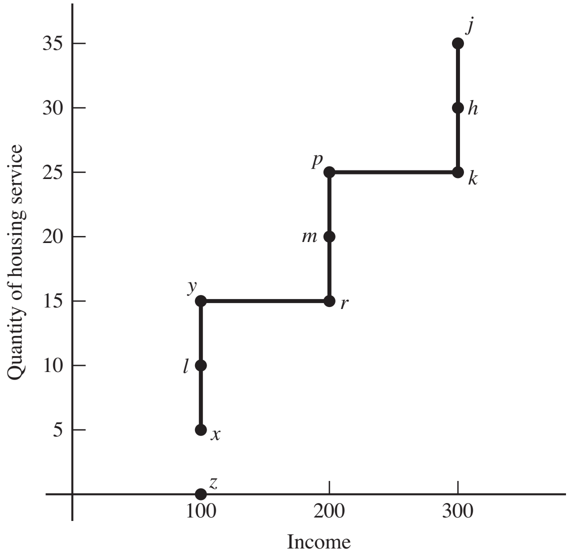
We can use the filtering model to explore the effects of government subsidies on new housing. Although new housing is typically occupied by high-income households, the subsidies hasten the filtering of used housing to lower income households, so everyone benefits from the subsidies.

Suppose the government subsidizes new housing, which we assume is occupied exclusively by high-income households. The subsidy decreases the price of new housing to high-income consumers, and they demand higher quality houses. Figure 14–6 has the same starting points as Figure 14–5, with the high-income household starting at point *h*. Suppose the ideal point for high-income households shifts from point *h* to point *j* (quality level 35). Suppose as before that a household tolerates up to a 10-unit mismatch between its ideal and actual quality level. Given the higher ideal quality level (35), a household will move once the quality level of its original house has dropped from 30 to 25. So the household will move after 5 years instead of 10, and vacate a dwelling with quality level 25 instead of 20. When households are given the option of subsidized new housing, they move sooner and the quality level of housing available for filtering increases.

What are the implications for middle-income and low-income households? Now a middle-income household chooses between moving to a house with quality level 25 (point *p*) and upgrading its old house with quality level 15 (point *r*). The excess supply of $q = 25$ houses will decrease their prices, providing households with an incentive to move and thus vacate their $q = 15$ houses (a move from point *r* to *p*). Similarly, low-income households will have an incentive to move into the $q = 15$ houses vacated by middle-income households (going from point *x* to point *y*).

The filtering process transmits the benefits of housing subsidies throughout the housing market. In this simple example, each type of household experiences a five-unit increase in housing quality. The subsidies shift the quality stepladder up by

FIGURE 14–6 The Effects of Subsidies for New Housing



A subsidy for new housing increases the ideal quality for the high-income household, and it vacates its original dwelling earlier, leaving a higher quality dwelling for filtering down to middle-income households. In general, a housing subsidy shifts the quality stepladder up, so everyone in the housing market gets higher quality housing.

five units: Once everyone has adjusted to the subsidy, the high-income household reaches point *j* instead of point *h*, while the middle-income household reaches *p* instead of *m*, and the low-income household reaches *y* instead of *l*.

The Effects of Growth Controls

Consider next the implications of growth controls that decrease the number of new houses that can be built. To simplify matters, suppose a city outlaws new housing, all of which would have been occupied by high-income households. The building ban will affect the filtering process, leading to higher prices and lower quality housing.

The building ban has a direct effect on high-income households, the potential occupants of new housing. Looking back at Figure 14–5, moving to a new house is no longer an option, so at the end of the decade when the house has deteriorated from quality level 30 to 20, the household must either bear the relatively high cost of upgrading, or tolerate a growing mismatch. In either case, the lack of new housing means that no houses will be vacated for filtering to middle-income households. As a result, middle-income households will either upgrade at a high cost or tolerate a mismatch, meaning that no houses will be vacated for filtering to low-income households.

In our simple model, the imposition of a building ban causes a switch from a market in which everyone moves to a market in which no one moves. Growth control is costly because it causes costly upgrading for some households and forces other households to tolerate bigger housing mismatches. In both cases, the costs of the building ban are borne by everyone in the housing market, not just those who are prevented from buying new houses at the top of the quality ladder.

In a more complex model of the housing market, people enter and leave the market, so some houses will change hands. Nonetheless, growth controls will generate higher prices for two reasons. First, the elimination of new housing will decrease the supply of housing in general, leading to higher prices for new and used housing. Second, prices will reflect the higher costs of upgrading houses to offset deterioration.

Filtering with Rising Income

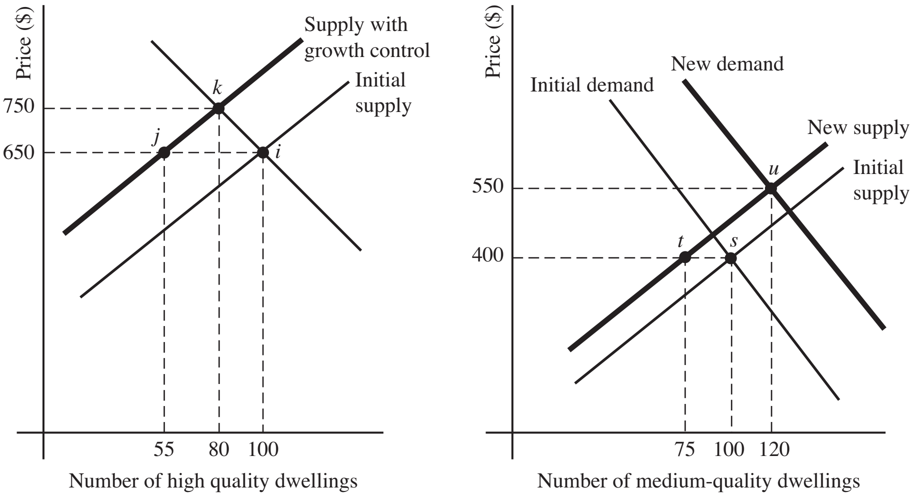
Our simple model of the filtering process assumes constant income over time for each income group. In a model with rising income, rising demand for housing means that filtering is even more beneficial. For example, suppose the ideal quality level for a high-income household increases from 30 to 40. To meet its greater demand by modifying its old house, the household would be forced to not only restore the original quality level, but also to increase it with substantial remodeling. As a result, the advantage of new housing over used housing would be even greater. Similarly, if the incomes of middle-income households are rising over time, they would find the houses vacated by high-income households more attractive relative to upgrading their old houses. The same logic applies to low-income households. In general, when income increases over time, filtering is more advantageous, reflecting the high cost of moving dwellings up the quality ladder.

The Price Effects of Growth Controls

Figure 14–7 provides a closer look at the price effects of a growth-control policy. Consider the market for rental dwellings, and imagine that there are three quality submarkets: high, medium, and low. The number of renters and the number of dwellings are fixed. Each property owner chooses a submarket for the dwelling. The owner of a high-quality dwelling has two options: (1) do nothing, and the dwelling moves down the quality ladder to medium quality, or (2) spend money on maintenance to keep the dwelling at the high-quality level. The owner of a medium-quality dwelling has three options: (1) do nothing, to the quality causing drop to the low level, (2) spend a moderate amount on maintenance to keep the medium quality level, and (3) spend a large amount on maintenance to move the dwelling up the quality ladder to the high level.

Panel A of Figure 14–7 shows the high-quality submarket, which includes new dwellings. If a growth-control policy outlaws new housing, the supply curve will shift to the left, with fewer dwellings supplied at each price. Comparing point *i* to point *j*, 45 new dwellings that would have been built at the original price of $650 are not built. The new supply curve intersects the demand curve at point *k*, with a higher

FIGURE 14–7 The Price Effects of a Building Ban



A: A ban on new housing shifts the supply curve for high-quality dwellings to the left, increasing the equilibrium price.
B: An increase in the price of high-quality dwellings (1) decreases the supply of medium-quality dwellings as fewer filter down from the high-quality market and (2) increases demand as consumers switch from the high-quality submarket to the medium-quality market. The equilibrium price rises.

price ($750) and a smaller quantity of dwellings (80 instead of 100). The net loss in dwellings is 20, which is less than the loss of 45 new dwellings. The higher price of high-quality housing gives property owners a greater incentive to spend money on maintenance to keep the dwellings at the high-quality level. In other words, increased maintenance means fewer dwellings move down the quality ladder, partly offsetting the loss of new housing from the building ban.

Panel B of Figure 14–7 shows the implications for the medium-quality submarket. The increase in the price of high-quality dwellings has two effects. First, as we saw in Panel A, fewer dwellings filter down to the medium-quality market. This is shown as a leftward shift of the market supply curve. Comparing point *t* to point *s*, 25 dwellings that would have filtered down stay at the high-quality level. On the demand side, the two types of dwellings are imperfect substitutes, and consumers move between the two submarkets. The increase in the price of high-quality dwellings (from $650 to $750), causes some consumers to switch to medium-quality dwellings. The demand curve for medium-quality dwellings shifts to the right, with a larger quantity demanded at each price. At the new equilibrium shown by point *u*, the price of medium-quality dwellings is $550, up from $400.

The general lesson from Figure 14–7 is that a supply restriction in the high-quality market causes higher prices in both markets. A growth control policy that restricts the supply of new housing reduces filtering, so its effects are transmitted to the medium-quality submarket. In addition, consumers fleeing higher prices in the high-quality market increase the demand for medium-quality housing. The combination of decreased supply (less filtering) and increased demand (fleeing consumers) generates higher prices in the medium-quality market. The same logic applies to the low-quality submarket. The increase in the price in the medium-quality submarket reduces the flow of apartments downward along the quality ladder and causes some consumers to flee to the low-quality market. The result is higher prices for low-quality housing.

SUMMARY

This chapter explores three characteristics that distinguish housing from other markets. Housing is heterogeneous and durable, and moving costs make it costly to change consumption. Here are the main points from the chapter.

1. The hedonic approach is based on the notion that a dwelling is composed of a bundle of housing services, each with an implicit price.
2. Housing is durable and the owner controls its position on the quality ladder by spending on maintenance, repair, renovation, and remodeling.
3. The supply of housing is relatively inelastic for long periods of time because the bulk of the housing stock is used.
4. The cost of moving is relatively large, so households change their housing consumption infrequently and make large changes when they move.
5. The filtering model explains how a dwelling moves down the quality ladder to households with progressively lower income.
6. Dwellings at different quality levels are related on the supply side because of filtering and related on the demand side because of consumer substitution. A policy that reduces the quantity of new, high-quality dwellings increases the equilibrium prices of all quality levels.

APPLYING THE CONCEPTS

For exercises that have blanks (_____), fill each blank with a single word or number. For exercises with ellipses (. . .), complete the statement with as many words as necessary. For exercises with words in square brackets ([increase, decrease]), circle one of the words.

1. **Churches, Incinerators, and Housing Prices**
   Recent hedonic studies of urban housing markets have concluded that the price of housing is higher closer to churches, and lower closer to incinerators. Suppose that price of a house one kilometer from a church is \$7,000 higher than an identical house two kilometers from a church, and the relationship is linear. The price of a house one kilometer from an incinerator is \$4,000 lower than an identical house two kilometers from an incinerator, and the relationship is linear. Consider a city where a church and an incinerator are two kilometers apart, with the church at $x = 0$ and the incinerator at $x = 2$. The price of housing at $x = 0$ is \$200,000.

a. Draw a housing-price curve for $x = 0$ to $x = 4$.
b. The slope of the housing-price curve is _____ per km for $x = 0$ to $x = 2$, and _____ per km beyond $x = 2$.
c. The road from the church to the fires of the incinerator is paved with progressively [more, less] expensive houses.

2. **Picking a Quality Level**
Suppose dwelling quality is measured on a scale of 1 to 10, and the monthly cost of producing a particular quality equals the square of the quality level: For quality level 1, the cost is $1; for quality level 2, the cost is $4, and so on. The monthly rent equals the price per unit of quality ($P = \$9$) times the quality level.
a. Use a two-panel graph like Figure 14–2 to show the profit-maximizing quality level.
b. At the profit-maximizing quantity of _____, marginal cost equals _____.

3. **Changes in Quality Level**
Using the two-panel graph in Figure 14–2 as a model, show the effects of two changes in the housing market.
a. An increase in price [increases, decreases] the slope of the _____ curve and shifts the _____ curve [upward, downward]. The profit-maximizing quality [increases, decreases].
b. An increase in the wages of housing-maintenance workers (plumbers, roofers, painters, carpenters) [increases, decreases] the slope of the _____ curve and shifts the _____ curve [upward, downward]. The profit-maximizing quality [increases, decreases].

4. **Stay or Move?**
Consider a household with income of $100 and housing consumption equal to 40 units of quality. The income elasticity of demand for housing is 0.75.
a. Use a graph like the one shown in Figure 14–4 to show the household's initial utility-maximizing choice (labeled *i*).
b. Suppose the household's income increases to $120. Show the household's new choice in the absence of moving costs (labeled *m* for "move") and in the presence of moving cost high enough that the household will not move (labeled *s* for "stay").
c. Suppose that given its moving cost, the household will move if the gap between its actual housing consumption and its ideal consumption (with moving cost = 0) is at least 30 percent. The household will choose point [*m*, *s*] because . . .
d. The threshold income level, the income just high enough to cause the household to move is _____ because . . .

5. **Moving and Consumer Surplus**
Consider a city where the demand for housing for the typical household is $Q = 1{,}000 - 1{,}000 \cdot P$, where Q is the square footage of housing demanded and P is the price per square foot.
a. Draw the demand curve for housing.
b. At the initial price of housing of $0.50 per square foot. the typical household chooses a quantity of _____ and gets a consumer surplus of _____.

c. Suppose the price decreases to $0.40. If the household does not move, consumer surplus = _____. If the household moves, consumer surplus = _____. If the cost of moving is $12, the typical household [will, won't] move because . . .

6. Stepladder versus Ramp

The real estate agents in Rampville are baffled. Although dwellings are subject to the normal decrease in quality over time, there is no filtering: No household has changed its house in the last 10 years. Provide an explanation for this phenomenon. Use a figure like Figure 14–5 to depict the housing market in Rampville.

7. Subsidies for Middle-Income Households

Consider the effects of housing subsidies on the filtering process. As before, assume that each household tolerates up to a 10-unit mismatch between its ideal and the actual quality of housing. All new housing is built for high-income households. Upgrading housing to a higher quality level is prohibitively expensive. Use Figure 14–5 as a starting point, with points *l*, *m*, and *h*. Suppose a subsidy program pays 20 percent of the housing expenses of middle-income households. The price elasticity of demand for housing is 1.0.

a. For middle-income households, the ideal quality of housing services increases from _____ to _____. A middle-income household will reach a 10-unit mismatch after _____ years. At that time, a high-income household has a mismatch of _____.

b. To induce the high-income household to move, the price of 24-unit housing [increases, decreases].

c. To induce a low-income household to buy the 14-unit dwellings vacated by middle-income households, the price of 14-unit dwellings [increases, decreases].

d. Draw a graph like Figure 14–6 to show the long-run effects of the subsidy program on the housing stepladder.

8. Affordable Housing

Suppose the national government pays for a program of affordable housing—subsidized new housing for low-income households. Assume that "affordable" housing is low-quality housing.

a. Use a graph to show the effects of affordable housing on the markets for low-quality and medium-quality housing.

b. The policy [increases, decreases] the price of low-quality housing because . . .

c. The policy [speeds up, slows down] the filtering process from the medium-quality to the low-quality market because . . .

d. The policy causes consumers to switch from the [low, medium] quality market to the [low, medium] quality market because . . .

e. The policy [increases, decreases] the price of medium quality housing because . . .

9. Hurricane Katrina

Hurricane Katrina destroyed a large fraction of the housing stock of New Orleans, with relatively large losses in the low-quality market. To simplify, suppose that the hurricane destroyed only low-quality houses.

a. Use a graph to show the effects of the hurricane on the markets for low-quality and medium-quality housing.
b. The hurricane [increased, decreased] the equilibrium price of low-quality housing because . . .
c. The hurricane [sped up, slowed down] the filtering process from medium-quality to low-quality markets because . . .
d. The hurricane caused consumers to switch from the [low, medium] quality market to the [low, medium] quality market.
e. The hurricane [increases, decreases] the price of medium quality housing because . . .

10. Mandated Energy Conservation

Consider a city where each year 10 households move from old houses (energy consumption 20 units per year) into new houses (energy consumption 12 units per year). Suppose a new law requires that all houses built in the future include an energy-saving device that adds $5,000 to construction cost and cuts energy consumption from 12 to 9 units.

a. Use a figure like Figure 14–7 to show the effects of the law on the markets for new and used housing.
b. According to Ms. Wizard, "I predict that the law will actually harm our energy-conservation efforts." Ms. Wizard is correct if . . .

REFERENCES AND ADDITIONAL READING

1. De Leeuw, Frank, and Nkanta Ekanem. "The Supply of Rental Housing." *American Economic Review* 61 (1971), pp. 806–17.
2. Ellwood, David, and Mitchell Polinski. "An Empirical Reconciliation of Micro and Grouped Estimates of the Demand for Housing." *Review of Economics and Statistics* 61 (1979), pp. 199–205.
3. Follain, James R., and Emmanuel Jiminez. "Estimating the Demand for Housing Characteristics." *Regional Science and Urban Economics* 15 (1985), pp. 77–107.
4. Kain, John, and John Quigley. "Measuring the Value of Housing Quality." *Journal of the American Statistical Association* 65 (1970), pp. 532–48.
5. Olsen, Edgar O. "A Competitive Theory of the Housing Market." *American Economic Review* 59 (1969), pp. 612–22.
6. Ozanne, L., and Raymond Struyk. "The Price Elasticity of Supply of Housing Services." In *Urban Housing Markets: Recent Directions in Research and Policy*, eds. L. S. Bourne and J. R. Hitchcock. Toronto: University of Toronto Press, 1978.
7. Quigley, John M. "What Have We Learned about Housing Markets?" In *Current Issues in Urban Economics,* eds. Peter Mieszkowski and Mahlon Straszheim. Baltimore: Johns Hopkins University Press, 1979.
8. White, Michelle. "Property Taxes and Urban Housing Abandonment." *Journal of Urban Economics* 20 (1986), pp. 312–30.

CHAPTER 15

Housing Policy

Fools build houses, and wise men buy them.

—Proverb

Almost any man worthy of his salt would fight to defend his home, but no one ever heard of a man going to war for his boarding house.

—Mark Twain

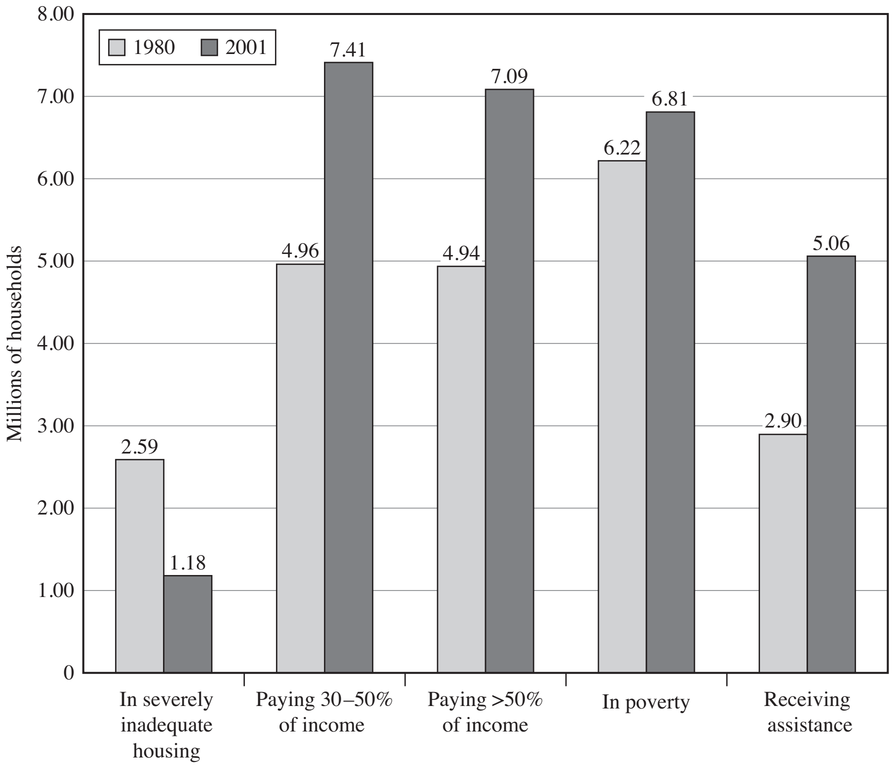
This chapter discusses housing policies that assist low-income as well as high-income households. For low-income households, the federal government spends about $30 billion per year to provide public housing, subsidize privately produced housing, and provide vouchers that households use to help pay for housing they choose themselves. The federal government also uses a number of community development programs to support local efforts to improve housing conditions and revitalize neighborhoods. For middle-income and high-income households, the government sacrifices about $66 billion in tax revenue per year to subsidize home ownership through the tax deduction for mortgage interest.

Figure 15–1 (page 387) shows the number of renter households with inadequate housing or a relatively high rent burden. Between 1980 and 2001, the number of renter households in "severely inadequate housing" dropped significantly, while the number paying more than 30 percent of income for rent increased from 9.9 million to 14.5 million. Only about 30 percent of the households that are eligible for public housing, subsidized housing, or vouchers actually receive assistance. There are long waiting lists for households to get into public housing, and voucher programs are not funded at a level necessary to serve all the eligible households.

PUBLIC HOUSING

In 1998, about 1.3 million households lived in public housing. In terms of budgetary costs, the federal government's outlays for public housing included $3.1 billion for operating subsidies (to cover the gap between the rent collected from tenants and the costs of operating the project) and $3.8 billion for capital expenditures (for

FIGURE 15–1 Number of Households in Inadequate or Unaffordable Housing



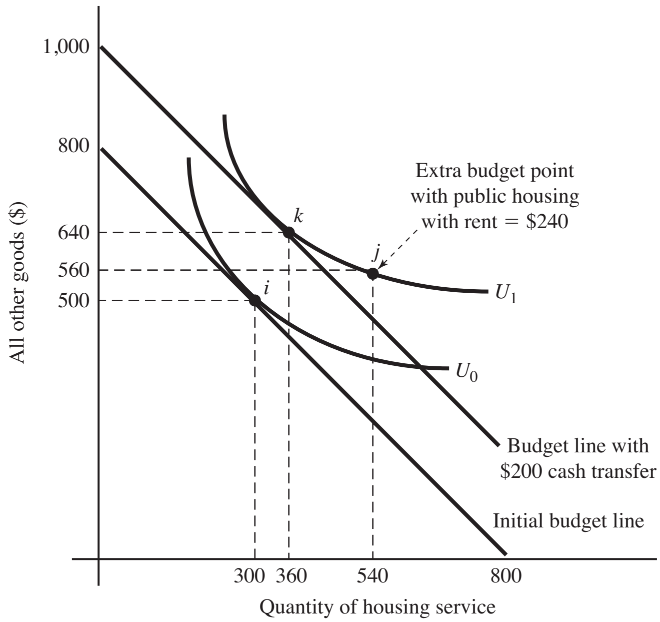
Source: Based on John Quigley. "A Decent Home: Housing Policy in Perspective." *Brookings-Wharton Papers on Urban Affairs* (2000), pp. 53–99; John Quigley and Steren Rafael. "Is Housing Affordable? Why Isn't It More Affordable?" *Journal of Economic Perspectives* 18 (2004), pp. 191–214.

repair, upgrading, and demolition). Public housing is managed by local housing authorities, subject to rules adopted by the federal government. The rent charged to a particular tenant can be no more than 30 percent of the household's income. In recent years, there has been no investment in new public housing, but funds have been allocated for the repair and renovation of decaying public housing.

Public Housing and Recipient Welfare

Figure 15–2 uses the consumer choice model to represent the housing choices of a low-income household. The choice model is reviewed in Section 4 of "Tools of Microeconomics," the appendix at the end of the book. The horizontal axis measures housing consumption (the quantity of housing service from a dwelling), and the vertical axis measures the monthly consumption of all other goods (in dollars). Suppose the household has a monthly income of $800 and the price of housing is $1 per unit of service. The initial budget line shows the household's options, with a

FIGURE 15–2 Utility Maximization with Public Housing



The government charges public housing tenants 30% of income ($240) for a dwelling with 540 units of service, so point *j* is added to the budget set and the recipient's utility level increases from U_0 to U_1. The recipient would reach the same utility level with a cash transfer of $200. The subsidy is $300 ($540 – $240 rent) so the value of public housing is two-thirds of the subsidy.

one-for-one trade-off between housing and other goods. The household maximizes utility at point *i*, where the indifference curve is tangent to the budget line. At this point, the marginal rate of substitution of housing and other goods equals the price ratio, so the household is doing the best it can. The household occupies a dwelling that generates 300 units of housing service at a cost of $300, leaving $500 to spend on other goods.

Suppose the typical dwelling in public housing generates 540 units of housing service. The government charges rent equal to 30 percent of the tenant's income, or $240 for our hypothetical household (0.30 times $800). Public housing adds point *j* to the household's budget set. In addition to all the points on or below the initial budget line, the household also has the option of 540 units of service for only $240, leaving $560 to spend on other goods. By accepting the offer of public housing, the household gets more of both goods and reaches a higher utility level (U_1 instead of U_0).

How much is public housing worth to the tenant? The value of public housing can be measured by answering the following question: What cash payment would make the household indifferent between the cash and the public housing offer (a 540-unit dwelling at a price of $240)? In other words, what cash payment would move the household to the indifference curve associated with public housing (U_1)?

In Figure 15–2, a \$200 payment shifts the budget line upward in a parallel fashion, and the household maximizes utility at point *k*, reaching the indifference curve U_1. In other words, the household is indifferent between \$200 in cash and a housing subsidy of \$300 (equal to the dollar value of the 540-unit dwelling (\$540) minus the rent paid (\$240)). The value of public housing is two-thirds of the subsidy. This is consistent with the results of studies that measure the value of public housing to recipients (Green and Malpezzi, 2003).

How does the cost of new public housing compare to the cost of private housing? Public housing is more expensive for two reasons. First, the private sector can build new low-income housing more efficiently than the public sector. Second, there is a plentiful supply of used low-quality housing, so even the least costly new housing costs more than used housing. Economists measure the production efficiency of public housing by the ratio of the market value of the dwelling divided by the production cost. According to Green and Malpezzi (2003), the production efficiency of public housing is 0.50, meaning that the production cost is twice the market value. In our example, if it costs the government \$1,080 to produce a dwelling worth \$540, the budgetary cost of getting to the public-housing point (point *j* in Figure 15–2) is \$840 (\$1,080 minus the \$240 rent paid by tenants).

What is the bang per buck of public housing? In other words, what is the recipient benefit per taxpayer dollar spent on public housing? As shown in Figure 15–2, recipients would be indifferent between \$200 in cash and public housing, so the bang from each public-housing dwelling is \$200. If the government subsidy per dwelling is \$840, the bang per buck is \$0.24 (equal to \$200/\$840).

Subsidies for Private Housing

One alternative to public housing is a system of subsidies to encourage the private sector to build and manage low-income housing. Under two programs, Section 236 and Section 8—Project Based, the government pays a property owner the difference between the household's rent and the "fair market rent." In most cases, the household's contribution is 30 percent of its income. The fair market rent is determined by either the cost of building and managing the property or the prevailing rent in the area. For example, suppose the fair market rent of an apartment is \$500, and an eligible household's income is \$800. In this case, the household would pay \$240 (30 percent of \$800) and the government would pay \$260, for a total payment of \$500.

Under these subsidy programs, the federal government signs long-term contracts to provide annual payments to property owners. The owner is guaranteed the fair market rent on all units occupied by eligible households. In 1998, 1.4 million households were assisted under the Section 8 program, and just under half a million were assisted under Section 236 (Quigley, 2000).

Although subsidies for private housing are more efficient than public housing, they still produce dwellings with market values less than their production cost. As reported by Green and Malpezzi (2003), estimates of the production efficiency of subsidized new private housing range from 0.61 to 0.85, with a median of about 0.75.

Low-Income Housing Tax Credit

The Tax Reform Act of 1986 instituted a program of tax credits for investment in affordable housing for low-income households. To qualify for tax credits, a project must set aside a fraction of the dwellings to be "rent restricted" (a maximum rent for the dwelling) and "occupant restricted" (a maximum income for tenants). There are two tests for these set-asides:

1. **The 20/50 test:** At least 20 percent of the rental dwellings must be occupied by households with income no greater than 50 percent of the median area income.
2. **The 40/60 test:** At least 40 percent of the rental dwellings must be occupied by households with income no greater than 60 percent of the median area income.

In both cases, the maximum rent is 30 percent of the qualifying income. A builder of low-income housing earns an annual credit of 9 percent of the project cost attributable to low-income housing. For example, a project with a $10 million cost for low-income housing generates an annual tax credit of $900,000, and the builder's federal tax liability drops by that amount. The builder can get the annual credit for up to 10 years, although the set-aside restrictions apply for 15 years.

The tax-credit program has been used in projects that had produced 700,000 low-income dwellings by 1999. The credits are allocated by state housing authorities, subject to restrictions established by the federal government. In 2002, each state was allowed to grant credits totaling $1.50 per capita, and the allocation rose to $1.75 per capita in 2003. In 2003, the cost of this tax expenditure in terms of lost revenue was about $3.5 billion (Office of Management and Budget, 2002).

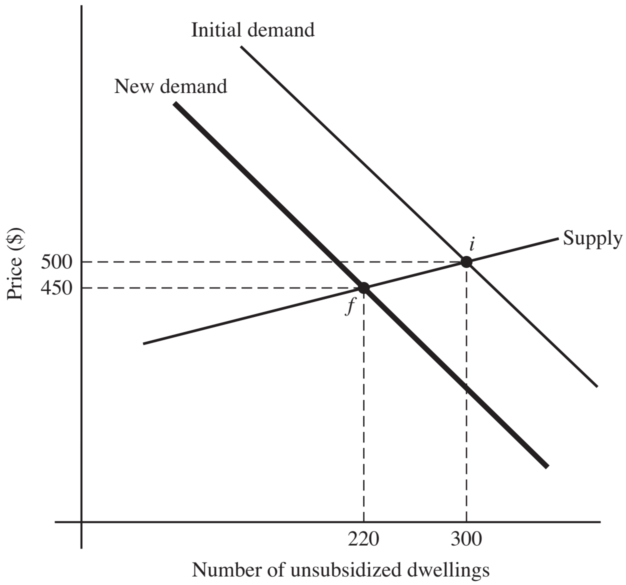
The revenue loss from the tax-credit program is large relative to the amount of housing produced. The complexity and riskiness of the program mean that investors demand relatively high rates of return on funds invested. In 1996, each dollar of federal subsidy produced only about $0.62 worth of housing, so the efficiency of the tax-credit program is not much different from that of public housing (Quigley, 2000). DiPasquale, Fricke, and Garcia-Diaz (2003) estimate that the cost of providing a one-bedroom rental unit under the tax-credit program is 19 percent higher than an equivalent unsubsidized unit, and the cost gap for a two-bedroom unit is 14 percent.

The Market Effects of Subsidized Housing

Government subsidies for privately produced housing currently support about 2.6 million dwellings. Does this mean that housing subsidies have increased the total housing stock by 2.6 million units? As we'll see, housing subsidies displace unsubsidized housing, so the net effect of the subsidies on the housing stock is relatively small.

Figure 15–3 (page 391) shows the effects of housing subsidies on the unsubsidized market. At the initial equilibrium before subsidies (shown by point *i*) the price is $500 and the quantity is 300 unsubsidized dwellings. If government subsidies generate 100 new housing units, 100 households leave the unsubsidized market, so

FIGURE 15–3 The Displacement Effect of Housing Subsidies



Housing subsidies decrease the demand for unsubsidized low-quality dwellings. The resulting excess supply decreases the equilibrium price from $500 to $450. The number of unsubsidized dwellings decreases from 300 to 220.

the demand curve shifts to the left by 100 dwellings. The resulting excess supply will cause the price to drop, and equilibrium is restored at point *f*, with a price of $450 and a quantity of 220. In other words, the number of unsubsidized dwellings decreases by 80, partly offsetting the effects of housing subsidies.

Let's take a closer look at the movement downward along the supply curve for unsubsidized housing as the price drops. This is low-income housing, so we are at the low end of the housing quality ladder. The quantity of low-quality unsubsidized dwellings decreases for two reasons.

- **More retirement.** A decrease in the price of low-quality dwellings reduces the profit on unsubsidized dwellings, so more dwellings are retired from the housing market, either converted to another use or abandoned.
- **Slower downward filtering.** A decrease in the price of low-quality housing relative to the price of medium-quality housing causes property owners to slow the movement of dwellings downward along the quality ladder. Owners spend more on maintenance and repair to hold their dwellings in the medium-quality submarket.

What are the facts on the displacement or "crowding out" of unsubsidized housing by subsidized housing? Murray (1999) estimates that in the long run, the reduction in the number of unsubsidized dwellings is at least one-third of the increase in subsidized dwellings. His "best estimate" is that there is a one-for-one crowding out, implying that housing subsidies do not increase the total housing stock in the long run. Similarly, a study of the low-income tax credit program by Malpezzi and Vandell (2002) found no evidence that the program increased the total housing stock.

HOUSING VOUCHERS

So far we have considered policies that help low-income households by increasing the supply of housing. Under a demand-side policy, low-income households are given housing coupons or vouchers that can be redeemed for housing. Like food stamps, housing vouchers allow recipients to make their own consumption choices. In 1999, about 1.6 million households received housing vouchers at a budgetary cost of $7 billion. (Some of the vouchers are called "rent certificates.") A voucher household must occupy a dwelling that meets minimum quality standards. The face value of a voucher is based on household income and the fair market rent in the metropolitan area. The formula is

$$\text{Face value} = \text{Fair market rent} - 0.30 \cdot \text{Income}$$

The fair market rent is defined as the 45th percentile of rents in the metropolitan area (45 percent of dwellings rent for less).

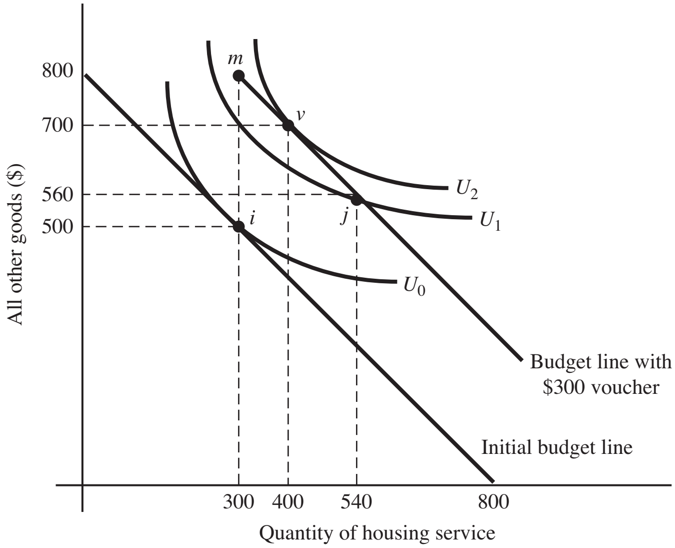
Vouchers and Consumer Welfare

We can use the consumer choice model to show the effects of a voucher on a recipient's budget decisions and utility. As in the example of public housing, income is $800 and the price of housing is $1 per unit of housing service. Suppose the fair market rent is $540 (a dwelling with 540 units of housing service). The face value of the voucher is $300 (equal to $540 – 0.30 · $800).

Figure 15–4 (page 393) shows the recipient's response to a housing voucher. The voucher program shifts the household's budget line to the right by $300. Point *m* is in the new budget set because the household could use the voucher to get $300 worth of housing (assuming the minimum standard is met) and spend all of its own income ($800) on other goods. As spending on housing rises above $300, there is a dollar-for-dollar trade-off between housing and other goods. We are assuming for the moment that vouchers do not affect the market price of housing.

The voucher increases housing consumption and household utility. The new utility-maximizing point is *v*, compared to the initial point *i*. Housing consumption increases from $300 to $400, and spending on other goods increases to $700. In other words, the household spends one-third of the voucher on housing, leaving two-thirds to spend on other goods. Comparing point *v* to the public-housing point (*j*), we see that utility is higher under the voucher program. From the recipient's perspective, vouchers are better because they provide more options, letting

FIGURE 15–4 Utility Maximization with Housing Voucher



A $300 voucher shifts the budget line to the right by $300 and the recipient chooses point *v*, with 400 units of housing service and $700 of other goods. The voucher generates higher utility than public housing because the voucher gives the recipient more options.

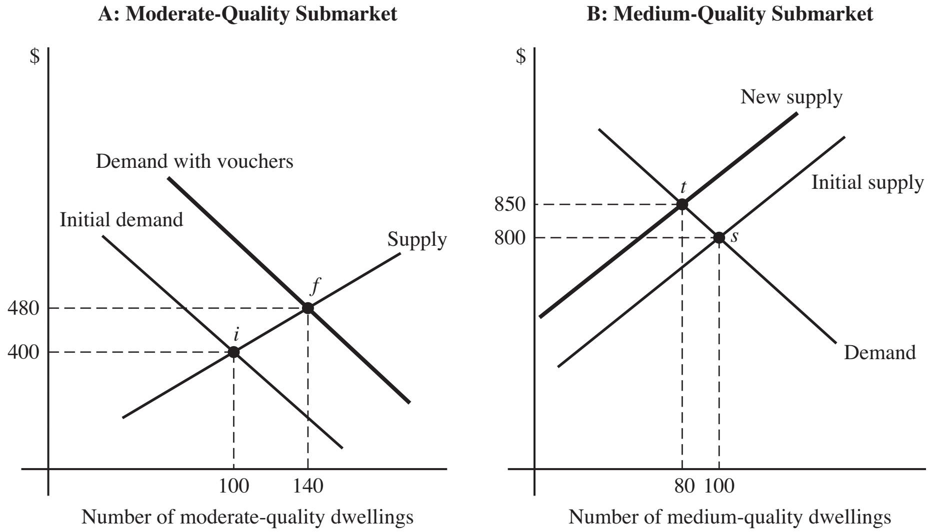
the recipient pick the utility-maximizing consumption bundle. In fact, if the utility-maximizing spending on housing is at least $300, the voucher is equivalent to a cash transfer of $300.

Market Effects of Vouchers

What are the market effects of housing vouchers? In Figure 15–4, vouchers increase the housing consumption from 300 to 400 units of housing service. What are the implications for housing prices and the welfare of recipients and other households? Consider a city with two income groups (low-income and middle-income) and three levels of housing quality (low, moderate, and medium). All low-income households live in low-quality (300 units of housing service) or moderate-quality (400 units of housing service) housing. All middle-income households live in medium-quality housing.

Figure 15–5 (page 394) shows the market effects of the voucher program. In Panel A, the initial equilibrium in the moderate-quality submarket is shown by point *i*, with a price of $400 and a quantity of 100 dwellings. The voucher program shifts the demand curve for moderate-quality housing to the right as most recipients switch from the low-quality submarket to the moderate submarket. The resulting excess demand for housing raises the price until equilibrium is restored at point *f*, with a

FIGURE 15–5 The Market Effects of Housing Vouchers



A: A voucher program increases the demand for moderate-quality housing and increases the equilibrium price (from \$400 to \$480).
B: The increase in the price of moderate-quality dwellings decreases the supply of medium-quality dwellings as more dwellings filter down to the moderate-quality submarket. The equilibrium price rises from \$800 to \$850.

price of \$480. The quantity of moderate-quality dwellings supplied increases by 40 dwellings for two reasons:

1. **More rapid filtering from medium-quality submarket.** The price of moderate-quality dwellings increases relative to the price of medium-quality dwellings, so dwellings move down the quality ladder more rapidly.
2. **Slower filtering to low-quality submarket.** The higher price of moderate-quality dwellings gives owners a greater incentive to keep their dwellings at the moderate level, so fewer dwellings filter down to the low level.

Panel B of Figure 15–5 shows the implications of housing vouchers for the medium-quality submarket. By speeding up the filtering process, vouchers decrease the supply of medium-quality housing, shifting the supply curve to the left. As a result, the equilibrium price increases from \$800 (point *s*) to \$850 (point *t*). In other words, medium-income households are hurt by the housing vouchers given to low-income households. Although the two types of households do not compete directly for housing in the medium-quality market, they compete indirectly through the filtering process.

What about the low-quality and high-quality submarkets? At the high end, the increase in the price of medium-quality housing increases the filtering rate from higher-quality levels, and the resulting decrease in supply of high-quality dwellings increases the price. At the low end, an increase in the price of moderate-quality dwellings slows the filtering down to the low-quality submarket, and the decrease in supply of low-quality dwellings increases the price. This means that low-income households that do not get vouchers are hurt by higher housing prices. In general, the filtering process means that the increase in demand for housing triggered by the vouchers increases the prices of housing of all quality levels.

A recent study of the 90 largest metropolitan areas estimates the effects of housing vouchers on the prices in different submarkets (Susin, 2002). The results suggest that vouchers increased the price of low-income housing by about 16 percent. This implies that the price elasticity of supply of low-income housing is relatively low—between zero and 0.38. The estimated price effect on middle-income housing was smaller (3.2 percent), and the effect for high-income housing was close to zero.

Portable Vouchers: Moving to Opportunity

A social experiment called Moving to Opportunity (MTO) was designed to test whether low-income households would fare better in neighborhoods with relatively low poverty rates. A group of public-housing tenants and other low-income households receiving rental assistance were randomly assigned to one of three groups:

1. **The MTO treatment group:** Received rent vouchers that initially could be used only in neighborhoods with poverty rates less than 10 percent. After one year, the vouchers could be used anywhere.
2. **A comparison group:** Received vouchers with no restrictions.
3. **A control group:** Received assistance tied to a specific public housing project.

The results from the first stage of research suggest that mobility improves the outcomes of low-income children (Goering, 2003). The mobile families (groups 1 and 2) moved into neighborhoods that had less poverty (15 percentage points lower on average), less segregation, less crime, and better schools. The children in the mobile groups had fewer behavioral problems, better school performance, and less juvenile crime. The adults in the mobile group reported better health and less stress and fear of crime. In contrast, there was no significant difference between the mobile groups and the control group (3) in terms of adult employment, hours worked, or use of public assistance.

COMMUNITY DEVELOPMENT AND URBAN RENEWAL

Dozens of programs and policies fit under the term *community development*. The mandate for federally supported community development calls for "systematic and sustained action by the federal, state, and local governments to eliminate blight, to conserve and renew older areas, to improve the living environment of low- and moderate-income families, and to develop new centers of population growth and economic activity." The two principal purposes of community development policies are to revitalize declining areas of the city and improve the housing of low-income households.

Urban Renewal

Urban renewal, the original community development program in the United States, was established under the Housing Act of 1949 and was eventually dropped in 1973. The national government provided local governments with the power and the money to demolish and rebuild parts of their cities. Local agencies acquired property under the right of eminent domain, cleared the site of "undesirable" uses (such as low-income housing and small businesses), and then either built a public facility or sold the site to a private developer at a discount. The federal government covered two-thirds of the loss incurred by local government. The private developer built housing, government buildings, or commercial establishments.

The urban renewal program displaced low-income households in favor of higher-income residents, public facilities, and commercial operations. A total of 600,000 dwellings were demolished and were replaced by 250,000 new dwellings, 120 million square feet of public facilities, and 224 million square feet of commercial space. The assessed value of property on the renewed sites increased by a factor of 4.6. The critics of urban renewal focus on its demolition aspects, pointing out that 2 million low-income people were displaced. The defenders of the program focus on its rebuilding aspects, pointing out that the new commercial developments provided jobs for the poor residents of the central city.

Recent Community Development Programs

More recent federal community-development programs have avoided many of the problems of the urban renewal program. The newer programs are executed on a smaller scale, so they displace fewer households. In addition, the modern programs place a greater emphasis on providing housing for low-income households.

The bulk of federal funding for community development is for Community Development Block Grants (CDBG). In 1997, the CDBG budget was $4.7 billion, with about 70 percent of the funds going to central cities and urban counties. The allocation formula for fund distribution favors cities with relatively old and overcrowded housing, high poverty rates, and slow economic growth. The funds are used to improve housing, support public services, promote economic development, and clear land for new development. As Connerly and Liou (1998) report, about 40 percent of CDBG funds are spent on housing programs, with the remainder divided into public works (20 percent), economic development (13 percent), and public services (10 percent) and other programs. The CDBG program is a relatively small part of the system of intergovernmental grants—funding never reaches 8 percent of the total grant budget.

Several recently developed grant programs provide flexibility to local governments (Quigley, 2000). Under the McKinney Act, funds are provided to address homelessness, with funds for emergency shelters and the rehabilitation of single-room occupancy dwellings. Under the HOPE IV program, local governments receive funds to renovate or demolish obsolete public housing projects. The HOME program provides funds to produce and preserve low-income housing. These three programs allow most of the decisions to be made at the local level, by local officials and members of nonprofit community organizations.

Homelessness

As we've seen, the federal government provides grants to local governments to address homelessness. The McKinney Homeless Assistance Act of 1987 defines a homeless person as someone who sleeps (a) outside, (b) inside in places not intended for sleeping (e.g., the lobby of a public building), or (c) in housing shelters (places providing temporary housing).

Table 15–1 shows the makeup of the U.S. homeless population, drawing on various studies between 2003 and 2005. The background and experiences of many of the homeless are not conducive to economic success. Among the frequently occurring problems among the homeless are low educational attainment, time spent in jail and prison, drug dependency, and mental health problems.

What causes homelessness? From an economic perspective, a person will be homeless if his or her income is low enough relative to the price of housing that it is not sensible—or not even possible—to purchase housing services. This simple theory is supported by studies of homelessness, which show that homeless rates are higher in areas with relatively high rent on low-quality housing (Honig and Filer, 1993; Green and Malpezzi, 2003). Honig and Flier estimate an elasticity of homelessness with respect to rent on low-quality housing of 1.25, meaning that a 10 percent increase in rent increases the homeless rate by 12.5 percent.

Several other factors contribute to homelessness. The homeless population is higher in areas with weak labor markets (slow employment growth), low levels of public assistance, and low institutionalization rates for the mentally ill. These other

TABLE 15–1 Makeup of U.S. Homeless Population

Group	Percent of Homeless Population	Date for Data
Children under 18	39	2003
Children under 5	16	2003
Single men	43	2005
Single women	17	2005
Families with children	33	2005
African American	49	2004
Caucasian	35	2004
Hispanic	13	2004
Native American	2	2004
Asian	1	2004

Source: National Coalition for the Homeless, *Who is Homeless? NCH Fact Sheet #3* (June 2006). http://www.nationalhomeless.org.

factors suggest that homelessness is not simply a housing problem, but a complex problem with many causes. Although it appears that current housing policies have little effect on homelessness (Early, 1998, 1999), there is evidence that the problem could be mitigated by policies that improve the functioning of the low end of the housing market (O'Flaherty, 1996; Green and Malpezzi, 2003).

WHICH HOUSING POLICY IS BEST?

Economists and policy makers have an ongoing debate about the relative merits of supply-side and demand-side housing policies. As we've seen, public and subsidized housing is more costly than unsubsidized housing, and it gives recipients fewer options. On the other side of the market, vouchers give recipients more options and generate housing at a relatively low cost, but they increase housing demand and housing prices. The higher prices are especially problematic for the 70 percent of low-income households who are eligible for housing assistance but don't receive any.

Which policy is best? Analyses by Apgar (1990) and Struyk (1990) suggest that this question is misguided. The "best" policy varies across metropolitan areas and within metropolitan areas, depending on market conditions. In areas with a plentiful supply of low-income housing and a relatively large elasticity of supply, the price effects of vouchers will be relatively small. In contrast, in areas with a relatively inelastic supply of low-income housing, vouchers will generate large price hikes, and carefully crafted supply-side policies could play an important role. Given the large cost of new low-income housing, Struyk (1990) advocates policies that preserve the existing supply of low-income housing, for example, rehabilitation grants and property-tax abatements. He also argues for replacing the tax-credit program with direct grants to local governments. The direct grants would give local officials the flexibility to decide the best policy mix of supply subsidies and vouchers.

SUBSIDIES FOR MORTGAGE INTEREST

In 2002, the federal government provided tax breaks to homeowners that reduced total tax revenue by $66 billion (Office of Management and Budget, 2002). Homeowners can deduct mortgage-interest payments from their gross income, so every dollar of mortgage interest decreases the tax liability by the household's marginal tax rate. For example, if the marginal tax rate is 28 percent, every dollar spent on mortgage interest decreases the tax liability by $0.28. This is an example of a "tax expenditure." Instead of giving money directly to homeowners (an expenditure), the government cuts their taxes. The budgetary consequence of a tax expenditure is the same as an explicit expenditure: In both cases, the government either spends less on other programs or increases other taxes.

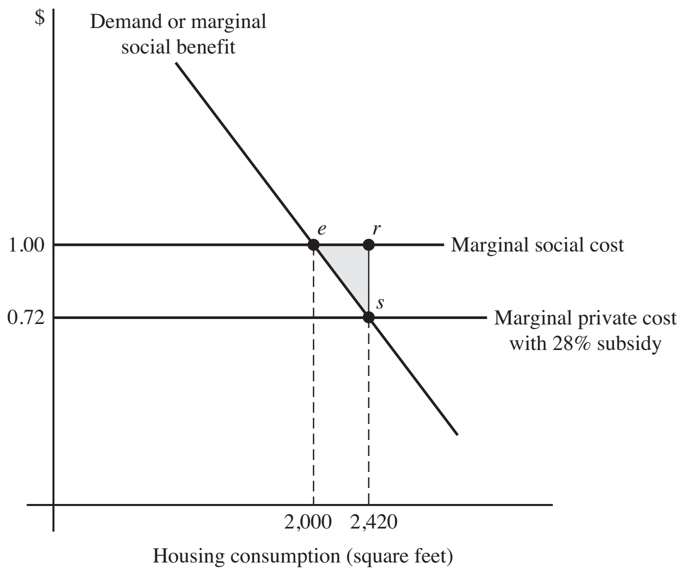
The household's benefit from the mortgage tax break increases with household income for two reasons. First, under a progressive tax system, the marginal tax rate increases with income, so the tax benefit per dollar of mortgage interest increases with income. Second, because the demand for housing increases with income, wealthier households have larger mortgage payments and thus larger deductions. In 2002, about 62 percent of the benefits of this tax break went to households with income above $100,000.

Mortgage Subsidy and Efficiency

We can use the marginal principle to explore the efficiency effects of the mortgage subsidy. The marginal principle is reviewed in Section 1.1 of "Tools of Microeconomics," the appendix at the end of the book. In Figure 15–6, the demand curve shows a household's willingness to pay for housing, or the marginal benefit of housing. The horizontal line at $1 shows the marginal social cost (the opportunity cost) of housing: By spending $1 on housing, society sacrifices $1 worth of investment in other capital (e.g., factories, machines, schools). As shown by point *e*, the marginal social cost equals the marginal benefit at 2,000 square feet, so that's the socially efficient quantity of housing. For housing consumption above this level, the marginal benefit is less than the opportunity cost, meaning that the money would be more efficiently spent on factories, machines, or schools.

The mortgage subsidy causes inefficiency because it creates a gap between the private and social cost of housing. Suppose the marginal tax rate for the household is 28 percent, so each dollar spent on housing cuts taxes by $0.28 and thus has a

FIGURE 15–6 The Mortgage Subsidy Increases Housing Consumption



The socially efficient level of housing consumption is shown by point *e*, where the marginal social benefit equals the marginal social cost. The mortgage subsidy reduces the private cost, leading to excessive consumption. The shaded triangle shows the social loss.

net cost to the consumer of only \$0.72. In Figure 15–6, the marginal private cost is \$0.72, and housing consumption increases to 2,420 square feet. This is socially inefficient because it violates the marginal principle: The marginal benefit from the extra housing is less than the opportunity cost (\$1). By spending more on housing, society has less to spend on factories, machines, and schools. The shaded triangle shows the loss resulting from excessive housing consumption. The computation of market surplus is reviewed in Section 1.2 of "Tools of Microeconomics," the appendix at the end of the book.

Mortgage Subsidy and Home Ownership

The mortgage subsidy also creates a bias toward home ownership as opposed to renting. Consider Bedrock, a town where everyone lives in identical rock houses, each of which has a market value of \$100,000. There are no maintenance or repair expenses, so the only cost for a property owner is the interest payment on the money borrowed to purchase the house. If the interest rate is 8 percent per year, the annual interest cost is \$8,000, so in a competitive market with zero economic profit, the annual rent on housing would be \$8,000. Barney owns rental property and makes zero economic profit, with \$8,000 rent matching his \$8,000 interest expense. If the government allows him to deduct interest payments from his gross income, his taxable income will be zero, so he will pay no taxes.

The mortgage subsidy creates a bias toward ownership because homeowners can deduct their interest expenses even though they don't declare any income from the property. For example, suppose Wilma moves out of rental housing and buys a house for \$100,000. She pays \$8,000 interest per year and she can deduct her interest costs, so her taxable income is \$8,000 less than it was when she rented. With a 28 percent marginal tax rate, her tax bill will decrease by \$2,240, so the benefit of owning rather than renting an identical dwelling is \$2,240. As Green and Vandell (1999) show, the mortgage subsidy has been an important factor in rising homeownership rates in the last several decades—from about half of households in 1945 to about two-thirds today.

The government could eliminate the tax bias toward homeownership in one of two ways. The simple and obvious response is to eliminate the mortgage-interest deduction for homeowners. An alternative is to change how homeowners calculate their income. A homeowner's imputed rental income is defined as the income earned from owning a dwelling and renting it to yourself. Alternatively, it is the money you could earn if you rented your dwelling to someone else. In our example, Wilma's imputed rental income is \$8,000 per year. Suppose she declared \$8,000 of imputed rental income as part of her income, and then deducted her \$8,000 mortgage cost. The two items would cancel one another, and the ownership bias would disappear. Her taxable income would be the same with renting and owning, so she would be indifferent between renting and owning.

What is the rationale for the mortgage subsidy? One possibility is that it could internalize a neighborhood externality. When I paint my peeling house or weed my lawn or otherwise improve the external appearance of my house, the neighborhood looks better, and the market values of neighboring houses increase. The problem with this line of reasoning is that the mortgage subsidy applies to all elements of housing consumption, not just the elements such as exterior painting and weeding that generate neighborhood externalities. A second possible rationale is to encourage ownership to promote stable communities. But the tax breaks are concentrated among high-income households, so the subsidy provides relatively small ownership incentives for households with below-average income.

RENT CONTROL AND RENT REGULATION

During World War II, the federal government instituted a national system of rent controls, establishing maximum rents for rental properties. New York City was the only city to retain rent controls after the war. During the 1970s, rent regulations were introduced in many cities, including Boston; Washington, DC; San Francisco; and Los Angeles. In contrast with pure rent control (a fixed maximum price), a policy of rent regulation provides for annual rent increases tied to inflation, and it often allows larger price hikes to offset higher costs and guarantee a "fair" or "reasonable" rate of return on investment (Arnott, 1995). Some types of rental housing are commonly exempted from regulation, including new housing and high-rent housing. Some regulation policies permit the deregulation of prices once they reach a certain level, and others allow unrestricted price increases when tenants change.

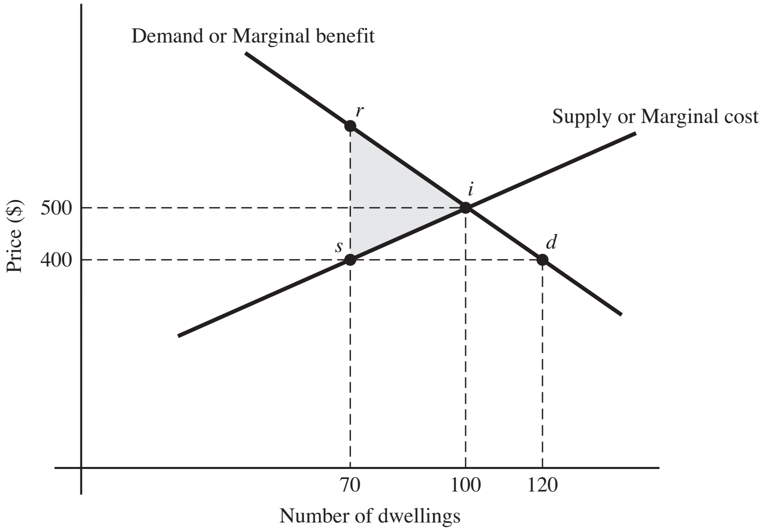
Figure 15–7 (page 507) shows the market effects of pure rent control. In the initial equilibrium shown by point *i*, the price is $500 and the quantity is 100 dwellings. If the maximum rent is set at $400, the quantity supplied decreases to 70 dwellings and the market moves to point *s*. The quantity demanded is 120 (shown by point *d*), so the excess demand is 50 dwellings. The exchanged quantity of housing drops from 100 to 70 dwellings.

We can use the notion of market surplus to measure the inefficiency of rent control. The concept of market surplus is reviewed in Section 2.5 of "Tools of Microeconomics," the appendix at the end of the book. Rent control decreases the quantity of housing below the equilibrium quantity and reduces the surplus of the housing market. The loss in value (the deadweight loss) is shown by the area between the demand (marginal benefit) curve and the supply (marginal cost) curve from the rent-control quantity of 70 to the market-equilibrium quantity of 100 (the shaded triangle).

The winners from rent control are the consumers who manage to get one of the rent-controlled dwellings at the artificially low price. The gains of the occupants are diminished by three responses to rent control:

- **Search costs.** Given the excess demand for housing, vacancy rates will be relatively low, and it will take a longer time to find a dwelling. The higher search cost at least partly offsets the benefit of a lower price.
- **Key money.** Competition among consumers may increase the effective price above the controlled price. In some rent-control cities, property owners charge tenants a large sum of money for the keys to the dwelling, and others impose

FIGURE 15–7 The Market Effects of Pure Rent Control



Pure rent control (a maximum rent at $400) decreases the quantity of housing supplied from 100 to 70 and causes permanent excess demand for housing at the maximum price: The quantity demanded (point *d*) exceeds the quantity supplied (point *s*). The lost surplus is shown by the shaded triangle *ris*.

nonrefundable deposits. In the city of Cairo, payments of key money increase the effective rent from about 38 percent of the market price to about 71 percent (Malpezzi, 1998).

- **Reduced quality.** The lower price of rental housing decreases the owners' incentives to maintain and repair property, so the quality of housing decreases.

Among the losers from rent control are households that are displaced by the policy. Rent control decreases the quantity of housing supplied as rental properties are converted to other uses (condominiums or commercial property) or retired from the housing market. In addition, households displaced by rent control in one municipality bid up the price of housing in other municipalities, generating costs for people outside the rent-control city.

Property owners lose under rent control. A decrease in the allowable rent decreases the profit that can be earned on a property, decreasing the market value of the property. In a study of New York City's rent-control program, Olson (1972) concluded that the loss of property owners was about twice as large as the gain to consumers. Toronto's rent-control program decreased the market value of apartment buildings by about 40 percent over a five-year period (Smith and Tomlinson, 1981).

How does the analysis of rent regulation differ from the analysis of pure rent control shown in Figure 15–7? The key difference between control and regulation is the flexibility of the regulated price, which increases with inflation and may increase with production cost. In addition, the price regulations may apply only while a particular tenant occupies the dwelling, allowing the owner to adjust the price with each new tenant. As a result, the price gap (the equilibrium price minus the regulated price) is likely to be smaller under regulation, so the effects on the quantity and quality of housing are likely to be smaller.

SUMMARY

This chapter explores the effects of housing policies that assist low-income households and subsidize the mortgage costs of middle-income and high-income households. Here are the main points of the chapter:

1. Public housing is more costly to produce than private housing and limits the choices of recipients.
2. Housing vouchers give recipients more options, but they increase the demand for housing, increasing prices for recipients and nonrecipients.
3. The mortgage subsidy increases housing consumption beyond the socially efficient level, and the bulk of the benefits go to high-income households.
4. In contrast with rent control (a fixed maximum price), a policy of rent regulation allows greater flexibility in the regulated price of housing.

APPLYING THE CONCEPTS

For exercises that have blanks (_______), fill each blank with a single word or number. For exercises with ellipses (…), complete the statement with as many words as necessary. For exercises with words in square brackets ([increase, decrease]), circle one of the words.

1. **Bang per Buck of Low-Income Housing Tax Credits?**
   Suppose you build a low-income house that qualifies for the low-income housing tax credit. The cost of building the house is $100,000.
   a. Over a 10-year period, your tax credit is _______, computed as…
   b. Based on Quigley's results, the market value of the house is _______, computed as…
   c. The market value is less than the building cost because…
2. **Nil Effects of Public Housing**
   According to Mr. Wizard, "If my single assumption about the market for privately produced housing is correct, the long-run market effects of public housing are nil. The building of 100 public housing dwellings will affect neither the equilibrium price of housing nor the equilibrium quantity of housing."
   a. Mr. Wizard's assumption is that the _______ curve for privately produced housing is _______.
   b. Illustrate your answer to (*a*) with a graph showing the effect of 100 public housing dwellings on the market for privately produced housing.

3. Computing the Market Effects of Vouchers

Suppose the government issues housing vouchers that increase the demand for housing in a city by 10 percent. The initial price of housing is $400. The price elasticity of supply of housing is 4.0 and the price elasticity of demand is −1.0.

a. Depict graphically the market effects of the housing vouchers.

b. Using the price change formula discussed in Chapter 5, the increase in demand ________ the equilibrium price of housing by ________ percent (from $400 to ________), computed as…

4. Filtering and Price Effects

The study by Susin (2002) suggests that vouchers increase the price of low-income housing by about 16 percent and increase the price of middle-income housing by about 3 percent. Suppose that her results for "low-income housing" apply to moderate-quality housing and her result for "middle-income housing" applies to medium-quality housing. The initial price of moderate quality is $500 and the initial price of medium quality is $1,000. Use graph like Figure 14–5 to show these effects.

5. Harmed by Vouchers?

Suppose that the poor households of city Spillover are enrolled in a voucher program. You are a middle-income renter in the city.

a. You will be [helped, harmed] by the voucher program because…

b. The help or harm from a voucher program will be relatively large if the elasticity of ________ is relatively [large, small].

c. How would your answer to (*a*) change if you owned your home?

6. The Price Effects of Vouchers and Recipient Welfare

Using Figure 15–4 as a starting point (*v* is the voucher point and *j* is the public-housing point), suppose the $300 program increases the price of housing, from $1 to $1.50 per unit of housing.

a. Draw the voucher budget line with a housing price of $1.50.

b. For a household that chooses 300 units of housing service, its spending on all other goods is ________, computed as…

c. The slope of the budget line is ________ per unit of housing, compared to ________ if vouchers don't affect housing prices.

d. Given the price effects of a voucher program, the typical recipient would be [better, worse] off with public housing because…

7. Ask Dr. Elastic

Suppose that you are interested in maximizing the welfare of the poor. You must choose between vouchers and subsidized public housing. You can ask Dr. Elastic, who knows every economic elasticity ever measured, a single question.

a. Your question is…

b. Vouchers will be better if the answer to the question is a [large, small] number.

c. Subsidized public housing will be better if the answer to the question is a [large, small] number.

8. **Deadweight Loss from the Mortgage Subsidy**
   Suppose the marginal value of a square foot of factory space is constant at \$1.00. The marginal benefit of a square foot of housing space is \$1.00 for 1,000 square feet and \$0.80 for 1,200 square feet. Suppose the government provides a 20 percent mortgage subsidy, cutting the net price of housing to consumers from \$1.00 to \$0.80 per square foot.
   a. Use a graph like Figure 15–6 to show the deadweight welfare loss resulting from the mortgage subsidy.
   b. The deadweight loss is _______, computed as…
   c. For the typical consumer, the consumer-surplus gain from the subsidy is _______, computed as…
   d. The revenue loss for the government is _______, computed as…
   e. The revenue loss is [greater, less] than the consumer-surplus gain because…
9. **Fruit Fly Economist**
   According to Frieda, the talking fruit fly, "A rent-control law implemented today will simply redistribute income from property owners to consumers. I will never see any deadweight loss from the rent-control program."
   a. After googling "fruit fly lifespan," draw a supply/demand graph to illustrate the logic of Frieda's statement.
   b. Frieda's statement is literally correct because…
   c. Contrast Frieda's view with the view of Dumbo, the talking elephant. Draw Dumbo's graph of the housing market, and show the the deadweight loss from rent control.
10. **Rent Control and Elasticity**
   Consider the effects of rent control in two cities, each of which imposes rent control with a maximum price of \$400, compared to an equilibrium price of \$500. In Elastic City, the price elasticity of supply of housing is 5.0. In Rigid City, the supply elasticity is 0.40.
   a. Compared to Elastic City, Rigid City will experience a [smaller, larger] reduction in the quantity of housing supplied and a [smaller, larger] deadweight loss from rent control.
   b. Illustrate your answers with two graphs, one for each city.

REFERENCES AND ADDITIONAL READING

1. Apgar, William. "Which Housing Policy Is Best?" *Housing Policy Debate* 1 (1990), pp. 1–32.
2. Arnott, Richard. "Time for Revisionism on Rent Control?" *Journal of Economic Perspectives* 9 (1995), pp. 99–120.
3. Burman, Leonard. "Low Income Housing Credit." In *Encyclopedia of Taxation and Tax Policy,* eds. Joseph Cordes, Robert Ebel, Jane Gravelle. Washington DC: Urban Institute Press, 1999, pp. 263–65.
4. Burt, Martha R. *Over the Edge: The Growth of Homelessness in the 1980s.* New York: Sage, 1992.

5. Connerly, Charles E., and Y. Thomas Liou. "Community Development Block Grants." In *The Encyclopedia of Housing,* ed. Willem Van Vliet. Thousand Oaks, CA: Sage Publications, 1998.
6. DiPasquale, Denise, Dennis Fricke, and Daniel Garcia-Diaz. "Comparing the Costs of Federal Housing Assistance Programs." *Federal Reserve Bank of New York Policy Review* (June 2003), pp. 147–66.
7. Early, Dirk. "The Role of Subsidized Housing in Reducing Homelessness: An Empirical Investigation Using Micro-data." *Journal of Policy Analysis and Management* 17 (1998), pp. 687–96.
8. Early, Dirk. "A Microeconomic Analysis of Homelessness: An Empirical Investigation Using Choice-Based Sampling" *Journal of Housing Economics* 8 (1999), pp. 312–27.
9. Goering, John. "The Impacts of New Neighborhoods on Poor Families: Evaluating the Policy Implications of the Moving to Opportunity Demonstration." *Federal Reserve Bank of New York Policy Review* (June 2003), pp. 113–40.
10. Green, Richard K., and Stephen Malpezzi. *A Primer on U.S. Housing Markets and Housing Policy*. Washington, DC: Urban Institute Press, 2003.
11. Green, Richard K., and Kerry D. Vandell. "Giving Households Credit: How Changes in the U.S. Tax Code Could Promote Homeownership." *Regional Science and Urban Economics* 29 (1999), pp. 419–44.
12. Gyourko, Joseph, and Peter Linneman. "Equity and Efficiency Aspects of Rent Control: An Empirical Study of New York City." *Journal of Urban Economics* 26 (1989), pp. 54–74.
13. Honig, Marjorie, and Randall K. Filer. "Causes of Intercity Variation in Homelessness." *American Economic Review* 83 (1993), pp. 248–55.
14. Malpezzi, Stephen. "Welfare Analysis of Rent Control with Side Payments: A Natural Experiment in Cairo, Egypt." *Regional Science and Urban Economics* 28 (1998), pp. 773–95.
15. Malpezzi, Stephen, and Kerry Vandell. "Does the Low-Income Housing Tax Credit Increase the Supply of Housing?" *Journal of Housing Economics* 11 (2002), pp. 360–80.
16. Murray, Michael. "Subsidized and Unsubsidized Housing Stocks 1935–1987: Crowding Out and Cointegration." *Journal of Real Estate Finance and Economics* 18 (1999), pp. 107–24.
17. National Coalition for the Homeless, *Who is Homeless? NCH Fact Sheet #3* (June 2006). http://www.nationalhomeless.org.
18. Office of Management and Budget. *Budget of the United States Government Fiscal Year 2002,* Table 8–1. Washington, DC: 2002.
19. O'Flaherty, Brendan. *Making Room: The Economics of Homelessness.* Cambridge, Mass: Harvard University Press (1996).
20. Ohls, James C. "Public Policy toward Low-Income Housing and Filtering in Housing Markets." *Journal of Urban Economics* 2 (1975), pp. 144–71.
21. Olson, Edgar. "An Econometric Analysis of Rent Control." *Journal of Political Economy* 80 (1972), pp. 1081–1100.

22. Poterba, James M. “Taxation and Housing: Old Questions, New Answers.” *American Economic Review* 82 (1992), pp. 237–42.
23. Quigley, John. “A Decent Home: Housing Policy in Perspective.” *Brookings-Wharton Papers on Urban Affairs* (2000), pp. 53–99.
24. Quigley, John, and Steven Raphael. “Is Housing Affordable? Why Isn’t It More Affordable?” *Journal of Economic Perspectives* 18 (2004), pp. 191–214.
25. Smeeding, Timothy M. “Alternative Methods for Evaluating Selected In-Kind Transfer Benefits and Measuring Their Effect on Poverty.” Technical Paper no. 50. Washington, DC: U.S. Bureau of the Census, 1982.
26. Smith, Lawrence, and Peter Tomlinson. “Rent Control in Ontario: Roofs or Ceilings.” *American Real Estate and Urban Economics Journal* 9 (1981), pp. 93–114.
27. Stegman, Michael A. “The Excessive Costs of Creative Finance: Growing Inefficiencies in the Production of Low-Income Housing.” *Housing Policy Debate* 2 (1991), pp. 357–73.
28. Struyk, Raymond J. “Comment on William Apgar’s ‘Which Housing Policy Is Best?’” *Housing Policy Debate* 1 (1990), pp. 41–51.
29. Susin, Scott. “Rent Vouchers and the Price of Low-Income Housing.” *Journal of Public Economics* 83 (2002), pp. 109–52.

PART SIX

Local Government

The final part of the book explains the role of local government and explores how citizens respond to local taxes and intergovernmental grants. As explained in Chapter 16, local governments provide local public goods, deal with natural monopolies, and respond to externalities. The chapter also explores local decision making through elections and explains why majority rule is unlikely to generate socially efficient choices. Chapter 17 looks at the revenue side of local government, focusing on the two largest revenue sources: the property tax and intergovernmental grants. As we'll see, the person who pays the property tax in a legal sense may shift the tax onto other people, so the economic burden is different from the legal burden. Local governments respond to intergovernmental grants by cutting taxes and shifting resources into other programs, so part of a grant is spent on other public goods and private goods.

CHAPTER 16

The Role of Local Government

Pedro: Do you think people will vote for me?
Napoleon Dynamite: Heck yes! I'd vote for you.
Pedro: Like what are my skills?
Napoleon Dynamite: Well, you have a sweet bike. And you're really good at hooking up with chicks. Plus you're like the only guy at school who has a mustache.

—FROM THE MOVIE *NAPOLEON DYNAMITE* (2004)

Democracy is the worst form of government except all the others that have been tried.

—WINSTON CHURCHILL

This chapter provides an overview of the local public sector. After presenting the facts on local government, we explore the role of local government in a federal system of government. We'll see why products such as public schooling, public safety, parks, and transit systems are produced by local governments rather than private firms or higher levels of government. We'll also see why voting with majority rule is unlikely to generate socially efficient decisions and explore different responses to the inefficiency of majority rule.

As shown in Table 16–1 (page 412), there are more than 87,000 local governments in the United States. In terms of total expenditures, the most important types of local governments are municipalities and school districts (each about a third of local government expenditures) and counties (about a quarter of expenditures). A special district serves a single function, such as fire protection, natural-resource management, or the administration of housing or community-development programs.

Table 16–2 (page 412) shows the per-capita spending for local government in general and municipalities. Education is the dominant program for local government, with nearly half of total expenditures, 90 percent of which goes to elementary and secondary education. The spending of municipalities is more evenly divided, with the largest expenditures on police, education, highways, sewerage, and fire protection.

TABLE 16–1 Types of Local Governments in 2002

Type of Local Government	Number
County	3,034
Municipal	19,429
Township and town	16,504
School	13,506
Special	35,052
Total	87,525

Source: U.S. Bureau of the Census. *2002 Census of Governments*.

TABLE 16–2 Expenditures per Capita for Local Government, 2002

	Local Government	Municipalities
Education	$1,537	$125
Police protection	196	129
Governmental administration	188	76
Hospitals	179	35
Highways	157	78
Interest on general debt	156	56
Public welfare	141	35
Sewerage	107	66
Health	104	21
Housing and community development	99	44
Fire protection	92	63
Parks and recreation	89	55
Correction	65	11
Solid waste management	58	34
Natural resources	19	1
Protective inspection and regulation	13	9
Parking facilities	4	4
Transit subsidies	1	1
Total	3,206	843

Source: U.S. Bureau of the Census. *2002 Census of Governments.*

THE ROLE OF LOCAL GOVERNMENT

What is the role of local government in the market economy? Musgrave and Musgrave (1980) distinguish between three roles for government:

1. **Stabilization.** The government uses monetary and fiscal policy to control unemployment and inflation.
2. **Income redistribution.** The government uses taxes and transfers to alter the distributions of income and wealth.
3. **Resource allocation.** The government makes decisions about what to produce and how to produce it. When the government actually produces a particular good or service, it makes these resource allocation decisions directly. When

the government subsidizes or taxes private activities, it influences the resource allocation decisions of the private sector.

How do local governments fit into this three-part scheme of governmental activity?

The national government has assumed the responsibility for stabilization policy for two reasons. First, although each local government could print its own money and execute its own monetary policy, such a system would be chaotic. Instead, the national government prints the money and manages a national monetary policy. Second, because a large fraction of local income is spent on goods produced outside the local area, local monetary and fiscal policies would be relatively weak and ineffective. Fiscal policy is more effective at the national level because a relatively small fraction of national income is spent on imports.

Consider next the distribution role of government. Local attempts to redistribute income will be frustrated by the mobility of taxpayers and transfer recipients. Suppose that a city imposes a tax on its wealthy citizens and provides transfer payments to the poor. To escape the tax, some wealthy households will leave the city, causing a decrease in total tax revenue. At the same time, some poor households will migrate to the relatively generous city, causing a decrease in the transfer payment per recipient. In combination, the flight of the wealthy and the migration of the poor will weaken the city's redistribution program. A national redistribution program is more effective because there is less mobility between nations than between cities.

The third role of government is resource allocation, which involves decisions that determine how an economy's resources are allocated to different goods and services. As Table 16–2 shows, local governments are responsible for providing several goods and services, including education, highways, police and fire protection, parks, and sewers. In the next three parts of the chapter, we explore three sorts of resource allocation by local government: providing local public goods, dealing with natural monopoly, and internalizing externalities.

LOCAL PUBLIC GOODS: EQUILIBRIUM VERSUS OPTIMUM

A local public good has three characteristics. First, it is nonrivalrous: The fact that one person benefits from a public good doesn't reduce the benefit for someone else. In contrast, a private good such as a hot dog is rivalrous because only one person can consume it. Many of the goods provided by local governments are impure or congestible in the sense that if enough people use the good, each person reduces the benefit to others. One example of an impure public good is a city park; if enough people use the park, they get in each others' way, with frisbees flying into birthday cakes. As we saw earlier in the book, streets and highways are subject to congestion during peak travel periods.

The second characteristic of a local public good is that it is nonexcludable. In other words, it is impossible or impractical to exclude people who do not pay for the good. Consider the park example. Although it may be possible to charge everyone for using a park and exclude people who don't pay, it would be costly to fence off the parks, install turnstiles, and monitor their use. Although it might be possible to bill people for fire and safety services, a system of user fees for the fire department and police would violate most people's notions of equity and fairness.

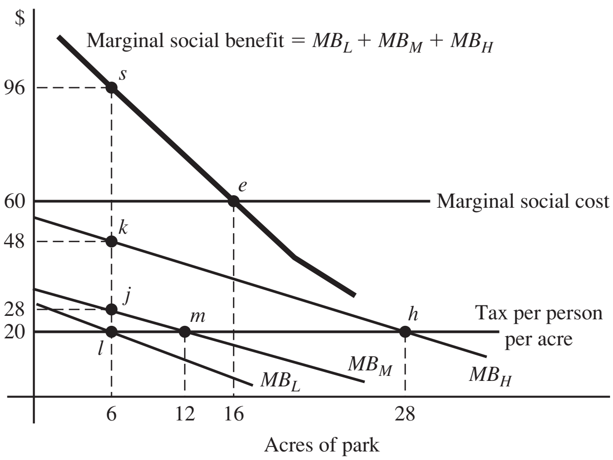
The third characteristic of a local public good is that its benefits are confined to a relatively small geographical area—a municipality or a metropolitan area. Unlike national defense, which generates benefits for the entire nation, most of the benefits of the local police force and local fire department go to local citizens. Similarly, local citizens get most of the benefits from local streets and highways. The appropriate size of the jurisdiction is determined by the "localness" of the public good—the geographical extent of the benefits from the good. The more widespread the benefits, the larger the jurisdiction required to contain all the beneficiaries.

The Efficient Quantity of Local Public Goods

In the chapter on neighborhood choice, we discussed the provision of city parks in a three-person city. Figure 16–1 uses Figure 8–2 as a starting point. We have three citizens who vary in their demand for park acres: Lois has a low demand (with marginal-benefit curve MB_L); Marian has medium demand (MB_M); Hiram has high demand (MB_H).

The optimum level of the local public good is the quantity at which the marginal social benefit equals the marginal social cost. The social benefit of a city park is the sum of the benefits going to the citizens, and the marginal social benefit is the sum of the individual marginal benefits. In Figure 16–1, the marginal social benefit

FIGURE 16–1 Optimum versus Equilibrium Local Public Good



The marginal social benefit is the sum of the marginal private benefits of the three citizens. The optimum acreage is 16 acres shown by the intersection of the marginal social benefit curve and the marginal social cost curve (point *e*). Under majority rule, the equilibrium is the preferred choice of the median voter (12 acres, shown by point *m*).

of the sixth acre is $96 (point *s*), equal to the sum of the marginal benefits of $20 for Lois, $28 for Marian, and $48 for Hiram. Similarly, for other park sizes, we add the individual marginal benefits to get the marginal social benefit. The marginal social-benefit curve is the vertical sum of the individual marginal-benefit curves.

We can use the marginal principle to identify the socially efficient park size. The marginal principle is reviewed in Section 1.1 of "Tools of Microeconomics," the appendix at the end of the book. At the efficient level, the marginal social benefit of park acreage equals its marginal social cost. In Figure 16–1, the marginal cost is $60 per acre, so the optimum acreage is 16 acres. For any amount less than this, citizens in the city would collectively be willing to pay more than $60 per additional acre, so an increase in size would increase social welfare. For example, suppose the city starts at six acres. At this point, the willingness to pay for parks (from the social-benefit curve) is $96 and the cost is only $60, so another acre of park would generate a net gain. In contrast, for any amount exceeding 16 acres, the aggregate willingness to pay for another acre would be less than the social cost, so a smaller park would be more efficient.

The Median Voter Picks the Equilibrium Quantity

As we saw in the chapter on neighborhood choice, the equilibrium quantity under majority rule is the preferred quantity of the median voter. If the government imposes a common head tax of $20 per person per acre of parks, each citizen faces a marginal cost of $20 per acre, and Marian the median voter prefers 12 acres of parks (point *m*). If the government holds a series of pair-wise elections between the preferred quantities of the three citizens, the median voter will win. The median voter always wins because she can get one other person to vote against any other option.

In Figure 16–1, the voting equilibrium is not the same as the optimum. The median voter prefers a quantity less than the optimum, and so the city chooses an inefficiently small park. If the city had a direct election of the median preference versus the optimum, the median preference would win because Lois would join Marian to defeat the optimum. The power and inefficiency of the median-voter result can be seen clearly by imagining that the marginal benefit of the high-demand consumer doubles. Such a change would increase the optimum park acreage but would not affect the voting equilibrium because the preference of the median voter hasn't changed.

Tiebout Model: Voting with Feet

Most metropolitan areas in the United States have dozens of municipalities, school districts, and other local governments. When citizens disagree about how much of a local good to provide, they can "vote with their feet," moving to jurisdictions with like-minded people. One of the implications of the Tiebout model is that interjurisdictional mobility (voting with feet) may prevent the inefficiencies associated with majority voting.

The simple version of the Tiebout model is based on five assumptions about local government and location choices:

1. **Municipal choice.** A household chooses the municipality (or school district or other local jurisdiction) that provides the ideal level of local public goods. There are enough municipalities to ensure that every household finds the perfect jurisdiction.
2. **Perfect information and mobility.** All citizens have access to all relevant information about the alternative municipalities, and moving is costless.
3. **No interjurisdictional spillovers.** There are no spillovers (externalities) associated with local public goods: All the benefits from local public goods accrue to citizens within the municipality.
4. **No scale economies.** The average cost of production is independent of output.
5. **Head taxes.** A municipality pays for its public goods with a head tax: If you have a head, you pay the head tax.

Under the Tiebout process, households will sort themselves into municipalities according to their demand for parks. Suppose three low-demand citizens form a municipality called Loisville. The marginal social benefit of six acres is three times \$20, or \$60, the same as the marginal social cost of park acreage. When each voter pays a tax of \$20 per acre, they will all prefer six acres, so they will vote unanimously for the optimum. Similarly, if three Marians form a municipality, they will choose the optimum for medium demanders—12 acres. This type of sorting eliminates the inefficiencies of majority rule because everyone in a homogeneous municipality has the same preferred level of the local public good.

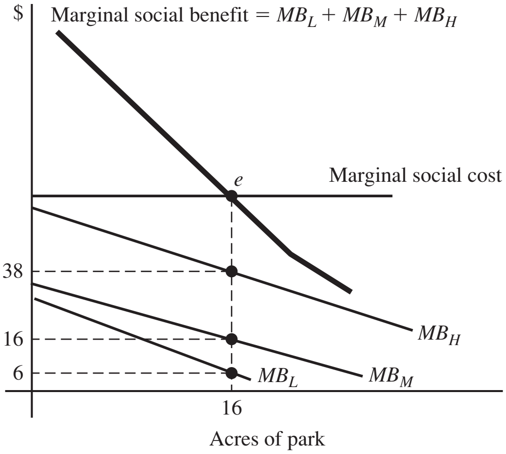
There is evidence that citizens sort themselves with respect to the demand for local public goods. Heikkila (1996) shows that communities in Los Angeles are relatively homogeneous with respect to the demand for local public goods. Fisher and Wassmer (1998) show that the greater the variation across households in the underlying demand for local public goods in a metropolitan area, the larger the number of municipalities and school districts in the metropolitan area. At the international level, the greater the ethnic diversity of a nation, the more decentralized its public sector (Panizza, 1999).

Benefit Taxation

The Tiebout response to diversity in demand for local public goods is to eliminate diversity by forming homogeneous municipalities. An alternative response is to match the diversity in demand with diversity in tax liabilities. Under the Lindahl approach (named after economist Erik Lindahl), taxes are proportional to the willingness to pay for local public goods.

Figure 16–2 (page 417) shows how Lindahl or benefit taxes work. Suppose the government knows the marginal-benefit curves of its citizens and can determine the optimum level of the public good—equal to 16 acres of parks in our example. The government allocates the cost of the public good to its citizens according to their willingness to pay (marginal benefit). Hiram's tax liability is \$38 per acre, while Marian's is

FIGURE 16–2 Lindahl or Benefit Taxation



Under benefit taxation, each household pays a per-acre tax equal to its marginal private benefit for the socially efficient park size (16 acres), and everyone prefers the efficient size. The benefit tax is \$6 per acre for Lois, \$16 for Marian, and \$38 for Hiram.

\$16, and Lois's is \$6. Faced with a marginal cost per acre of \$38, Hiram's preferred park size is 16 acres, the optimum size. Similarly, Lois has a marginal cost of \$6, and prefers the optimum size too. Under benefit taxation, citizens with larger benefits pay higher taxes, and diversity in demand is matched by diversity in tax liabilities.

Is the benefit principle practical? One problem is the government doesn't know its citizens' marginal-benefit curves, so it can't precisely determine the appropriate taxes. The government can't simply ask its citizens to reveal their willingness to pay because each citizen has an incentive to understate their willingness to pay—and thus pay lower taxes. But for some public goods such as fire protection or public safety, the benefit from local public goods may be roughly proportional to property value, so a property tax could serve as a rough benefit tax. Similarly, if the benefits from local public goods increase with income, an income tax could serve as a rough benefit tax.

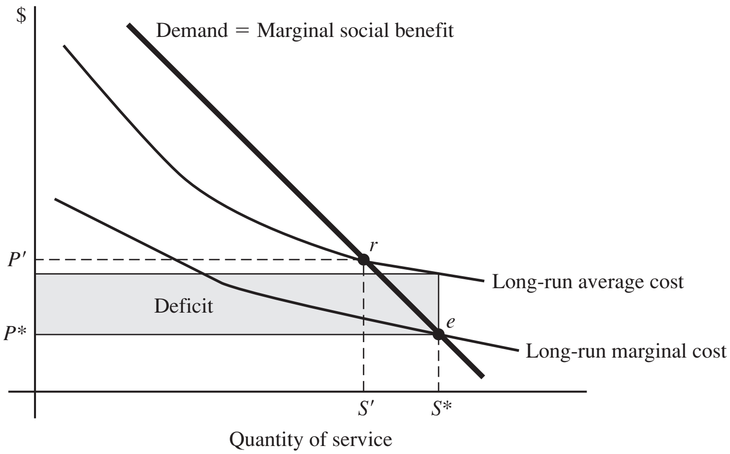
NATURAL MONOPOLY

Local governments operate natural monopolies such as water systems and waste-disposal systems. A natural monopoly occurs if the increasing returns to scale in production are large relative to the demand for the product. Recall the fourth principle of urban economics:

Production occurs with increasing returns to scale

The provision of sewage services involves a large network of pipes and a large treatment plant, so the indivisible inputs are large and expensive. In Figure 16–3,

FIGURE 16–3 Natural Monopoly in Sewage Services



Sewage service is a natural monopoly, with a negatively sloped long-run average cost curve. The socially efficient output occurs where the marginal social benefit equals the marginal cost (point *e*). At this quantity, the price is less than the average cost, generating a deficit. The government can provide the services and cover the deficit with general tax revenue. Alternatively, the government could regulate a private firm.

the long-run average-cost curve for sewage services is negatively sloped over a wide range of output, reflecting the substantial costs of these indivisible inputs. As output increases, the average cost decreases as the cost of these indivisible inputs is spread over larger quantities.

We can use the marginal principle to identify the socially efficient level of sewage service. The marginal principle is reviewed in Section 1.1 of "Tools of Microeconomics," the appendix at the end of the book. The efficient output is the quantity at which the marginal benefit of service equals the marginal cost. The demand curve is a marginal-benefit curve, so the efficient point is where the demand curve intersects the marginal-cost curve (point *e*), and the efficient quantity is S^* units of service. One problem is that a firm producing the efficient output will lose money. To get consumers to purchase S^* units of sewage service, the firm must charge a price of P^*, which is less than the average cost of producing S^*. The shaded area shows the deficit that occurs at the socially efficient quantity of output.

The government has several options in responding to the deficit problem. First, the government could produce the service itself, charging a price of P^* and covering the deficit with general tax revenue. Second, the government could subsidize a private firm to provide sewage service: The firm would charge the efficient price (P^*) and the city would cover the firm's deficit with tax revenue. Third, the government could allow the private firm to charge P' instead of P^*. At the higher price, the quantity sold would be S' and the average cost would equal the price, so the firm would cover all of its cost. Under this scheme, the output (S') is less than the socially efficient output.

Local governments have the same options for other natural monopolies. In the chapter on mass transit, we explored the pricing and regulation of mass transit and discussed transit deficits. For other natural monopolies such as water systems and solid-waste disposal, a local government can either produce the service itself or regulate a private producer.

EXTERNALITIES

Another role for government is to internalize externalities. Recall the third axiom of urban economics:

Externalities cause inefficiency

In earlier chapters, we have seen several sorts of externalities and explored ways to internalize them. The three externalities from driving—congestion, pollution, and traffic accidents—can be internalized with taxes on driving. A tax on peak-period driving internalizes the congestion externality, while a tax on pollution internalizes pollution externalities. A per-mile tax that depends on the age of the driver can internalize accident externalities. In this chapter, we consider externalities that come from education and public safety (police and fire protection).

Public Education Externalities and Vouchers

As we saw earlier in the book, education generates external benefits because it makes people better team workers, improves the democratic process, and decreases crime. One option for government is to take responsibility for providing education. A system of free compulsory education could encourage citizens to consume more education. An alternative approach is to subsidize private education, using tax credits or education vouchers to cover part or all the costs associated with private education.

Under a pure education-voucher system, each child is issued a voucher or coupon that can be used to pay for either public or private schooling. The face value of the voucher would be equal to the current cost per pupil for public schools (for example, $6,000), allowing a family to choose either the public school or a private school charging up to $6,000. The school would collect vouchers from its patrons and redeem them from the state government. To qualify for the voucher program, a school would be required to teach basic cognitive skills and civics, and admit students without regard to race, sex, or religion.

What are the possible consequences of education vouchers? If vouchers force public schools to compete with private schools for students, this competition could make public schools more efficient and more responsive to parental concerns. If so, achievement could increase, even for the students who remain in public schools. The opponents of vouchers suggest that wider school choice is likely to increase segregation with respect to income, race, and academic ability. Although a parent who receives a voucher can pick any school—public or private—not all parents will exercise this option. It appears that parents with the most education and the highest incomes are more likely to switch schools (Levin, 1997), so the peer environment—and achievement level—of students from low-income families may deteriorate.

Most proposals for education vouchers would provide vouchers for low-income families only. In recent years, there have been experiments with income-targeted vouchers in Milwaukee; Cleveland; New York; Dayton; and Washington, DC. By targeting low-income families, the voucher programs do not cause the sort of income segregation that would occur with a universal program. In fact, if the face value of the voucher is high enough, it would encourage private schools to accept more low-income students, promoting integration rather than segregation.

How do targeted vouchers affect achievement? Using data from Milwaukee, Rouse (1998) shows that students in the voucher program had higher math test scores but about the same reading scores. In addition, low-income students attending special public schools with reduced classes did just as well or better than the students who used vouchers to attend private schools. This suggests that the achievement advantage of voucher (private) schools comes in part from their smaller classes.

As we saw in our earlier discussion of the education production function, the teacher is a key factor in student achievement. As shown by Rivkin, Hanushek, and Kain (1998), most of the differences between schools are explained by differences in the quality of teachers, not by differences in school organization or other education resources. What really matters are teachers, so vouchers could increase achievement if the greater competition for students causes schools to hire better teachers.

Externalities from Public Safety Programs

Consider next the externalities that result from public safety. Spending on police services generates positive and negative externalities. Both externalities result from the fact that criminals are mobile: They can move from one jurisdiction to another.

- **Capturing externality.** The positive externality occurs when one municipality uses resources to capture a criminal. By getting a criminal off the street, the municipality generates benefits for surrounding municipalities: The marginal social benefit of police spending exceeds the marginal local (municipal) benefit.
- **Chasing externality.** The negative externality occurs when a municipality's crime-fighting activities cause a criminal to move to another jurisdiction. In this case, police spending just moves crime around, so the marginal local benefit of police spending exceeds the marginal social benefit.

The greater the mobility of criminals, the larger the jurisdiction required to contain all the people who are affected by crime fighting activities. In the United States, the typical response to these positive and negative externalities is to provide police services through municipal governments.

The other public safety service, fire protection, also generates externalities. Fires can spread from one house to another, so the marginal social benefit of fire protection exceeds the marginal private benefit. In most metropolitan areas, fire protection is provided by local governments, while some municipalities contract with private firms to provide fire protection.

FEDERALISM AND METROPOLITAN GOVERNMENT

Under the federal system of government, the responsibility for providing public goods is divided between the national, state, and local governments. Some goods, such as defense and space exploration, are provided at the national level. Others, such as education and police protection, are provided at the local level. Oates (1972) discusses the advantages and disadvantages of the local provision of public goods.

1. **Diversity in demand.** As we saw in the discussion of the Tiebout model, local governments can accommodate diverse demands for local public goods and thus promote efficiency.
2. **Externalities.** For some locally provided products, benefits spill over to people outside the municipality or school district. In this case, local voters will ignore the benefits of outsiders, so they will make inefficient choices.
3. **Scale economies.** If there are scale economies in the provision of public goods, a system of small local governments has a relatively high production cost.

The local provision of a public good is efficient if the advantages outweigh the disadvantages. In other words, local provision is efficient if (1) diversity in demand is relatively large, (2) externalities are relatively small in a geographic sense, and (3) scale economies are relatively small.

What are the facts on scale economies in the provision of local public goods? There have been dozens of studies of the relationship between production costs and jurisdiction sizes. The evidence suggests that there are moderate scale economies in the provision of water and sewage services. Because these services are capital intensive, average cost decreases as population increases. In contrast, studies of other local public goods (police protection, fire protection, schools) suggest that scale economies are exhausted with a relatively small population—about 100,000. Many small cities use intergovernmental contracts and joint service contracts to join forces and exploit scale economies in the provision of public services.

The most important trade-off associated with local service provision is between diversity of demand and externalities. Metropolitan government will be more efficient than municipal government if interjurisdictional spillovers are large relative to diversity in demand. In this case, the advantages of a small local government (the ability to accommodate diverse demands for local public goods) are relatively small, and the disadvantages (the inefficiencies associated with externalities that cross municipal boundaries) are relatively large. Therefore, a metropolitan system of government will be more efficient.

One solution to the spillover problem is a system of subsidies from a higher level of government. If the municipality receives a subsidy equal to the marginal external benefit of the public good, it bases its spending decisions on the marginal social benefit of the good. In the next chapter, we'll explore the effects of intergovernmental grants on local spending.

Another response to spillovers is to grant governmental bodies the power to deal with specific urban problems that cross municipal boundaries. Many economists and geographers believe that metropolitan areas—not municipalities or states—are the most important spatial units in today's economy. In the words of Anthony Downs (1998), it would be sensible to establish policy-making organizations for the entire metropolitan area

> because the various spatial sections of each metropolitan area are linked together in a series of densely interlocking networks. These networks transcend the boundaries of most individual communities but are not as intensive at the larger state level.

Among these networks are streets and highways, water systems, sewage-disposal systems, school systems, airsheds, and watersheds. Some of the problems that cross jurisdictional boundaries are highway congestion, air pollution, crime, and low educational achievement. In the current political system, the power to deal with these problems is divided among many small jurisdictions, most of which contain only a small fraction of the people affected by the problems. Two metropolitan areas—Portland, Oregon, and the Twin Cities in Minnesota—have governmental bodies with the power to deal with problems that cross municipal boundaries.

A CLOSER LOOK AT THE MEDIAN VOTER RESULT

Earlier in the book we explained the median-voter result in the context of a direct election with three citizens. In this part of the chapter, we'll take a closer look at voting, showing the general applicability of the median-voter result and its limitations. Many local jurisdictions—including most central cities—have large and heterogeneous populations, and decisions about local public goods are determined by voting with ballots, not by voting with feet.

A Series of Budget Elections

The median-voter result is applicable to a wide variety of election formats. Consider a school district that holds a series of elections to determine its budget. The district proposes a budget and holds an election in which citizens vote yes or no. If a particular budget proposal fails to receive a majority of votes, the school board decreases its proposed budget by $10 and then holds another election. This process continues until a majority of citizens vote in favor of the proposed budget. Under this election system, the school district chooses the largest budget that receives majority support.

Table 16–3 shows the preferences of the voters in the school district. The preferred budget of voter A is $49, while B's is $56, and so on. Suppose the district starts with a proposed budget of $90. The voters know that if the $90 budget fails to

TABLE 16–3 The Median Voter in a Series of School Budget Elections

Voter	Preferred Budget	Vote with $90 Budget	Vote with $80 Budget	Vote with $70 Budget
A	$49	N	N	N
B	56	N	N	N
C	63	N	N	N
D	70	N	N	Y
E	77	N	Y	Y
F	84	N	Y	Y
G	91	Y	Y	Y

get majority support, the next proposed budget will be $80. A citizen will vote yes if the $90 budget is better than an $80 budget, that is, if the citizen's preferred budget is greater than $85. For example, voter F, with a preferred budget of $84, will vote against the $90 budget because an $80 budget is closer to his preferred budget. Voter G is the only person who votes for the $90 budget, so the first budget proposal loses by a vote of 6 to 1. The second proposal ($80) fails by a vote of 4 to 3 because voters A through D, with preferred budgets less than $75, would prefer a $70 budget. In contrast, the preferred budget of the median voter ($70) wins by a vote of 4–3. The median voter is joined by voters with higher preferred budgets to approve the median voter's preferred budget.

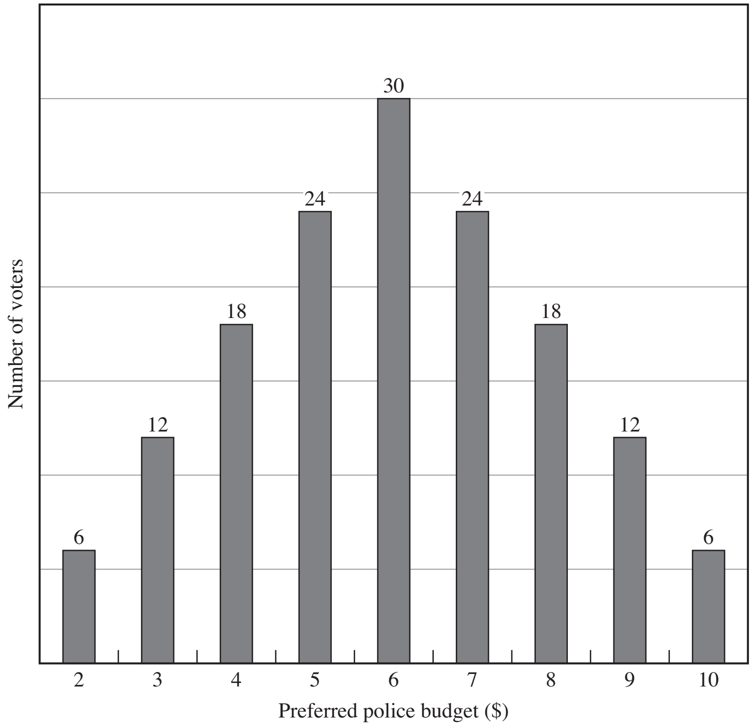
In this example, the school district decreased its proposed budget in $10 increments, from $90 down to the preferred budget of the median voter. The same result would occur if the board decreased the proposed budget in increments of $1. Similarly, the same result occurs if the school district reverses the direction of the budget sequence. If it starts out with a low budget and works its way upward, the median voter still wins.

The Median Voter in a Representative Democracy

In a representative democracy, decisions about budgets are made by elected officials. Citizens make budgetary decisions indirectly, by electing people whose budget philosophies are consistent with their own preferences. Consider a city that provides a single local public good (police services). There are two candidates for mayor, Penny (a low spender) and Buck (a big spender). The only issue in the election is the police budget, which is set by the mayor. Each citizen votes for the candidate whose proposed police budget is closest to the citizen's preferred budget.

Figure 16–4 (page 424) shows the voters' distribution of budget preferences. The horizontal axis measures the police budget, and the vertical axis measures the number of votes for a given budget. For example, six citizens have a preferred budget of $2, while 12 have a preferred budget of $3, and so on. The distribution of budget preferences is symmetric, and the median budget (the budget that splits the rest of the voters into two equal halves) is $6.

FIGURE 16–4 The Median Voter in Representative Democracy



If Penny proposes a police budget of $4 and Buck proposes a budget of $8, the election will result in a tie: The two candidates split the votes of people with a preferred budget of $6, while Penny gets voters with lower preferred budgets and Buck gets voters with larger preferred budgets. By moving toward the median budget, Penny can increase her chance of being elected. In equilibrium, both candidates propose the budget of the median voter ($6).

If the candidates take a natural approach and stake out different budget territories, the result could be a tie vote. Suppose Penny starts with a proposed budget of $4 and Buck proposes $8. Penny will get a total of 75 votes:

- 60 votes from people with preferred budgets less than or equal to $5.
- 15 of the 30 voters with a preferred budget of $6. This is halfway between the candidates' proposed budgets, so the two candidates split the votes.

The distribution of voters is symmetric, so Buck gets 75 votes too (60 from voters with preferred budgets greater than or equal to $7 and 15 from voters who prefer $6). In other words, the election results in a tie vote.

This is not an equilibrium because each candidate has an incentive to move toward the median budget. Penny could increase her chance of being elected by increasing her proposed budget to $5. If she does, she will get all 30 votes of the citizens with a $6 preferred budget because her $5 proposal is now closer than

Buck's $8. Penny will win by a vote of 90 (equal to 6 + 12 + 18 + 24 + 30) to 60 (equal to 24 + 18 + 12 + 6). If Buck responds by decreasing his proposed budget to $7, the election results in a tie vote again. Penny and Buck will continue to revise their proposed budgets until each candidate's budget is close to the preferred budget of the median voter ($6).

This example shows that the median-voter result occurs in a representative democracy. In equilibrium, both candidates propose a budget equal to the preferred budget of the median voter. Since both candidates propose the same budget, it doesn't matter which candidate actually wins the election. In either case, the median voter determines the size of the police budget.

Implications of the Median-Voter Rule

The median-voter rule has some important implications. First, as we saw earlier, there is no reason to expect voting to generate the socially efficient level of a local public good. The second implication concerns our ability to predict the outcome of an election. To predict the outcome, we need to first identify the median voter and then estimate his or her preferred budget. As a practical matter, it may be difficult to identify the median voter. One approach is to assume that the desired spending depends on income, so the person with the median income is the median voter. Of course, if the desired spending depends on other variables (household size, age, political philosophy), the predictions from this approach will be a rough approximation.

The third implication of the median-voter rule is that we can use the results of elections to estimate the elasticities of demand for local public goods. Consider two cities, one with a small police budget ($100 per capita) and a low median income ($1,000), and a second with a large police budget ($125 per capita) and high median income ($1,200). Assume that the "price" of police services (the opportunity cost of money spent on police) is the same in the two cities. The income elasticity of demand for police services is defined as the percentage change in the police budget divided by the percentage change in income. City L, with 20 percent higher income, has a 25 percent larger police budget, so the income elasticity of demand is 1.25 (25 percent divided by 20 percent). Table 16–4 summarizes the results of empirical

TABLE 16–4 Income and Price Elasticities of Demand for Local Public Goods

Public Good or Service	Price Elasticity	Income Elasticity
Total expenditures	−0.23 to −0.56	0.34 to 0.89
Education	−0.07 to −0.51	0.24 to 0.85
Parks and recreation	−0.19 to −0.92	0.99 to 1.32
Public safety (police and fire)	−0.19 to −1.0	0.52 to 0.71
Public works	−0.92 to −1.0	0.79

Source: Robert Inman. "The Fiscal Performance of Local Governments." In *Current Issues in Urban Economics*, eds. Peter Mieszkowski and Mahlon Straszheim. Baltimore: Johns Hopkins University Press, 1979.

studies based on the median-voter model. The income elasticities for most local public goods are less than 1.0.

If the price of local public goods varies across municipalities, we can use the median-voter model to draw the demand curve for local public goods and compute the price elasticity of demand. To plot the demand curve for local spending, we need information on price (the opportunity cost of local spending) and quantity (the local spending level). As shown in Table 16–4, the demands for local public goods are price-inelastic; the price elasticities are all less than or equal to 1.0 in absolute value.

Limitations of the Median-Voter Model

The median-voter model has a number of unrealistic assumptions. Although it provides a useful framework for thinking about local government decisions, three assumptions limit the model's applicability:

1. **No ideology.** Politicians care only about winning elections, so they slavishly adhere to voter preferences. Alternatively, a candidate could base her positions on ideology and use election campaigns to persuade voters that her position is the correct one, playing the role of a leader, not a follower.
2. **Single issue.** If there are several election issues (e.g., the police budget, the park budget, policies for the homeless), candidates will offer bundles or package deals to voters, and the notion of a median voter disappears.
3. **All citizens vote.** In real elections, only a fraction of eligible voters actually cast ballots. The benefit of voting will be relatively small if (a) the candidates are so close to one another that it makes little difference who wins (voter indifference), or (b) the best candidate is so far from the citizen's position that the citizen is alienated from the election process (voter alienation). If some citizens abstain from voting, the median-voter result will not necessarily occur.

SUMMARY

This chapter explores the role of local government in a federal system of government. Here are the main points of the chapter:

1. The role of local government is resource allocation—providing local public goods, operating natural monopolies, and internalizing local externalities.
2. The inefficiency of majority rule encourages citizens to vote with their feet, forming municipalities with citizens who share their preferences for local public goods.
3. An alternative to foot voting is benefit taxation, under which tax liabilities are proportional to the benefit of a local public good.
4. The local provision of a public good is efficient if (a) diversity in demand for local public goods is relatively large, (b) externalities are relatively small, and (c) scale economies are relatively small.

5. The median-voter model is applicable to sequential budget elections as well as voting for representative government.

APPLYING THE CONCEPTS

For exercises that have blanks (_____), fill each blank with a single word or number. For exercises with ellipses (. . .), complete the statement with as many words as necessary. For exercises with words in square brackets ([increase, decrease]), circle one of the words.

1. **U-Read and Blockbuster Books?**
   Why do we have public libraries? Why don't people rent books from firms, just as they rent trucks, roto tillers, and DVDs? We have U-Haul, so why not U-Read? We have Netflix, so why not Netbooks?
2. **Hattie Hates Parks**
   Consider the example shown in Figure 16–1. Suppose Lois leaves town and is replaced by Hattie, whose marginal benefit of park acreage is negative: −$16 per acre.
   a. Use a graph like the one in Figure 16–1 to show the socially efficient park size. The efficient park size is _____ acres.
   b. Predict the outcome of pair-wise elections between different park sizes.

	Votes for Smaller	Votes for Larger
0 versus 12	_____	_____
12 versus 28	_____	_____

   c. Under majority rule, the city chooses a park with _____ acres, which is [smaller, larger] than the optimum park.
   d. A Lindahl tax scheme will generate unanimous support for the socially efficient park size if Hiram pays _____ per acre, Marian pays _____ per acre, and Hattie . . .
3. **Paying for Flood Protection**
   Suppose a city could spend $2,100 on a dike that would decrease the probability of flooding in a three-house neighborhood from 0.03 to 0.02. A flood would destroy all three houses, and flood insurance is not available. The market values of the three houses are $60,000, $120,000, and $240,000.
   a. The social benefit of the dike = _____ = _____ + _____ + _____
   b. The dike is socially [efficient, inefficient] because . . .
   c. The benefit: Cost ratio is _____.
   d. Design a taxing scheme that would generate unanimous support for an efficient dike.

4. The Power of the Median Voter

Using Figure 16–1 as a starting point, suppose Hiram's willingness to pay for parks triples.

a. The optimum park acreage will [increase, decrease, not change] because . . .

b. The park acreage chosen under majority rule will [increase, decrease, not change] because . . .

5. Private Schools and Budget Elections

Using Table 16–3 as a starting point, consider the effect of private schools on a local school election. Suppose that households F and G (in Table 16–3) enroll their children in private schools. Their preferred budgets are $44 (F) and $51 (G). Predict the election outcome under two cases:

a. If F and G abstain from voting in the school elections, the winning budget will be _____ because . . .

b. If F and G vote in the elections, the winning budget will be _____because . . .

6. Voter Changes

In the example shown in Table 16–3, the winning budget is $70. Predict how the following changes will affect the winning budget. In each case, use as the starting point the distribution of voters shown in Table 16–3.

a. If four elderly voters move into the school district, with preferred budgets of $20, $25, $30, and $35, the winning budget will be _____ because . . .

b. If the desired spending of each household increases by $5, the winning budget will be _____ because . . .

7. Cops: Chasing Criminals

Consider a metropolitan area with many municipalities. In Chaseville, the marginal-benefit curve for police has a vertical intercept of $140 and a slope of −$10 per police officer. The marginal cost of a police officer is constant at $60.

a. Use a graph to show how to determine how many police officers to hire.

b. Chaseville will hire _____ police officers because . . .

c. Suppose police officers chase criminals to other municipalities, increasing the cost of crime elsewhere. Each police officer hired by Chaseville increases the cost of crime elsewhere by $30. The socially efficient number of officers in Chaseville is _____ because . . .

d. You have been hired by the other municipalities to negotiate an agreement with Chaseville to hire the socially efficient number of police officers. As a group, these other municipalities are willing to pay up to _____ for such an agreement, computed as . . .

e. Chaseville is willing to accept as little as _____ to hire the socially efficent number of officers, computed as . . .

f. The willingness to pay exceeds the willingness to accept because . . .

8. Fleeburg

Consider Fleeburg, a city that is subject to outmigration. After graduating from high school, 20 percent of high-school graduates leave the area.

Depict graphically the equilibrium spending on education and the optimum spending.

9. **Alienated Voters**
Consider the example of voting in a representative democracy shown in Figure 16–4. Suppose a citizen abstains from voting if the best candidate has a budget position that is more than $2 from the citizen's preferred budget. In other words, a citizen votes only if the gap is less than or equal to $2.
   - **a.** If each candidate adopts the position of the median voter, the tally will be _____ for Buck and _____ for Penny.
   - **b.** If Buck increases his budget position to $8, the vote tally will be _____ for Buck and _____ for Penny.
   Suppose the distribution of voters changes. For budgets from $3 through $9, there are 10 voters for each preferred budget level. For the two extremes ($2 and $10), there are 12 voters at each budget. The median budget remains the same ($6).
   - **c.** If each candidate adopts the position of the median voter, the tally will be _____ for Buck and _____ for Penny.
   - **d.** If Buck increases his budget position to $8, the vote tally will be _____ for Buck and _____ for Penny.

10. **Torn by Indifference?**
Consider the example of voting in a representative democracy shown in Figure 16–4. Suppose half of voters abstain from voting if the difference between the two candidates is too small. Specifically, half of voters of each type will abstain if the difference between the two candidates is less than $2. If the difference is greater than or equal to $2, everyone votes.
   - **a.** If each candidate adopts the position of the median voter, the tally will be _____ for Buck and _____ for Penny.
   - **b.** If Buck increases his budget position to $8, the vote tally will be _____ for Buck and _____ for Penny.

REFERENCES AND ADDITIONAL READING

1. Downs, Anthony. "The Devolution Revolution: Why Congress Is Shifting a Lot of Power to the Wrong Levels." *Brookings Policy Brief,* no. 18. Washington, DC: Brookings Institution, 1998.
2. Fisher, Ronald C., and Robert W. Wassmer. "Economic Influences on the Structure of Local Government in U.S. Metropolitan Areas." *Journal of Urban Economics* 43 (1998), pp. 444–70.
3. Heikkila, Eric J. "Are Municipalities Tieboutian Clubs?" *Regional Science and Urban Economics* 26 (1996), pp. 203–26.
4. Inman, Robert P. "The Fiscal Performance of Local Governments: An Interpretive Review." In *Current Issues in Urban Economics,* eds. Peter Mieszkowski

and Mahlon Straszheim. Baltimore: Johns Hopkins University Press, 1979, pp. 270–321.

5. Levin, Henry. "Educational Vouchers: Effectiveness, Choice and Costs." *Journal of Policy Analysis and Management* 17, no. 3 (1997), pp. 373–92.
6. Musgrave, Richard A., and Peggy B. Musgrave. *Public Finance in Theory and Practice*. New York: McGraw-Hill, 1980.
7. Oates, Wallace E. *Fiscal Federalism*. New York: Harcourt Brace Jovanovich, 1972.
8. Panizza, Ugo. "On the Determinants of Fiscal Centralization: Theory and Evidence." *Journal of Public Economics* 74 (1999), pp. 97–140.
9. Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. "Teachers, Schools, and Academic Achievement." NBER Working Paper Number 6691 (1998).
10. Rouse, Cecilia. "Schools and Student Achievement: More Evidence from the Milwaukee Parental Choice Program." *Federal Reserve Bank of New York Economic Policy Review* (March, 1998), pp. 61–76.
11. U.S. Bureau of the Census. *Census of Government 2002*.

CHAPTER 17

Local Government Revenue

Sonja: What are you suggesting, passive resistance?
Boris: No, I'm suggesting active fleeing.
—From the Movie *Love And Death* (1975)

This chapter explores the economics of the two most important revenue sources of local government, the property tax and intergovernmental grants. We address two key questions about these revenue sources:

1. **Who bears the cost of the property tax?**

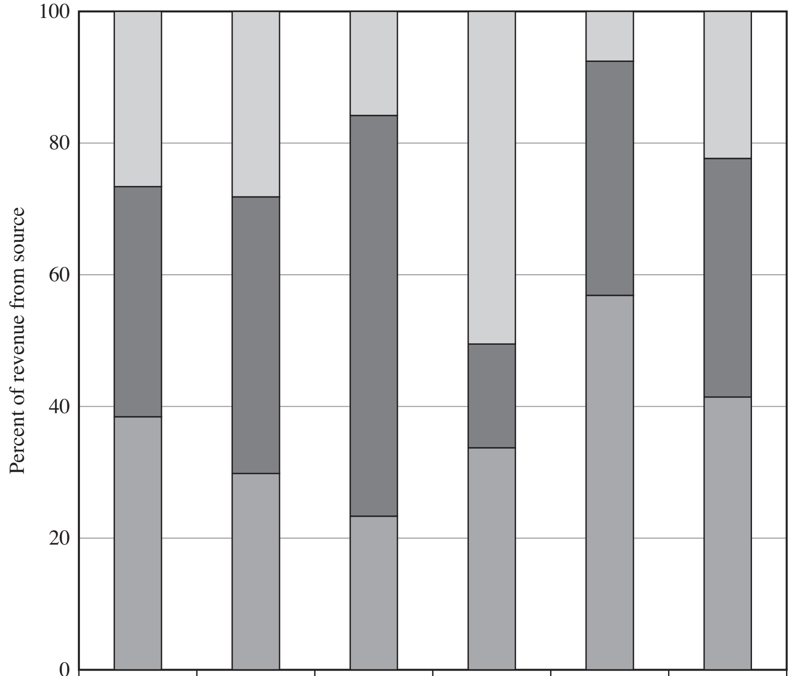
We'll use the model of supply and demand to show that the person who pays the tax in a legal sense shifts the tax to landowners, capital owners, and consumers.

2. **How does a local government respond to a grant for a particular program such as special education?**

We'll use the model of the median voter to show that a local government will use part of a grant to cut taxes and increase spending on other local public goods. Figure 17–1 (page 432) shows the distribution of revenue for different types of local governments. For local governments as a whole, 41 percent of revenue comes from intergovernmental grants, 36 percent comes from local taxes, and the remaining 22 percent comes from charges and general revenue. School districts are heavily dependent on intergovernmental grants, while special districts are heavily dependent on charges and general revenue.

Figure 17–2 (page 433) shows the distribution of local government revenue from different taxes. The property tax generates about three-fourths of local tax revenue and about half of municipal revenue. The sales tax generates roughly three times as much revenue as the individual income tax. Taxes on corporate income generate a small fraction of local tax revenue.

FIGURE 17–1 Revenue Shares of Local Government



	County	Municipal	Township	Special Dist.	School Dist.	Total
Charges and miscellaneous	27	28	16	51	8	22
Taxes	35	42	61	16	36	36
Intergovernmental grants	38	30	23	34	57	41

Source: U.S. Bureau of Census. *Census of Governments 2002*.

WHO PAYS THE RESIDENTIAL PROPERTY TAX?

No one likes to pay taxes, and once a tax is imposed, people change their behavior to try to avoid paying the tax. As a result, most taxes are at least partly shifted onto someone else. As we'll see, the property tax is shifted onto landowners, capital owners, and consumers.

Table 17–1 (page 433) shows the tax rates on residential property for selected cities, each of which is the largest city in its state. The effective tax rate is defined as the tax liability of a property divided by its market value. Because many local governments assess property at less than its full market value, the effective tax rate is typically less than the nominal tax rate (tax divided by the assessed value). As shown in the table, the effective tax rate ranges from 0.37 percent of market value per year in Honolulu to 3.86 percent in Bridgeport, Connecticut.

FIGURE 17–2 Revenues from Different Taxes

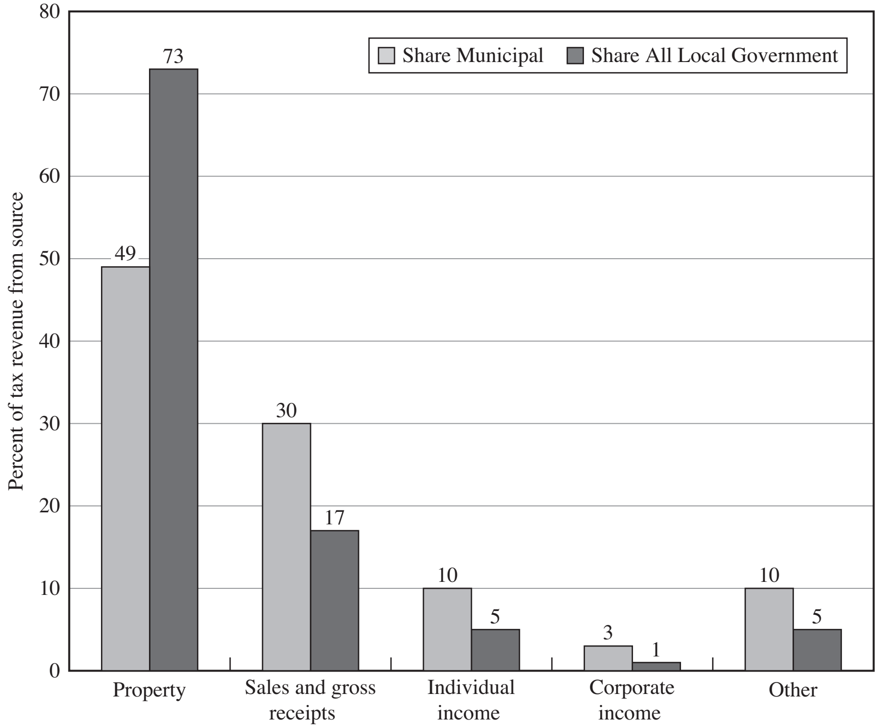


TABLE 17–1 Effective Property Tax Rates in Selected Cities, 2002

City	Effective Tax Rate	City	Effective Tax Rate
Bridgeport, CT	3.86%	Boston	1.10
Newark	2.95	Minneapolis	1.27
Milwaukee	2.67	Los Angeles	1.08
Des Moines	2.28	Phoenix	1.82
Houston	2.62	Chicago	1.69
Philadelphia	2.64	New York	0.93
Jacksonville	1.96	Denver	0.56
Memphis	1.76	Honolulu	0.37
Portland	1.46		

Source: U.S. Census Bureau. *Statistical Abstract of the United States*. Washington, DC (2004).

The property tax is an annual tax on residential, commercial, and industrial property. The total value of a particular property is the value of the structure plus the value of land. For example, suppose a property has a market value of $100,000, with $80,000 for the structure and $20,000 for the land. With a 1 percent property tax, the annual tax liability will be $1,000, equal to $800 for the structure plus $200 for the land.

Consider the residential city of Taxton, where all land is used for rental housing in the form of mobile homes. The rental housing industry is perfectly competitive, and, in equilibrium, each firm makes zero economic profit. Housing firms produce rental housing with two inputs, structure (capital) and land:

- **Structures.** A mobile home is a form of physical capital that housing firms rent from capital owners who live elsewhere. A mobile home can be moved costlessly from one city to another.
- **Land.** Housing firms rent the land under the mobile homes from absentee landowners. The lot size is fixed.

The housing firm rents housing (mobile home and land) to consumers. The initial (pretax) housing rent is \$5,000 per year, equal to \$4,000 for the structure rent and \$1,000 for land rent.

We assume that the property tax is paid in a legal sense by housing firms. To simplify matters, suppose the property tax is \$800 per mobile home and \$200 per standard lot. In other words, the property tax is a unit tax rather than a tax based on value. We are interested in the effect of the property tax on four types of people: owners of housing firms, housing consumers, landowners, and capital owners.

The Land Portion of the Property Tax

Consider first the land portion of the property tax. In Figure 17–3 (page 435), the supply of land is perfectly inelastic, with a fixed supply of 900 lots. The demand for land comes from housing firms, who use it as an input to rental housing. The demand curve intersects the supply curve at point *i*, generating an initial land rent of \$1,000 per lot.

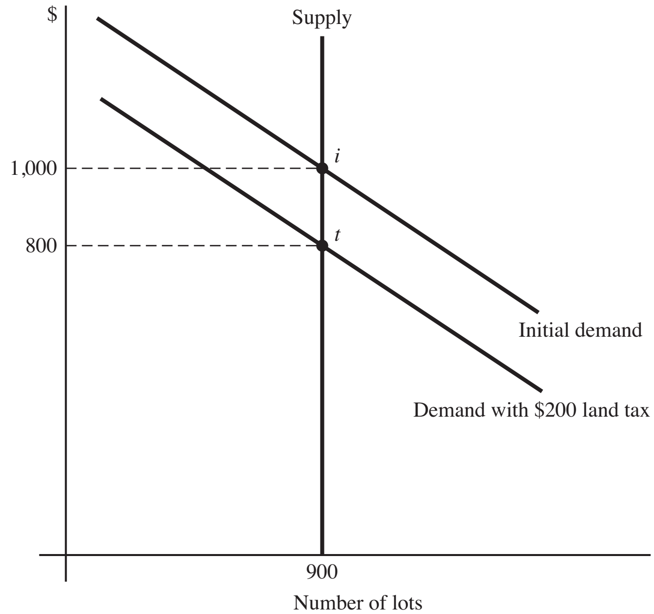
The demand curve shows how much a housing firm is willing to pay to landowners for one lot. If the firm pays a tax of \$200 per lot, it is willing to pay \$200 less to the landowner. In Figure 17–3, the \$200 land tax shifts the demand curve downward by \$200, and the new equilibrium is shown by point *t*, with a rent of \$800. The tax decreases land rent paid to landowners by the full amount of the tax. In other words, housing firms shift the entire land tax backwards onto landowners. This happens because the supply of land is perfectly inelastic. If landowners refused to cut land rent by \$200, the net price of land to housing firms (rent plus the tax) would rise above \$1,000. As a result, the quantity of land demanded would be less than the fixed supply of 900 lots. The resulting excess supply would cause rent to decrease until it reached \$800.

Structure Portion: A Partial-Equilibrium Approach

Consider next the structure portion of the property tax. We will start with partial-equilibrium analysis, looking at the effect of the tax in one input market (structures) in one city (Taxton). The analysis is partial because it ignores the effects of the tax on other markets and other cities.

Figure 17–4 (page 436) shows the initial equilibrium in the structure market. The initial supply curve for mobile homes is horizontal at \$4,000 per structure.

FIGURE 17–3 Market Effects of the Land Portion of the Property Tax



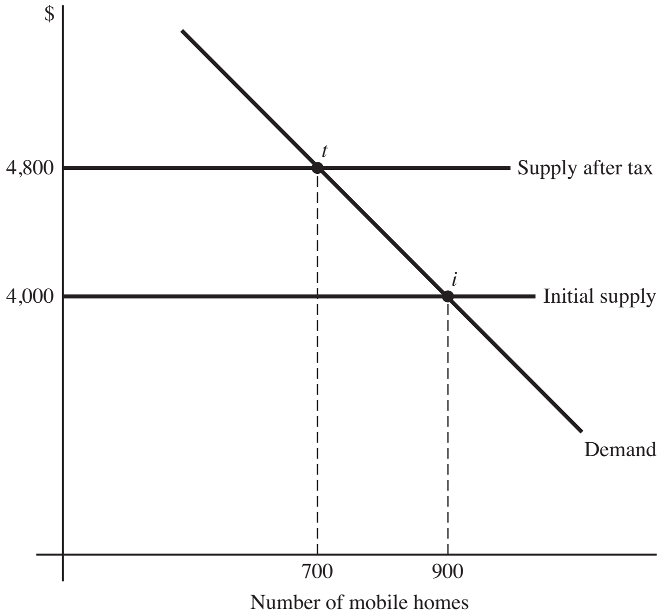
A land tax of $200 per lot decreases the amount a housing firm is willing to pay to a landowner by $200. Because the supply is perfectly inelastic, the equilibrium rent drops by $200, meaning that the landowner pays the entire tax.

As explained in Section 2.2 in the appendix "Tools of Microeconomics," a supply curve is also a marginal-cost curve. Recall that a housing firm rents mobile homes from capital owners, and the annual payment—the return to capital—is the firm's only cost of providing a structure to consumers. Therefore, the structure supply curve shows the housing firm's cost per mobile home. The horizontal supply curve indicates that the return to capital (the firm's cost per mobile home) is fixed at $4,000. To get a mobile home, a housing firm must pay the capital owner $4,000, regardless of how many mobile homes are used in Taxton. The initial supply curve intersects the demand curve at point *i*, with a quantity of 900 structures and a structure rent of $4,000.

In Figure 17–4, a structure tax of $800 shifts the supply curve upward by the amount of the tax. The tax increases the housing firm's marginal cost by the amount of the tax. For each structure, the housing firm pays $4,000 to the capital owner and $800 to the government, so the firm's marginal cost of a structure is now $4,800. The new equilibrium is shown by point *t*, with a price of $4,800 and a quantity of 700 dwellings. In other words, the entire structure tax is passed forward onto consumers, who pay $800 more for housing.

So housing firms don't pay *any* of the property tax? They shift the land portion backward onto landowners and shift the structure portion forward to consumers.

FIGURE 17–4 The Partial-Equilibrium Effects of the Structure Tax



The supply curve for structures shows the firm's cost per mobile home, the cost of renting a mobile home from a capital owner. The structure tax shifts the supply curve upward by the amount of the tax. If the return to capital is fixed, the supply curve for structures is horizontal, and the tax increases the structure rent by the full amount of the tax.

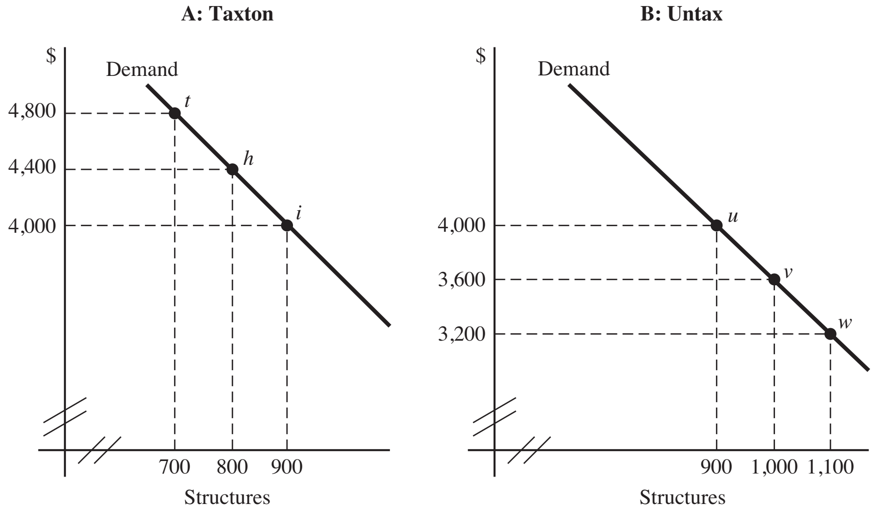
They get the money to pay the $1,000 tax by paying $200 less for land and charging $800 more to consumers. This of course is not unique to the housing market, but is the normal consequence of a tax on a competitive industry. A tax is shifted backward onto input suppliers and forward onto consumers, leaving producers with zero economic profit, just as they had before the tax.

Structure Portion: A General-Equilibrium Approach

So far we have looked at the effects of the structure tax in the taxing city. The partial-equilibrium analysis ignores the effects of the tax on people outside the city. Figure 17–4 shows the need for a more general approach. The tax decreases the quantity of structures in Taxton from 900 to 700. Where do the mobile homes go? What are the economic consequences? A general-equilibrium approach answers these questions.

We can extend our example by introducing a second city in the region, one without a property tax. Before Taxton imposes its property tax, the two cities are identical: Each city has 900 mobile homes and the equilibrium rent is $4,000.

FIGURE 17–5 General-Equilibrium Effects of the Structure Tax



The partial-equilibrium outcome is shown by point *t* in Taxton. Consumers pay a structure rent of $4,800, including $4,000 return to structure capital and a $800 structure tax. The 200 structures that flee Taxton increase the number of structures in Untax (point *w*), decreasing the return to capital to $3,200. General equilibrium requires equal capital returns in the two cities, shown by points *h* in Taxton (structure rent = $4,400; return to structure capital = $3,600) and *v* in Untax (structure rent = return to structure capital = $3,600). The structure tax decreases the common return to capital from $4,000 to $3,600.

We assume that mobile homes can be moved costlessly between the two cities but cannot leave the region. In other words, the regional supply of structures (capital) is fixed.

Our general-equilibrium analysis must account for the 200 mobile homes that flee the tax in Taxton. In Figure 17–5, the partial-equilibrium outcome is shown as point *t* in Panel A. There are 700 structures in Taxton (down from 900 at point *i*), and consumers pay $4,800 per structure, enough to cover a return to capital of $4,000 and an $800 tax. In Panel B, if 200 structures flee to the untaxed city (Untax), we move downward along the demand curve from point *u* to point *w*, and the return to capital decreases from $4,000 to $3,200. The return to capital decreases because to get consumers in Untax to rent the additional mobile homes, the housing firm must cut the structure rent. Otherwise, some mobile homes will be vacant and generate zero rent.

The flight of the mobile homes from Taxton to Untax generates a gap between the return to capital in the two cities. In Taxton, the return is $4,000: Housing firms collect $4,800 from consumers, pay the $800 tax, and pay $4,000 to capital owners. In Untax, the increased supply of mobile homes decreases the return to $3,200. This is not an equilibrium because capital owners have an incentive to move their mobile homes to Taxton. Equilibrium requires the same return to capital in the two cities.

The second flight of the mobile homes—from Untax to Taxton—is shown by the movement upward along the Untax demand curve and downward along the Taxton demand curve. If 100 mobile homes make the trip, we go from point *w* to point *v* in Untax, and the return increases to $3,600. In Taxton, we go from point *t* to point *h*, and the return decreases to $3,600 (the $4,400 rent charged to consumers minus the $800 tax). With points *v* and *h*, the return to capital is equalized, so we have an equilibrium, with 800 structures in Taxton and 1,000 structures in Untax.

Our conclusion is that the structure tax is paid by capital owners throughout the region. A tax of $800 per structure in one city decreases the return on capital by $400 per structure throughout the region. The tax is fully shifted backward onto capital owners because the regional supply of capital is assumed to be fixed. Recall that the land tax is fully shifted backward to landowners because the supply of land is fixed. The same basic logic applies to the structure tax: If an input is fixed in supply, owners of the input will bear the tax.

What about consumers? Let's assume that consumers can move costlessly between the two cities. Housing firms make zero economic profit, so the housing rent is just high enough to pay the firm's cost:

Housing rent = Return to capital + structure tax + land rent

Assume for the moment that land rent is fixed at $1,000, as shown in the second row of Table 17–2. As a result, housing rent is $5,400 in Taxton ($3,600 + $800 + $1,000) compared to only $4,600 in Untax ($3,600 + $1,000). As a result, consumers have an incentive to move from Taxton to Untax.

Locational equilibrium for consumers requires the same housing rent in the two cities. Recall the first axiom of urban economics:

Prices adjust to generate locational equilibrium

In this case, the price of land will adjust to equalize housing rent and make consumers indifferent between the two cities. In the third row of Table 17–2, the gap in housing rent can be closed if land rent decreases to $600 in Taxton and increases to $1,400 in Untax. With these changes in land rent, housing rent is $5,000 in each city, so consumers will be indifferent between the two cities. This means that the structure tax causes landowners in Taxton to lose $400 per lot, while landowners in Untax gain $400 per lot.

TABLE 17–2 General-Equilibrium Effects of the Structure Tax with Two Cities

	Taxton				Untax		
	Return to Capital	Structure Tax	Land Rent	Housing Rent	Return to Capital	Land Rent	Housing Rent
Initial	$4,000	$ 0	$1,000	**$5,000**	$4,000	$1,000	**$5,000**
Before change in land rent	$3,600	$800	$1,000	**$5,400**	$3,600	$1,000	**$4,600**
After change in land rent	$3,600	$800	$ 600	**$5,000**	$3,600	$1,400	**$5,000**

So who bears the cost of the structure portion of the property tax? Recall that for the moment we are assuming the supply of capital (structures) in the region is fixed.

- Capital owners bear the tax. The return to capital falls by $400 per structure in both cities.
- Landowners in the region experience zero-sum changes in rent, with landowners in the untaxed city gaining at the expense of landowners in the taxed city.
- Consumers pay the same price for housing, so they do not bear any part of the tax.
- Housing firms make zero economic profit. In the taxed city, they get the money to pay the $800 tax by paying $400 less to capital owners and $400 less to landowners. In the untaxed city, they pay $400 less to capital owners but pay $400 more to landowners.

Changing the Assumptions

The simple general-equilibrium model uses a number of assumptions to make the basic results transparent and clear-cut. If we modify some of these assumptions, things are not so tidy.

One of the key assumptions is that the total supply of capital (structures) is fixed. In fact, we expect that a lower return on capital will reduce the quantity of capital supplied. For example, some of the structures that flee the tax could be withdrawn from the market rather than simply moving to the untaxed city. If so, the initial excess supply of structures in the untaxed city won't be as large, so the return on capital won't drop as far. In equilibrium, housing rent will be greater than $5,000, meaning that part of the structure tax will be shifted to consumers, leaving a smaller burden on capital owners.

A second key assumption is that consumers are perfectly mobile between the two cities. As a result, any intercity differences in housing rent are eliminated by changes in land rent. If instead residents are immobile, the gap in housing rent will persist. Looking back at Table 17–2, with perfectly immobile consumers, we will be stuck in the second row, with consumers in Taxton paying $800 more for housing. To summarize, when consumers are perfectly mobile, there will be zero-sum changes in land rent; when consumers are perfectly immobile, there will be zero-sum changes in housing rent. Between these two extremes, when consumers are mobile but not perfectly mobile, both housing rent and land rent will change.

A third key assumption is that there are only two cities in the region. If there were 10 cities instead, the effects of Taxton's structure tax would be spread over five times as much capital. As a result, the decrease in the return to capital would be one-fifth as large: The return to capital would drop by $80 instead of $400. Table 17–3 (page 440) shows the implications for shifting the structure tax. To equalize housing rent between the cities, the price of land would increase by $80 in the untaxed cities and decrease by $720 in the taxing city. Notice that the changes in land rent in the region sum to zero: Nine cities experience an $80 rise and one experiences a $720 decline.

TABLE 17–3 The Structure Tax with 10 Cities

	Taxton				Untaxed Cities		
	Return to Capital	Structure Tax	Land Rent	Housing Rent	Return to Capital	Land Rent	Housing Rent
Initial	$4,000	$ 0	$1,000	$5,000	$4,000	$1,000	$5,000
Before change in land rent	$3,920	$800	$1,000	$5,720	$3,920	$1,000	$4,920
After change in land rent	$3,920	$800	$ 280	$5,000	$3,920	$1,080	$5,000

FROM MODELS TO REALITY

To explain the effects of the property tax on different sorts of people, we have used a number of modeling artifices that may seem to limit the applicability of the results. But in fact we can easily apply the lessons from the artificial model to real markets. Consider the lessons for property owners and policy makers.

What about Rental Property Owners and Homeowners?

Our model of the housing market has four economic actors: consumers, owners of housing firms, landowners, and capital owners. In a real rental housing market, these roles are merged into two: Housing firms own property (land and structures), and consumers rent housing from the firm. In the homeowner market, the roles are merged into one, with consumers as property owners. What does the general-equilibrium model say about the burden of the property tax for rental property owners and homeowners?

Property owners in a taxing city lose as owners of land and capital. They lose as landowners because (1) the land portion of the tax decreases land rent and (2) part of the structure portion is shifted onto land. In addition, like other capital owners in the region, they lose because the return to capital decreases. In general, the property tax decreases the market value of property. This is sensible because the property now carries a tax liability, so potential buyers are willing to pay less for it.

Although property owners in other cities don't pay the tax in a legal sense, they are affected by it. They gain as landowners because land rent in their city rises to equalize housing rents. Like other capital owners, they lose as the regionwide return on capital decreases. So the net effect on their income and the market value of their properties is ambiguous.

A Practical Guide for Policy Makers

We've explored the effects of the residential property tax with different models and assumptions. Suppose an elected official asks, Who actually pays the property tax? The appropriate response depends on the official's perspective. We'll consider a city perspective and a national perspective.

TABLE 17–4 Who Pays the Structure Portion of the Property Tax?

Tax Imposed by a Single City

Effects in the taxing city

1 Mobile households: Landowners receive lower land rent.
2 Immobile households: Consumers pay higher housing rent.

Effects in an untaxed city

1 Mobile households: Landowners receive higher land rent.
2 Immobile households: Consumers pay lower housing rent.

Regional effects

1 Capital owners receive lower return on capital.
2 Mobile households and fixed capital: zero-sum changes in land rent.
3 Immobile households and fixed capital: zero-sum changes in housing rent.
4 Mobile households and variable supply of capital: Consumers pay higher housing rent; the reduction in the return on capital is smaller.

Tax Imposed by All Cities (a National Property Tax)

1 Fixed supply of capital: Entire tax borne by capital owners.
2 Variable supply of capital: Part of tax shifted to housing consumers.

Consider a mayor who wants to predict the effect of her city's structure tax on citizens in her city. Her city is one of 50 cities in a regional economy. As we saw in Table 17–3, in a 10-city region, the regionwide return to capital decreases by 1/10 of the tax ($80) and land rent in the taxing city decreases by 9/10 of the tax ($720/$800). For a region with 50 cities, the return to capital will decrease by 1/50 of the tax, leaving 49/50 for land. So the mayor can assume that most of the tax will be borne by local landowners. Of course, if households are not perfectly mobile, housing consumers will share the burden of the structure tax with local landowners, as explained earlier.

Consider next a president who wants to predict the effect of a uniform property tax across cities in the nation. With the same tax rate in all cities, structures have nowhere to flee from one city's property tax. If the national supply of capital is fixed, the entire tax will be borne by the owners of capital. In this case, capital owners cannot shift the tax to anyone else because they don't respond to the tax: They don't move their capital between cities, and they don't decrease the total amount of capital in the nation. Of course, if the supply of capital is variable rather than fixed, capital owners can shift the tax to households in the form of higher housing rent throughout the nation.

Table 17–4 summarizes our analysis of the structure portion of the property tax. It shows who bears the burden of the tax under different assumptions about the type of tax, household mobility, and the supply of capital. The table also distinguishes between the taxing city and other cities.

What about the Business Property Tax?

The basic logic we have used to examine the residential property tax applies to the property tax on business property (commercial and industrial). Of course, instead of housing services as the output of the taxed industry, we have products such as books, haircuts, clothing, and manufactured goods. When a single city imposes a business property tax, the general-equilibrium approach shows that the structure portion of the tax will be borne by the owners of capital throughout the region as capital flees the taxing city. As in the case of the residential property tax, the effect on consumers depends on their mobility—their ability to switch to sellers in untaxed municipalities.

Tax exporting is the process of getting people outside the municipality to pay taxes. A city can use the business property tax to shift taxes to outsiders if they consume some of the city's products. There are limits, of course. As the price of a city's export goods increases, the quantity demanded decreases, decreasing the tax base. In addition, firms have an incentive to move to cities with lower property taxes. Tax exporting is more lucrative when a city has a unique production advantage that makes it a superior location for export firms.

THE TIEBOUT MODEL AND THE PROPERTY TAX

As we saw earlier in the book in the chapter on neighborhood choice, citizens sort themselves with respect to their demands for local public goods. If local public goods are financed with a property tax, households will also sort themselves with respect to housing consumption. This has important implications for the property tax, as Table 17–5 demonstrates.

Consider a metropolitan area where households have the same preferred level of local public goods ($6,000), but live in houses with different market values. The first row of Table 17–5 shows what happens in a mixed municipality with a tax rate of 0.02 (2 percent of value). Juan, who lives in a $100,000 house, pays a property tax of $2,000. At the other extreme, Thurl lives in a $500,000 house and pays five times as much. Thurl is paying more than her share of taxes and has an incentive to form a municipality with other people in expensive houses.

The last three rows in the table show what happens when citizens sort themselves into municipalities according to house value. A municipality full of expensive houses needs a tax rate of only 0.012 to generate the $6,000 necessary to support $6,000 worth of public services per household. With Juan-type households in one municipality (second row) and Tupak types in another municipality (third row),

TABLE 17–5 Municipality Formation for Tax Purposes

		Tax Bill For		
Outcome	Tax Rate	Juan ($100k house)	Tupak ($300k house)	Thurl ($500k house)
Mixed municipality	0.02	$2,000	$6,000	$10,000
All $100k houses	0.06	$6,000	—	—
All $300k houses	0.02	—	$6,000	—
All $500k houses	0.012	—	—	$6,000

everyone has the same tax liability, even though they have different house values. People in less expensive houses have higher tax rates, allowing them to pay $6,000 in taxes.

Because households sort themselves into homogeneous communities, the property tax is a user fee, not a conventional tax. A household's property tax liability is determined by its consumption of the local public good, not by its property value. In the Tiebout world, households get what they pay for, and the question of who pays the property tax is simple: Just as a consumer pays $2 to get a hot dog, a household pays a property tax of $6,000 to get $6,000 worth of local public goods. There is no tax shifting because the tax is a user fee.

How realistic is the Tiebout model and the user-fee view of the property tax? Given the large number of municipalities in the typical metropolitan area, households can choose from a wide variety of municipalities and local governments. But the sorting process is by no means perfect, even in suburban areas. The Tiebout model is clearly inapplicable to central cities, where a single municipality serves a large and diverse population. In large central cities, the property tax is not a user fee, but a conventional tax.

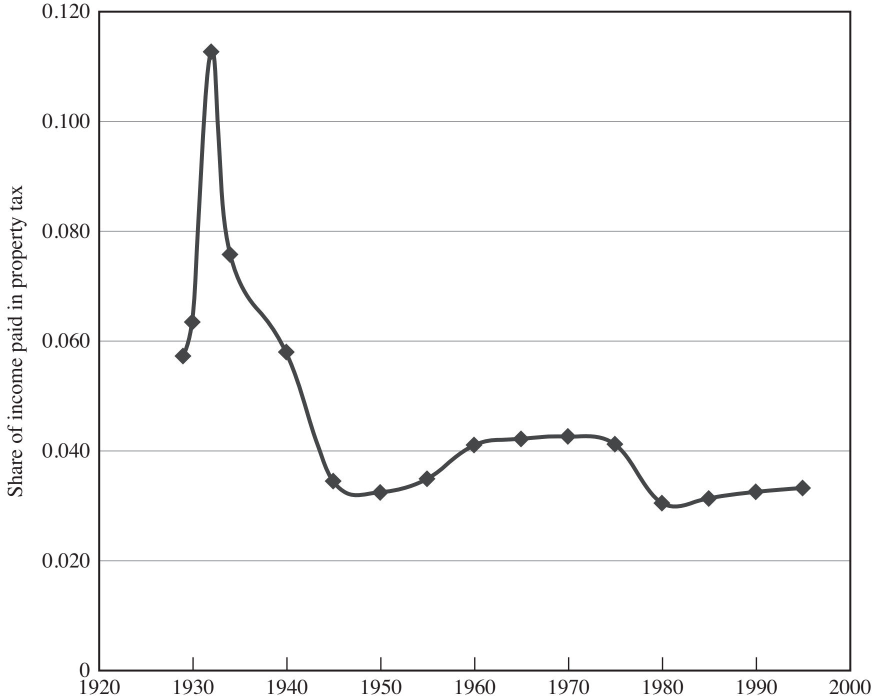
LIMITS ON PROPERTY TAXES

Limits on property taxes started in the 1870s and are currently in force in 44 states. About two-thirds of states limit the tax rate for specific types of local government, and about a quarter limit the tax rate for local government as a whole. Most of the rate limits fall in the range of 10 to 20 mills (1 to 2 percent of assessed value). About half of the states limit the annual growth rate of property tax revenue, with most limits in the range of 4 to 6 percent. Some states peg the growth rate to the inflation rate. In many states, local governments have the option to override a state limit with voter approval.

The first property tax revolt came during the Great Depression, a result of a mismatch between property tax liabilities and citizens' willingness to pay for local public services. As shown in Figure 17–6 (page 444), the share of income absorbed by the property tax doubled between 1929 and 1932, reaching 11.3 percent in 1932. During this three-year period, personal income was cut in half while property tax revenue decreased by only 9 percent. The decrease in citizens' ability to pay property taxes nearly tripled the delinquency rate. Fearing massive defaults on municipal bonds, the business community supported protax campaigns by paying for lapel buttons, mass mailings, and parades. The parades featured the descendants of canine war heroes, who barked and carried signboards urging people to pay their taxes.

In 1933, over 3,000 local tax leagues were agitating for tax reform. The clear message was that local government should scale back its operations to reflect lower income during the Great Depression. In the words of one agitator, "I buy less food, less tobacco, less recreation, and I'd like to buy less government." (Beito, 1989, page 18). In mass meetings organized by the tax leagues, citizens demanded the elimination of local services, including weed inspectors and county nurses.

FIGURE 17–6 Property Tax Revenue as a Share of Income



Source: Arthur O'Sullivan. "Limits on Local Property Taxation." Chapter 7 in *Property Taxation and Local Government Finance*, ed. Wallace E. Oates. Cambridge MA: Lincoln Institute, 2001.

The tax revolt of the 1930s resulted in the passage of tax limits that reduced the tax burden. In 1932 and 1933, a total of 16 states passed tax limits, with most of the measures setting a maximum overall rate for local property taxes. As shown in Figure 17–6, the share of income absorbed by property taxes dropped between 1932 and 1940. The decrease in the tax share resulted from a combination of income growth and the tax limits. By 1940, personal income had almost reached the level observed in 1930, while the share of income absorbed by the property tax was 5.8 percent, compared to 6.3 percent in 1930.

The modern tax revolt started in 1978 with the passage of Proposition 13 in California. As shown in Figure 17–6, during the period 1960 to 1975, the share of national income absorbed by the property tax was high by recent historical standards, about 4.2 percent, compared to 3.4 percent during the late 1940s and 1950s. By 1995, dozens of states had enacted new tax limits, and the share of income absorbed by property taxes dropped to 3.3 percent, the level observed in the 1940s and 1950s.

In contrast with the earlier tax revolt, the supporters of modern tax limits expected local governments to provide the same level of service with less money. In California, 38 percent of the citizens believed that state and local governments could absorb a 40 percent cut in tax revenue without cutting services. In Massachusetts, 82 percent of the supporters of Proposition 2 1/2 believed that the proposition would cut taxes without reducing the quality of local public services. In Michigan, three-fourths of the supporters of the Headlee Amendment expected the government to absorb the revenue cut by simply becoming more efficient.

In the 1990s, two states changed their property-tax systems. Illinois established limits on the growth rate of property tax revenues in the Chicago metropolitan area. The maximum growth rate is the maximum of the inflation rate or 5 percent per year. In 1995, Michigan reformed its entire education finance system. The state cut property tax revenue in half and offset the revenue loss by increasing sales taxes, tobacco taxes, and real estate transfer taxes.

There is evidence that modern tax limits reduce property taxes. As shown in Figure 17–6, the share of income paid in property taxes has fallen since the onset of the modern revolt in 1978. Property tax limits reduce real per-capita tax revenue by 3 percent to 6 percent (Advisory Committee on Intergovernmental Relations, 1995; Shadbegian, 1998). There is also some evidence of revenue substitution, with revenue from other sources at least partly offsetting the loss of property taxes. One response is to increase intergovernmental grants from state government. A second response is to increase nontax revenue from fees and charges. Shadbegian (1999) estimates that for each $1 reduction in county tax revenue, there is a $0.27 increase in miscellaneous revenue.

INTERGOVERNMENTAL GRANTS

This part of the chapter explores the economics of intergovernmental grants, examining how local governments respond to transfers of funds from higher levels of government. As we saw earlier in the chapter, intergovernmental grants provide about two-fifths of the revenue of local government and about one-fourth of the revenue of municipalities. Over half of this grant money goes to education, and the rest supports other local programs such as public welfare, housing and community development, highways, and health and hospitals. At the municipal level, about one-fifth of grant money supports the general operations of local government, and another fifth supports education. The two redistributional programs—public welfare and housing and community development—together get about a quarter of the grant money received by municipalities.

Why don't local governments pay their own way, supporting their spending programs with local taxes? First, intergovernmental grants can be used to internalize interjurisdictional spillovers, as discussed in the previous chapter. Second, if the desired spending on local public goods rises faster than the local tax base (e.g. property values and retail sales), there will be a mismatch between desired spending and local revenue. At the national level, tax revenue increases more rapidly with income, providing an opportunity to transfer surplus funds to local governments. Of course, a more straightforward response to the mismatch problem would be to increase local tax rates.

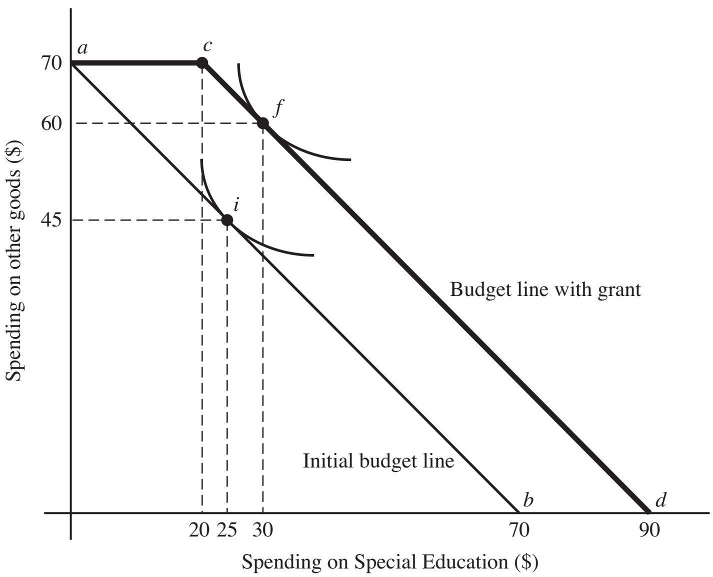
We will explore the responses of local government to two types of grants. A lump-sum grant is a fixed grant, independent of a local government's spending on a local public good. In contrast, under a matching grant, a higher level of government matches local spending, for example $1 of grant money for every $1 spent locally.

Lump-Sum Grants

Most lump-sum intergovernmental grants come with strings attached. The money from a conditional or categorical grant must be spent on a specific program. Conditional grants are provided for education, public welfare, health and hospitals, highways, housing, and community development. Within each expenditure group are program-specific grants. For example, education grants to local governments include specific grants for remedial reading, school libraries, special education, and other programs. We will use a grant for special education as an example.

We can use the consumer choice model to explore the effects of grants. The choice model is reviewed in Section 4 of "Tools of Microeconomics," the appendix at the end of the book. Figure 17–7 shows the budget line for Marian, the median voter in Grantburg. For every dollar spent on special education, there is one less dollar to spend on other goods, including local public goods and private goods. The indifference curves show Marian's trade-off between special education and other

FIGURE 17–7 Local Government Response to a Lump-Sum Grant



A lump-sum grant of $20 per capita shifts the budget line of the median voter from *ab* to *acd*, and the utility maximization point moves from point *i* to point *f*. The grant increases the spending on the target program (special education) by $5 and increases spending on other goods by $15.

goods. Given the initial budget line (*ab*) and her indifference curves, Marian's utility is maximized at point *i*. We know from earlier chapters that under majority rule, the city will choose the preferred budget of the median voter. In this case, the city spends $25 per household on special education, leaving Marian $45 for other goods.

Suppose the state gives the city a lump-sum grant of $20 per capita for special-education programs. The grant shifts the budget line from *ab* to *acd*. Point *c* is in the new budget set because Marian could spend all of her own money on other goods and use the $20 grant to support special education. For spending on special education above $20, there is a dollar-for-dollar trade-off between special education and other goods. The new utility-maximizing point is point *f*, meaning that the grant increases Marian's desired spending on special education to $30 (up by $5) and her desired spending on other goods to $60 (up by $15). In other words, one-fourth of the grant is spent on special education, and the rest is spent on other goods.

Why does a conditional grant of $20 increase spending on the target program by less than $20? The city can spend part of the grant on other goods because it decreases its own contribution to special education. Before the grant, $25 of local tax money was spent on special education. After the grant, total spending on special education is $30, and the city can combine the $20 grant with just $10 of local tax money. The grant frees up $15 worth of local tax money, which can be spent on other local public goods and private goods.

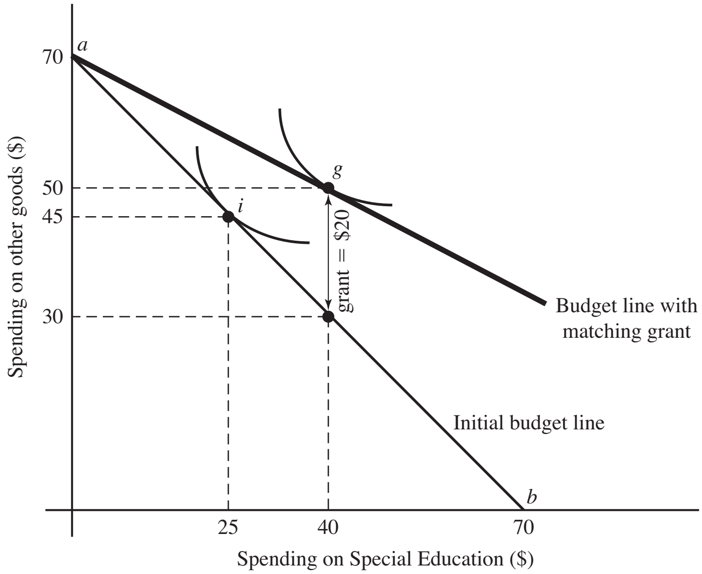
Matching Grants

Under a matching grant, a higher level of government contributes some amount for every dollar of local spending on a particular local public good. For example, under a one-for-one matching grant, the higher level of government gives one dollar in grant money for every dollar spent by local government. A matching grant decreases the opportunity cost of local public goods: With a one-for-one match, local citizens sacrifice only $0.50 in private goods to get a dollar's worth of local public goods ($0.50 of local spending plus a grant of $0.50).

Figure 17–8 (page 448) shows the effect of a one-for-one matching grant for special education. The grant decreases the slope of Marian's budget line, from $1 worth of other goods per dollar on special education to $0.50. Given the new budget line, Marian's utility-maximizing point moves from *i* to point *g*, and spending on special education increases from $25 to $40. Under the one-for-one grant, $20 of the city's $40 special-education budget comes from the state government.

The matching grant provides a greater stimulus to special education than an equivalent lump-sum grant. Although the state transfers the same amount for each type of grant ($20), the matching grant increases spending on special education to $40, while the lump-sum grant increases spending to only $30 (Figure 17–7). Both grants increase Marian's real income by $20, increasing her demand for special education and other goods. The matching grant also has a substitution effect because it cuts the opportunity cost (price) of special education in half. The decrease in the relative price of special education causes consumer substitution of special education for other goods.

FIGURE 17–8 Local Government Response to a Matching Grant



A one-for-one matching grant tilts the budget line of the median voter outward, and the utility maximization point moves from point *i* to point *g*. The grant increases the spending on the target program (special education) by $15 and increases spending on other goods by $5. The matching grant provides a bigger stimulus than a lump-sum grant because it decreases the opportunity cost of spending on the target program.

What about spending on other goods? Under a one-for-one matching grant, the local contribution to a $40 special-education budget is $20. This leaves $50 to spend on other goods, including other public goods and private goods, up from $45 before the grant. In other words, the city spends one-fourth of the $20 matching grant on other goods. Like a lump-sum grant, a matching grant increases spending on other goods as the local government cuts its own contribution to the program covered by the grant.

Up to this point, we have assumed that there is no upper limit on the matching grant. In many cases, the government specifies a maximum grant amount, and this type of grant is called a closed-ended matching grant. If the desired spending after the grant is less than the limit, the limit is irrelevant and the closed grant is equivalent to the open grant. If however, the desired spending exceeds the limit, the constraint is binding, and a closed grant generates a lower level of spending than an open grant.

Summary: The Stimulative Effects of Grants

As explained earlier in the chapter, local governments use intergovernmental grants to increase spending on local public goods and other goods, including private goods. Spending on private goods can increase because the local government can cut taxes.

What fraction of grant money is used for additional local spending, and how much is left over to increase the consumption of private goods? Empirical studies of the local response to grants conclude that each dollar from a nonmatching grant increases local government spending by roughly \$0.40 (Oates, 1999). In contrast, an additional dollar of household income increases local spending by about \$0.10. In other words, a local grant provides a bigger stimulus for local spending. This is known as the flypaper effect: The grant money sticks where it first hits (the local government) rather than being passed on to households in the form of lower taxes.

What explains the flypaper effect? The most prominent theory assumes that government bureaucrats want to maximize their budgets (Filimon, Romer, Rosenthal, 1982). If the bureaucrats hide grant money from citizens, voters are more likely to approve larger budgets. In states that have direct votes on local budgets, ballots often have information about the tax base, but rarely have information about intergovernmental grants.

Welfare Reform: Matching Grants to Lump-Sum Grants

A key component of the welfare-reform plan adopted in 1996 is the replacement of federal matching grants with lump-sum grants (also known as block grants). Under the old system, each state picked a level of welfare spending, and the federal government used matching grants to support local efforts. For low-income states, the federal rebate per dollar spent on welfare was \$0.78, so from a state's perspective, each dollar spent on welfare cost the state only \$0.22. The rebate was lower for high-income states, with a one-for-one match for the highest income states. Under the new grant system, the federal grant no longer depends on how much the state spends on welfare. There are no matching funds, so the state's price of a dollar spent on welfare is \$1.00.

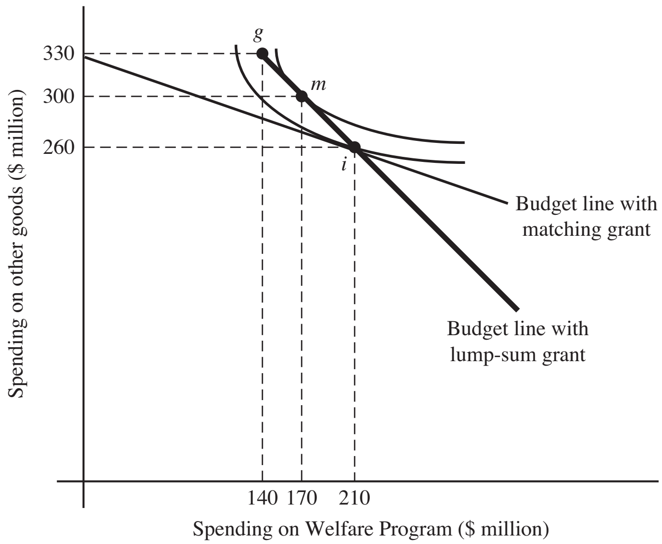
Figure 17–9 (page 450) uses the consumer choice model to show the effects of welfare reform on the budget choices of a low-income state. The budget line for the median voter under the matching grant is relatively flat, reflecting the low local price of welfare spending. The voter's initial preference (and thus the state's initial choice) is shown as point *i*, with \$210 million on welfare and \$260 million on other goods. The new lump-sum grant is \$140 million, so the new budget line is shown by the line connecting points *g*, *m*, and *i*. The lump-sum grant is large enough that the median voter has the option of picking its initial point *i*. If the initial point is possible, will the state choose it?

Under the lump-sum grant, the state will actually spend less on welfare programs. To maximize the utility of the median voter, the state picks the point where the slope of the indifference curve (the marginal rate of substitution) equals the slope of the budget line (the price ratio):

Utility-maximizing rule: Marginal rate of substitution = Price ratio

For the initial choice (point *i*), the marginal rate of substitution equals the price ratio, \$0.22. The switch to the lump-sum grant increases the price of welfare spending to \$1, so to maximize utility, the median voter moves to point *m*, where the marginal rate of substitution is 1.0. In other words, the median voter chooses to spend less on welfare

FIGURE 17–9 A Switch from a Matching Grant to a Lump-Sum Grant Decreases Spending



The replacement of a matching grant with a $140 lump-sum grant moves the utility maximizing point from point *i* to point *m*. The policy change reduces spending on the target program (welfare) because the opportunity cost of spending on welfare increases from $0.22 to $1.00.

programs and more on other goods. Comparing point *m* to point *i*, the state will spend $40 million less on welfare programs and $40 million more on other goods (other public goods and private goods). The switch to a lump-sum grant increases the price of welfare spending, causing a substitution effect that decreases welfare spending.

The predicted changes in welfare spending are large. For a low-income state, the price hike from $0.22 to $1.00 is projected to decrease welfare spending by 40 to 66 percent (Inman and Rubinfeld, 1997). For a high-income state, the price hike is smaller (from $0.50 to $1.00), and the switch to lump-sum grants is projected to decrease welfare spending by 1 to 18 percent. Congress was apparently aware that welfare reform would cause states to cut their welfare spending. The law requires states to continue to spend at least 80 percent of the amount spent under the old matching-grant policy.

SUMMARY

The two largest sources of local government revenue are the property tax and intergovernmental grants.

1. The supply of land is fixed, so the land portion of the property tax is borne by landowners.

2. Under the general-equilibrium analysis of the structure portion of the property tax, if the supply of capital (structures) is fixed at the regional level, the tax is borne by capital owners throughout the region.
3. The structure tax generates zero-sum changes in land rent or housing rent across cities. If consumers are perfectly mobile between cities, they are unaffected by the structure tax, but landowners in the untaxed city gain at the expense of landowners in the taxing city.
4. In the Tiebout world of household sorting, a household's property tax bill is independent of its housing consumption, so the property tax is a user fee.
5. The model of the median voter predicts that part of a categorical grant is spent on other local public goods and private goods.
6. A matching grant decreases the opportunity cost of spending on the targeted good, so it provides a greater stimulus than a lump-sum grant.

APPLYING THE CONCEPTS

For exercises that have blanks (_____), fill each blank with a single word or number. For exercises with ellipses (. . .), complete the statement with as many words as necessary. For exercises with words in square brackets ([increase, decrease]), circle one of the words.

1. **A Tax on Mobile Home Pads**
   The residents of mobile home parks own their dwellings and rent pads (the land under the mobile home) from landowners. In Padville, all land is initially occupied by mobile homes, and each resident rents one padacre (a standard pad). Each landowner owns one padacre. Initially, there are 100 residents, and the price of land is $200 per padacre. Suppose the city imposes a tax of $40 per padacre, regardless of how the land is used. The tax is paid, in legal terms, by the land user (the resident).
   a. Use a supply/demand graph to illustrate the effects of the tax on the land market.
   b. The land tax will [increase, decrease, not change] the amount paid by a resident to a landowner and will [increase, decrease, not change] the net cost of land to a resident, defined as the amount paid to the landowner plus the $40 tax.
   c. The land tax will [increase, decrease, not change] the income of landowners.
   d. The tax is paid in economic terms, by [residents, landowners]. This is consistent with the _____ principle in an earlier chapter.
2. **Tax Revenue versus Total Burden**
   Consider the analysis of the land tax in Figure 17–3 and the partial-equilibrium analysis of the structure tax in Figure 17–4.
   a. For the land tax, the loss to landowners is _____, computed as . . . . The revenue from the land tax is _____, computed as . . . .
   b. For the structure tax, the loss to consumers is _____, computed as . . . . The revenue from the structure tax is _____, computed as . . . .

c. The structure tax generates a deadweight loss because . . .
d. In contrast, the land tax does not generate a deadweight loss because . . .

3. **Passive Resistance versus Active Fleeing?**
The motto for this chapter comes from a famous movie scene in which Boris is about to be forced into fighting in a war. After the exchange with his friend Sonja, Boris scrambles out the door and tries to outrun military recruiters. How is this scene related to the issue of who bears the burden of a tax?

4. **Catatonia versus Fleetland**
In the state of Catatonia, there are two cities (Cat1 and Cat2), and people don't move from one city to another. In the state of Fleetland, residents are perfectly mobile between the state's two cities (Flee1 and Flee2). You just discovered that one city in each state (Cat1 and Flee1) will impose a structure tax next week, and you are the only person who knows about the upcoming taxes. You currently own 10 acres of land in each of the four cities.
a. If you want to keep a total of 20 acres in Catatonia, what if anything should you do?
b. If you want to keep a total of 20 acres in Fleetland, what if anything should you do?

5. **Effects the Property Tax on Different Individuals**
Consider the general-equilibrium view of the structure portion of the property tax. Based on the Taxton-Untax example in the chapter, compute the long-run effects of the structure tax on the following individuals. Assume that the regional supply of capital is fixed and consumers are perfectly mobile between cities in the region.
a. Rene, a renter in the taxing city, [gains, loses, isn't affected] $_____, computed as . . .
b. Landry, who owns land (three lots) in Taxton [gains, loses, isn't affected] $_____, computed as . . .
c. Loren, who owns land (two lots) in Untax [gains, loses, isn't affected] $_____, computed as . . .
d. Cap, who owns five structures in Taxton, [gains, loses, isn't affected] $_____, computed as . . .
e. Talulah, who owns five structures in Untax, [gains, loses, isn't affected] $_____, computed as . . .

6. **Consumer Mobility and Tax Shifting**
Consider the general-equilibrium view of the structure portion of the property tax.
a. Consumer mobility is good for _____ consumers and bad for _____ consumers because . . .
b. Consumer mobility is good for _____ landowners and bad for _____ landowners because . . .

7. **Corner Solution**
Using Figure 17–7 as a starting point, suppose the initial utility-maximizing spending on special education is $5 rather than $25. As before, the lump-sum

conditional grant for special education is \$20. The income elasticity of demand for special education is 1.0.

a. Show the response of the median household to the grant, with the new utility-maximizing point labeled *j*.

b. The grant increases spending on special education by _____ and increases spending on other goods by _____.

c. The lesson from this example is that the stimulative effect of a grant is relatively large when . . .

8. Education Lottery

Consider a city that initially spends \$20 million of its \$100 million budget on public schools, a choice consistent with the preferences of the median voter. The income elasticity of demand for public schools is 1.0. Suppose the city gets \$15 million from a new state lottery, and by law must spend all \$15 million on public schools.

a. Use a graph like Figure 17–7 to predict the effects of the lottery money on the city's spending on public schools and other goods.

b. Spending on public schools changes from _____ to _____ and spending on other goods changes from _____ to _____.

c. In other words, _____ percent of the lottery money goes to schools, and _____ percent goes to other goods.

9. Librarian Grant

Consider the hiring of city librarians. The daily wage of a librarian is \$100, and the city initially hires seven librarians for a total of \$700, leaving \$1,800 of its \$2,500 budget for other goods. Under a lump-sum conditional grant, the state gives the city \$500 to spend on librarians.

a. Suppose the grant causes the city to increase the number of librarians to 10. Use a graph like Figure 17–7 to show the city's response to the grant, with the pregrant choice labeled *i* and the postgrant choice labeled *c* (for conditional grant). If the city has 10 librarians, spending on other goods is _____.

b. Suppose the state switches to a matching grant for librarians, with a dollar-for-dollar match. Draw the budget line with the matching grant.

c. The matching-grant budget line [goes through, lies above, lies below] point *c*. The switch from the lump-sum grant to the matching grant [increases, decreases, does not change] the number of librarians because . . .

REFERENCES AND ADDITIONAL READING

1. Advisory Commission on Intergovernmental Relations. *Tax and Expenditure Limitations on Local Government.* Washington, DC: U.S. Government Printing Office (1995).
2. Beito, David T. *Taxpayers in Revolt*. Chapel Hill, NC: University of North Carolina Press, 1989.
3. Citrin, Jack. "Do People Want Something for Nothing: Public Opinion on Taxes and Spending." *National Tax Journal* 32 (1979), pp. 113–30.

4. Courant, Paul, Edward Gramlich, and Daniel Rubinfeld. “Why Voters Support Tax Limitations: The Michigan Case.” *National Tax Journal* 38 No. 1 (1980), pp. 1–20.
5. Courant, Paul N., and Susanna Loeb. “Centralization of School Finance in Michigan.” *Journal of Policy Analysis and Management* 16 (1997), pp. 114–35.
6. Duncombe, William, and John Yinger. “Why Is It So Hard to Help Central-City Schools?” *Journal of Policy Analysis and Management* 16 (1997), pp. 85–113.
7. Dye, Richard, and Therese McGuire. “The Effect of Property Tax Limitation Measures on Local Government Fiscal Behavior.” *Journal of Public Economics* 66 (1997), pp. 469–87.
8. Early, Dirk. “An Empirical Investigation of the Determinants of Street Homelessness.” *Journal of Housing Economics* 14 (2005), pp. 27–47.
9. Early, Dirk. “The Determinants of Homelessness and the Targeting of Housing Assistance.” *Journal of Urban Economics* 55 (2004), pp. 195–214.
10. Evans, William N., Sheila E. Murray, and Robert M. Schwab. “Schoolhouses, Courthouses, and Statehouses after Serrano.” *Journal of Policy Analysis and Management* 16 (1997), pp. 10–31.
11. Filimon, R., T. Romer, and H. Rosenthal. “Asymmetric Information and Agenda Control: The Bases of Monopoly Power and Public Spending.” *Journal of Public Economics* 17 (1982), pp. 51–70.
12. Inman, Robert P., and Daniel L. Rubinfeld. “Rethinking Federalism.” *Journal of Economic Perspectives* 11:4 (Fall 1997), pp. 43–65.
13. Oates, Wallace. “An Essay on Fiscal Federalism.” *Journal of Economic Literature* 37 (1999), pp. 1120–149.
14. O’Sullivan, Arthur. “Limits on Local Property Taxation.” Chapter 7 in *Property Taxation and Local Government Finance,* ed. Wallace E. Oates. Cambridge MA: Lincoln Institute, 2001.
15. Sexton, Terri, Steven M. Sheffrin, and Arthur O’Sullivan. “Proposition 13: Unintended Effects and Feasible Reforms.” *National Tax Journal* 52 (1999), pp. 99–112.
16. Shadbegian, Ronald. “Do Tax and Expenditure Limitations Affect Local Government Budgets? Evidence from Panel Data.” *Public Finance Review* 26 (1998), pp. 218–36.
17. Shadbegian, Ronald. “The Effect of Tax and Expenditure Limitations on the Revenue Structure of Local Government, 1962–87.” *National Tax Journal* 52 (1999), pp. 221–36.
18. Zodrow, George R. “The Property Tax as a Capital Tax: A Room with Three Views.” *National Tax Journal* 54 (2001), pp. 139–56.

APPENDIX

Tools of Microeconomics

This appendix reviews some of the basic tools of microeconomics used in various parts of the book. The appendix is divided into five sections:

1. **Marginal decision making.** The marginal principle tells us to pick the level of an activity where the marginal benefit equals the marginal cost.
2. **Product market.** The model of supply and demand shows how consumer prices are determined. We look at the market equilibrium and explore the issue of market efficiency.
3. **Labor market.** The model of labor supply and demand shows how wages and employment levels are determined. We look at the market equilibrium and discuss efficiency issues.
4. **Consumer choice.** The consumer choice model shows how consumers maximize their utility when subject to constraints imposed by their income and consumer prices.
5. **Input choice.** The input choice model is the production analog of the consumer choice model. It shows how firms pick the cost-minimizing combination of production inputs.

1. THE MARGINAL PRINCIPLE

The marginal principle provides a simple decision-making rule that helps individuals and organizations make decisions. The marginal benefit of some activity is the extra benefit from a small increase in the activity: for example, the extra revenue from keeping a barbershop open for one more hour. The marginal cost is the additional cost from a small increase in the activity: for example, the additional expense incurred by keeping a barbershop open for one more hour. Therefore, the *marginal principle* can be defined as follows:

If the marginal benefit of an activity exceeds the marginal cost, do more of it. If possible, pick the level at which the marginal benefit equals the marginal cost.

Applying the marginal principle to the barber's problem, the barber should stay open for one more hour if the extra revenue from the additional hour is at least as large as the extra cost.

Thinking at the margin enables us to fine-tune our decisions. We can use the marginal principle to determine whether a one-unit increase in a particular variable would make us better off. Just as a barber could decide whether to keep the shop open for one more hour, you could decide whether to study one more hour for a psychology midterm, and a firm could decide whether to hire one more worker. When we reach the level where the marginal benefit equals the marginal cost, the fine-tuning is done.

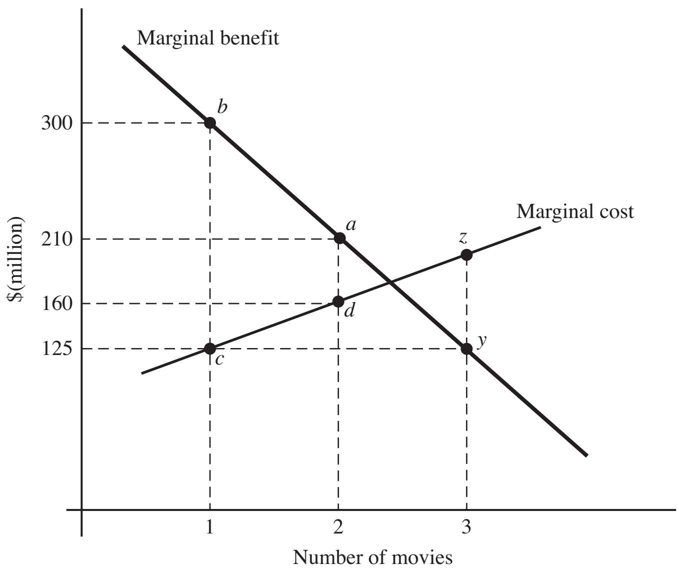
1.1 Example: How Many Movie Sequels?

To illustrate the marginal principle, consider a movie producer who must decide how many sequels to produce. Suppose the original version of a movie is successful enough that we expect a sequel to be profitable too. If the sequel turns out to be profitable, the producer then has to decide whether to make a third movie, then a fourth, and so on. The producer could use the marginal principle to figure out when to stop, thus avoiding the banana problem: Beginning spellers know how to start spelling banana, but often don't know when to stop—ba-na-na-na-na-na. . . .

Figure A–1 has two curves, one showing the marginal benefit of movies in a series, and a second showing the marginal cost. Consider the benefit curve first. A general rule of thumb in the movie business is that a sequel generates about 30 percent less revenue than the original, and revenue continues to drop for additional movies. In Figure A–1, the marginal benefit is the revenue generated by each movie, which drops from \$300 million for the first (original) movie (point *b*) to \$210 million for the second movie (point *a*), to \$125 million for the third movie (point *y*).

Consider next the cost curve in Figure A–1. The typical movie costs about \$50 million to produce and \$75 million to promote. At point *c* on the cost curve, the marginal cost of the first

FIGURE A–1 The Marginal Principle



The marginal benefit of movies in a series decreases as revenue drops for each additional movie, while the marginal cost increases because actors demand higher salaries. The marginal benefit exceeds the marginal cost for the first two movies, so it is sensible to produce two, but not three movies.

movie is $125 million. The marginal cost increases with the number of movies because film stars typically demand higher salaries to appear in sequels. For example, Angelina Jolie was paid more for *Tomb Raider 2* than for *Tomb Raider,* and the actors in *Charlie's Angels 2, Legally Blond 2, and Bad Boys 2* received raises too. The marginal-cost curve is positively sloped, with a cost of $160 million for the second movie (point *d*) and an even higher cost for the third (point *z*), reflecting the rising cost of hiring movie stars for sequels.

In this example, the first two movies are profitable, but the third is not. For the first movie, the marginal benefit ($300 million) exceeds the marginal cost ($125 million), generating a profit of $175 million. Although the second movie has a smaller benefit and a bigger cost, it is profitable because the marginal benefit still exceeds the marginal cost by $50 million ($210 − $160). In contrast, the marginal cost of the third movie ($195 at point *z*) exceeds the marginal benefit ($125 at point *y*), so the third movie is a losing proposition. In this example, the producer should stop after the second movie.

Although this example shows that only two movies in a series are profitable, other outcomes are possible. If the revenue from the third movie were higher or the cost were lower, the marginal benefit could exceed the marginal cost, and making a third movie would be profitable. Indeed, there are many examples of movies with multiple sequels and prequels, including *The Pink Panther, Star Wars,* and *Rocky*. Conversely, there are many examples of profitable movies that didn't generate any sequels. In these cases, the expected drop-off in revenues and run-up in costs were large enough to make a sequel unprofitable. In Figure A–1, if the marginal-benefit and marginal-cost curves were much steeper, the marginal benefit for the second movie would be less than the marginal cost, so a sequel would not be profitable.

1.2 Measuring the Surplus

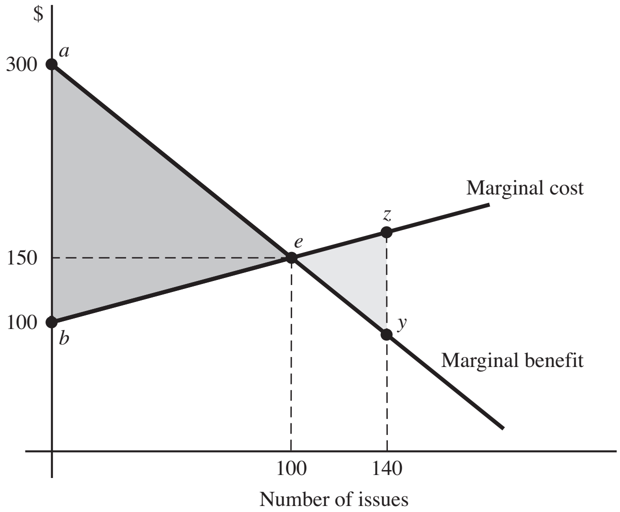
The marginal approach allows us to compute the net benefit or surplus from a particular activity. In the movie example, the surplus is the sum of the profits from the two movies produced. The profit is shown by the gap between the marginal benefit and the marginal cost. The gap is $175 million for the first movie ($300 − $125) and $50 million for the second ($210 − $160), so the total surplus is $225 million. In general, to compute the surplus or net benefit, we add the gaps between the marginal-benefit and marginal-cost curves across the quantity produced.

Figure A–2 (page 458) shows how to compute the surplus when the quantity produced is larger. We can change our example to consider serial literature or comic books. In 1836 and 1837, Charles Dickens wrote monthly installments of *The Pickwick Papers,* and his decision each month was whether to write another installment. More recently, the producers of comic books decided each month whether to issue another installment of *Superman* or *Donald Duck*. In Figure A–2, as the number of issues increases, the marginal benefit decreases and the marginal cost increases. The marginal principle is satisfied at point *e*, with 100 issues.

The area between the marginal-benefit and marginal-cost curves provides a good approximation of the actual surplus from the serial. To compute the actual surplus, we would add the surpluses (the gaps between the marginal benefit and marginal cost) for the first issue, the second issue, and so on up to the 100th issue. The darkly shaded area (triangle *aeb*) is a good approximation because the number of issues is relatively large.

What happens if we go beyond the point that satisfies the marginal principle? For example, suppose the serial producer suffers from the banana problem (he doesn't know when to stop) and

FIGURE A–2 Computing the Surplus



The net benefit or surplus from an activity is the area between the marginal-benefit and marginal-cost curves up to the quantity chosen. Triangle *aeb* is the surplus associated with satisfying the marginal principle at point *e*. Triangle *ezy* is the loss from going too far.

produces 140 issues. The loss associated with an excessive quantity is shown by the gap between the marginal-cost and marginal-benefit curves beyond the point that satisfies the marginal principle. For example, for the 101st through the 140th issue this loss is shown by the darkly shaded area (triangle *ezy*). The net benefit or surplus from the entire serial is the area of triangle *aeb* (the surplus from producing the first 100 installments) minus the area of triangle *ezy* (the loss from going too far).

2. EQUILIBRIUM AND EFFICIENCY IN A PRODUCT MARKET

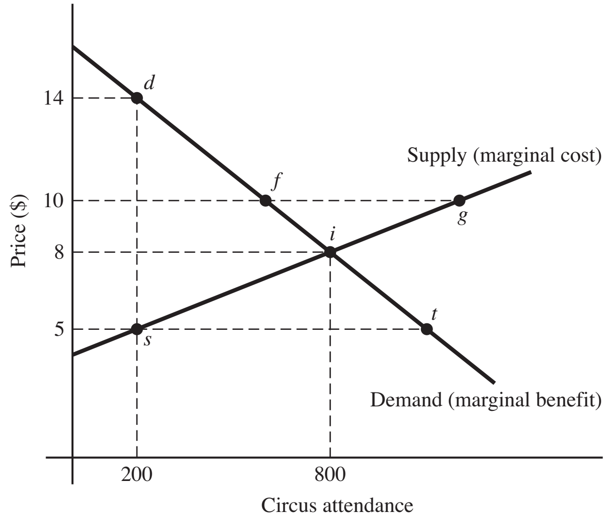
Economists use the model of supply and demand to determine equilibrium prices and quantities. In this book, we use the model in several chapters to explore the effects of public policies on product prices and quantities.

2.1 The Demand for a Product

Figure A–3 shows a market demand curve for seeing the circus. The demand curve is negatively sloped, indicating that an increase in the price decreases the quantity of people who see the circus. This occurs for two reasons:

- **Substitution effect.** An increase in the price of admission increases the cost of seeing a circus relative to the cost of other consumer goods. As a result, consumers will cut back on circus attendance in favor of seeing more movies, reading more books, or going to zoos or comedy clubs.

FIGURE A–3 Supply, Demand, and Equilibrium in the Circus Market



The demand curve is a marginal benefit curve, and the supply curve is a marginal cost curve. Equilibrium occurs at point *i*, where the quantity demanded equals the quantity supplied. The equilibrium price is $8. For a lower price (e.g., $5), there is excess demand, as shown by points *s* and *t*. For a higher price (e.g., $10), there is excess supply, as shown by points *f* and *g*.

- **Income effect.** An increase in price means that a consumer can no longer afford the original bundle of circuses and other goods (food, housing, entertainment). In other words, an increase in price decreases the consumer's real income. The consumer must cut back on something, and the circus is one candidate for cutting back. If circus attendance is a "normal" good, its consumption increases when real income increases and decreases when income drops.

The demand curve of a "normal" good is negatively sloped because the income effect reinforces the substitution effect. Both effects tend to reduce the quantity demanded when the price increases.

The demand curve shows the marginal benefit of consuming a good, so it is also a marginal-benefit curve. To see this, consider point *d*, which indicates that when the price is $14, a total of 200 people attend circuses. If the price were slightly higher (for example, $14.02), the 200th person won't go to the circus because the benefit is less than the $14.02 cost. But when the price drops to $14.00, this consumer goes to the circus because now the benefit exceeds the $14.00 cost. So in this case the marginal benefit of the circus for the 200th consumer must be $14.01, or as an approximation, $14. Similarly, the 800th person goes to the circus when the price drops to $8, so the marginal benefit for the 800th consumer is $8.

2.2 The Supply of a Product

Figure A–3 also shows the market supply curve for circus performances. The supply curve is positively sloped, indicating that an increase in price increases the number of people who can see the circus: Firms put on more shows and have them in larger venues so that more people can watch. The higher the price, the larger the quantity supplied.

The supply curve is also a marginal-cost curve, showing the marginal cost of entertaining people in circuses. To see this, consider point *s*, which tells us that when the price is \$5, firms are willing to perform for a total of 200 people. If the price were slightly lower (for example, \$4.98), no firm would serve the 200th person because the \$4.98 price doesn't cover the firm's marginal cost. When the price rises to \$5, a firm serves the 200th customer because the price is now high enough to just cover the cost. So in this case the firm's marginal cost must be \$4.99, or as an approximation, \$5. Similarly, a firm will serve the 800th consumer when the price reaches \$8, so the marginal cost of serving the 800th consumer is about \$8.

Why is the supply curve positively sloped? The supply curve is a marginal-cost curve, so it is positively sloped because of rising marginal cost. Consider the long-run, a period long enough that circus firms can change all their inputs, including labor and capital. The appeal of circus performers such as trapeze artists, elephants, and jugglers is their rarity. As the number of circus performances increases, circuses need more of these scarce inputs, and bidding among competing circuses pushes up the input prices. For example, the scarcity of bearded ladies means that as the circus industry expands, they earn higher wages.

The general idea is that a supply curve is positively sloped because as an industry expands, firms bid up the prices of scarce inputs. In the book, we discuss several markets subject to rising input prices, including housing and gasoline.

- **Housing and land prices.** The scarce input in the production of housing is land, and as the number of houses built increases, so does the price of land.
- **Gasoline and crude oil.** The scarce input in the production of gasoline is crude oil, the price of which rises with the total production of gasoline.

2.3 Equilibrium in the Product Market

As in other markets, equilibrium in the product market is shown by the intersection of the supply curve and the demand curve. In Figure A–3, this happens at point *i*, with a price of \$8 and a quantity of 800 circus viewers. At any other price, the quantity demanded will differ from the quantity supplied, resulting in pressure to increase or decrease the price.

Consider first what happens when the price is below the equilibrium level. For example, at a price of \$5, the quantity demanded (point *t*) exceeds the quantity supplied (point *s*), so there will be excess demand. Some consumers who want to see a circus at the relatively low price will be unable to do so. The long lines and disappointed consumers will produce pressure to increase the price. As the price increases, we move upward along the supply curve as firms stage more performances in bigger venues. At the same time, we move upward along the demand curve: Fewer consumers will want to see a circus at the higher price. The price will continue to rise until the excess demand is eliminated at point *i*.

Consider next what happens when the price is above the equilibrium level. For example, at a price of \$10, there is excess supply, with the quantity supplied (point *g*) exceeding the quantity demanded (point *f*). In other words, there aren't enough consumers to fill the circus tents. Competition among firms will cause the price to drop. As the price decreases, we move downward along the demand curve because the lower price encourages more consumers to see the circus. At the same time, we move downward along the supply curve as circus firms reduce the number of performances, perform in smaller tents, or go out of business. The price will continue to drop until the excess supply is eliminated at point *i*.

2.4 Shifting the Curves

The demand curve shows the relationship between the price and the quantity of circus services demanded, *ceteris paribus* (all other things held fixed). What are the *cetera* (other things) that are *paria* (fixed in value) in drawing the curves? Once we've identified the other things, we have a list of other (nonprice) variables whose values determine the position of the curve. When the value of one of these other variables changes, the curve shifts to a new position.

On the demand side of the market, several variables are held fixed in drawing a particular demand curve. When the value of one of the variables changes, the curve shifts.

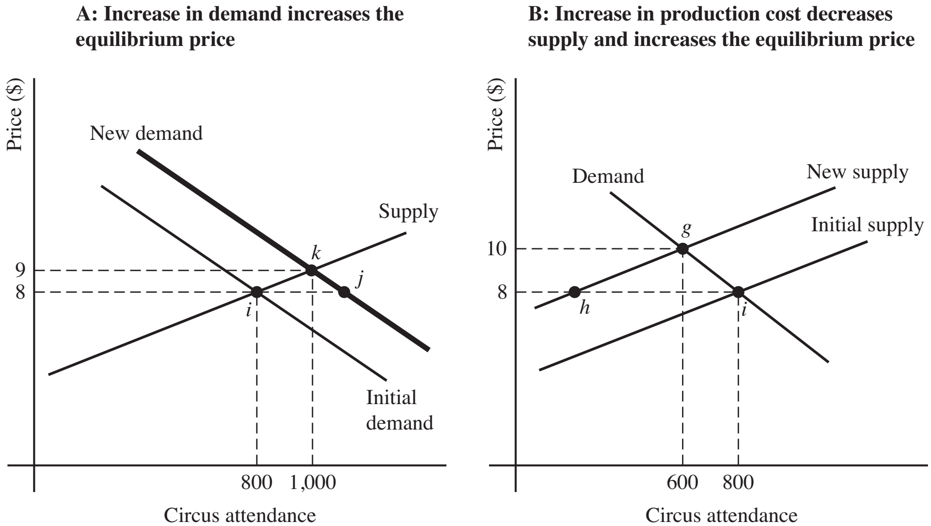
- **Consumer income.** An increase in income increases the demand for all "normal" goods, shifting the demand curve to the right.
- **Price of substitute products.** An increase in the price of a substitute good such as movies or books decreases the relative price of a circus and increases circus demand, shifting the demand curve to the right.
- **Price of complementary products.** An increase in the price of a complementary good such as peanuts or popcorn increases the total cost of an afternoon at the circus, decreasing the demand for circuses and shifting the demand curve to the left.
- **Preferences or tastes.** A change in preferences such as a greater desire to see jugglers and bearded ladies shifts the demand curve to the right.
- **Population.** If the market is defined geographically, an increase in the number of people shifts the demand curve to the right.

Panel A of Figure A–4 (page 462) shows the effects of an increase in demand. The demand curve shifts to the right, and at the original price of $8, there is excess demand: The quantity demanded (shown by point *j*) now exceeds the quantity supplied (point *i*). In the new equilibrium (shown by point *k*), the price is $9 and the quantity is 1,000.

On the supply side of the market, a number of variables are held fixed in drawing the supply curve:

- **Input prices.** Anything that increases the cost of producing a given quantity of output increases the marginal cost of production, shifting the supply curve upward. The sources of higher production costs include higher prices for raw materials (animal feed and fuel), higher capital costs (for tents and cages), and higher wages at each total output level. When production costs increase, firms are willing to supply less output at a given price, so the supply curve shifts to the left.
- **Labor productivity.** An increase in labor productivity means less labor time is required to produce each unit of output, so production costs drop, shifting the supply curve downward and to the right.
- **Technology.** Innovations that cut production costs shift the supply curve downward and to the right.

FIGURE A–4 Changes in Supply and Demand



A: An increase in demand shifts the demand curve to the right, causing excess demand that increases the equilibrium price from $8 (point *i*) to $9 (point *k*).

B. An increase in production cost shifts the supply curve upward and to the left, causing excess demand that increases the equilibrium price from $8 (point *i*) to $10 (point *g*).

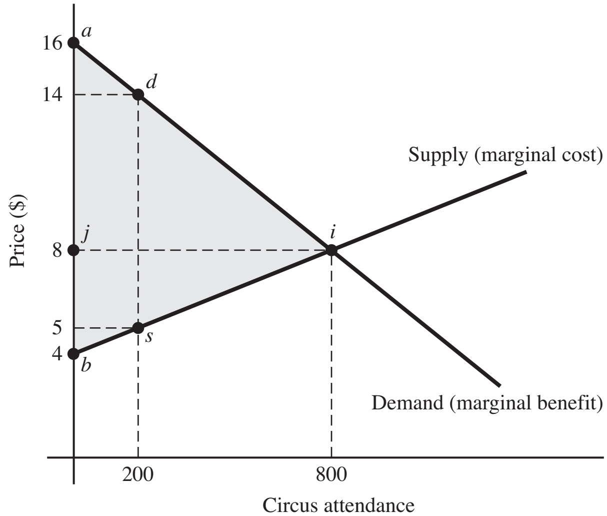
Panel B of Figure A–4 shows the effects of an increase in production cost. The supply curve shifts up and to the left, and at the original price of $8, there is excess demand: The quantity demanded (shown by point *i*) now exceeds the quantity supplied (point *h*). In the new equilibrium (shown by point *g*) the equilibrium price is $10 and the equilibrium quantity is 600.

2.5 Market Surplus

Earlier in the appendix, we used the marginal-benefit and marginal-cost curves to measure the surplus from an activity. We can use the supply and demand curves to measure the surplus or total value of the circus market. The demand curve shows the marginal benefit of the circus for the individual consumer. Assuming there are no external benefits from attending the circus, this is also the marginal social benefit of the circus. The supply curve shows the marginal cost of the circus, and assuming there are no external costs, it also shows the marginal social cost of the circus.

The market equilibrium shown in Figure A–5 maximizes the market surplus because it satisfies the marginal principle. At point *i*, the marginal-benefit curve (demand curve) intersects the marginal-cost curve (supply curve), and the market surplus is the shaded triangle *aib* between the demand curve and the supply curve. This is the best we can do. If we were to stop short of the equilibrium quantity, the marginal benefit of one more circus patron would exceed the marginal cost, so we could increase the surplus by moving toward the market equilibrium. If we were to go beyond the equilibrium quantity, the marginal cost would exceed the marginal benefit, so we could increase the surplus by moving back toward the market equilibrium.

FIGURE A–5 The Market Equilibrium Maximizes the Market Surplus



If there are no external costs or benefits, the market equilibrium satisfies the marginal principle, and the market equilibrium maximizes the market surplus, measured as the area between the demand (marginal benefit) and supply (marginal cost) curves. Consumer surplus is shown by triangle *aij* and producer surplus is shown by triangle *jib*.

2.6 Consumer Surplus and Producer Surplus

We can divide the market surplus into two surpluses, one gained by consumers and a second gained by producers. The surplus for an individual consumer equals the gap between the marginal benefit of consuming a product and the price paid for the product. The demand curve shows the marginal benefit to consumers. In Figure A–5, the 200th consumer has a marginal benefit of \$14 (point *d*) and pays a price of \$8, so that consumer's surplus is \$6.

We can add the surpluses of individual consumers to get the market consumer surplus. In Figure A–5, the market consumer surplus is shown by the area between the demand (marginal benefit) curve and the dashed line at the equilibrium price of \$8. In other words, the market consumer surplus equals the area of triangle *aij*. The area of this triangle is half its height times its base, or $\$3,200 = 0.50 \cdot (\$16 - \$8) \cdot 800$.

The producer surplus is a measure of the net benefit of the market for producers. The producer surplus for an individual producer equals the price received for the product minus the marginal cost of producing it. The supply curve shows the marginal cost of producing the product. In Figure A–5, the firm that serves the 200th consumer has a marginal cost of \$5 (point *s*) and gets a price of \$8, so its producer surplus is \$3. We can add the surpluses for different producers to get the market producer surplus. In Figure A–5, the market producer surplus is shown by the area between the supply (marginal cost) curve and the dashed line at the equilibrium price of \$8. In other words, the market producer surplus is the area of triangle *jib*. The area of this triangle is half its height times its base, or $\$1,600 = 0.50 \cdot (\$8 - \$4) \cdot 800$.

2.7 Inefficiency with Externalities

When the production of a product generates external costs, the market equilibrium does not maximize the market surplus. Recall the third axiom of urban economics:

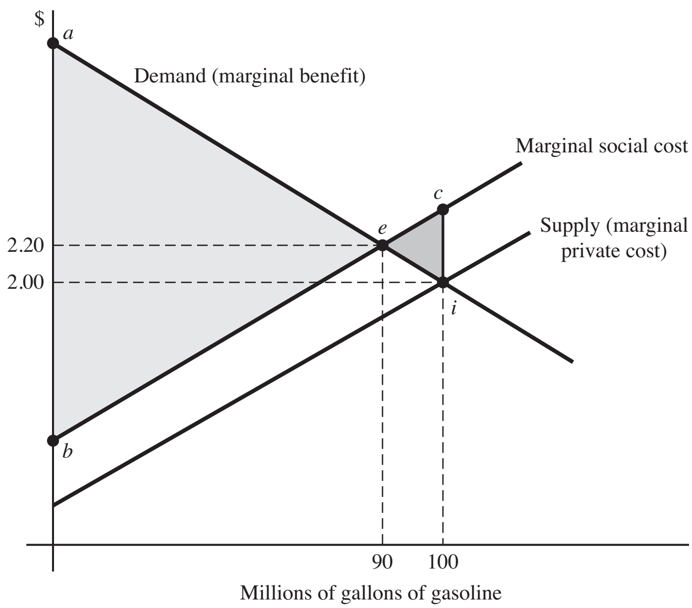
Externalities cause inefficiency

As we saw in Figure A–5, in a market without externalities, the market equilibrium is efficient in the sense that it maximizes the total surplus of the market. Things are different when there are externalities.

Consider the market for gasoline. Using gasoline as a car fuel causes air pollution and generates greenhouse gases, so the marginal social cost of gasoline exceeds the marginal private cost of producing it. A gasoline producer pays its input suppliers, including workers and crude oil suppliers, and the supply curve (marginal-cost curve) includes these costs. The supply curve does not incorporate the cost of emissions, however, so the marginal social cost of gasoline consumption exceeds the marginal private cost. In Figure A–6, the marginal social cost curve lies above the supply curve, with a gap equal to the marginal external cost of pollution.

We can use Figure A–6 to show the socially efficient quantity of gasoline. To satisfy the marginal principle, we find the quantity at which the marginal benefit of gasoline (shown by the demand curve) equals the marginal social cost. This happens at point *e*, with 90 million gallons of gasoline.

FIGURE A–6 External Cost and Inefficiency



The marginal principle is satisfied at point *e*, so the total surplus of the market is shown by triangle *aeb*. The market equilibrium, shown by point *i*, generates an excessive quantity, with a loss from producing too much, shown by the triangle *eci*.

The total surplus of the market is the area between the demand curve and the marginal social cost curve, up to the socially efficient quantity. This is shown as the lightly shaded triangle *aeb*.

What happens if we go beyond point *e* and reach the market equilibrium at point *i*? For the last 10 million gallons, the marginal social cost exceeds the marginal benefit, so the surplus of the market decreases as the quantity increases. The loss associated with producing too much is shown by the darkly shaded triangle *eci*. This is the area between the marginal social cost curve and the demand curve, from the efficient quantity (90 million gallons) to the equilibrium quantity (100 million gallons). Point *i* violates the marginal principle, and triangle *eci* measures the social loss or dead-weight loss from going too far.

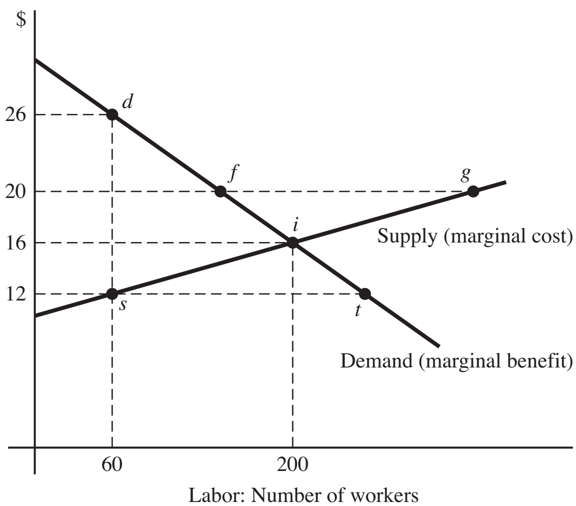
3. THE LABOR MARKET

Economists use the model of labor supply and demand to determine equilibrium wages and employment. In this book, we use the model in several chapters to explore various issues in the labor market.

3.1 The Demand for Labor

Figure A–7 shows a market demand curve for labor. The demand for labor comes from firms and other producers, and the demand for labor is derived from the demand for products. The labor

FIGURE A–7 Labor Market Equilibrium



The demand curve is a marginal benefit curve, and the supply curve is a marginal cost curve. Equilibrium occurs at point *i*, where the quantity demanded equals the quantity supplied. The equilibrium wage is $16. For a lower wage (e.g., $12), there is excess demand, as shown by points *s* and *t*. For a higher wage (e.g., $20), there is excess supply, as shown by points *f* and *g*.

demand curve is negatively sloped, indicating that an increase in the wage decreases the quantity of labor demanded. This occurs for two reasons:

- **Substitution effect.** An increase in the wage causes firms to substitute other inputs (capital, land, materials) for the relatively expensive labor.
- **Output effect.** An increase in the wage increases production costs, increasing the prices of the products produced with labor. Consumers respond to higher prices by purchasing less output, so firms produce less and hire fewer workers.

The demand curve is negatively sloped because an increase in wages generates both a substitution effect and an output effect.

The demand curve is also a marginal-benefit curve. To see this, consider point *d*, which tells us that when the wage is $26, a total of 60 workers will be hired. If the wage were slightly higher (for example, $26.02), the firm would not hire the 60th worker because the wage exceeds the firm's benefit from the worker (the value of output produced). But when the wage drops to $26, the firm hires the worker because now the benefit is just above the wage. So in this case the marginal benefit of hiring the worker must be $26.01, or as an approximation, $26. Similarly, the firm hires the 200th worker when the wage drops to 16, so the marginal benefit of the 200th worker is $16.

3.2 The Supply of Labor

The supply of labor comes from workers who have the skills required for a particular job. In Figure A–7, the market supply curve is positively sloped, indicating that an increase in the wage increases the quantity of labor supplied. The supply curve shows the number of workers at different wages and implicitly assumes that each worker works the same number of hours, independent of the wage. This assumption simplifies matters because we don't have to keep track of hours worked by each worker, just the number of workers. The empirical evidence on labor supply suggests that an increase in the wage has a negligible effect on the aggregate hours worked: Some people work more and others work less, but on average, people work about the same number of hours.

Why is the labor supply curve positively sloped? If we ignore space and geography for the moment, the positive slope results from the fact that people have different opportunity costs of work time. At the low end of the supply curve, the 60th worker joins the labor market when the wage reaches $12, reflecting a relatively low opportunity cost of work time. Further up the supply curve at point *i*, the 200th worker joins when the wage reaches $16, reflecting a higher opportunity cost. In general, as the wage rises, the market attracts workers with progressively higher opportunity costs.

In an urban context, geography matters, and the supply curve reflects the migration of workers between cities. As one city's wage rises, the city becomes more attractive relative to other cities in the region. As a result, workers will migrate to the city, increasing the quantity of labor supply as the city moves upward along its labor supply curve. In this context, the positive slope indicates that a higher wage attracts more workers to a city.

The supply curve is also a marginal-cost curve for labor. To see this, consider point *s*, which tells us that when the wage is $12, a total of 60 workers are willing to work in the market. If the wage were slightly lower (for example, $11.98), the 60th person wouldn't work because the opportunity cost of working exceeds the wage. But when the wage rises to $12, the person joins the workforce because now the wage exceeds the opportunity cost. So in this case, the marginal cost of the 60th worker must be $11.99, or as an approximation, $12. Similarly, the 200th worker joins the market when the wage reaches $16, so the marginal cost is about $16.

3.3 Equilibrium in the Labor Market

As in other markets, equilibrium in the labor market is shown by the intersection of the supply curve and demand curve. In Figure A–7, this happens at point *i*, with a wage of $16 and 200 workers. At any other wage, the quantity demanded will differ from the quantity supplied, resulting in pressure to change the wage.

Consider first what happens when the wage is below the equilibrium level. For example, at a wage of $12, the quantity demanded (point *t*) exceeds the quantity supplied (point *s*), so there will be excess demand for labor. Some firms will be unable to hire as many workers as they want, and competition among firms for a relatively small number of workers will bid up the wage. As the wage increases, we move upward along the supply curve because more workers enter the market, attracted by the higher wage. At the same time, we move upward along the demand curve, because firms demand fewer workers at the higher wage. The wage continues to rise until the excess demand is eliminated at point *i*.

Consider next a wage above the equilibrium level. As shown by points *f* and *g*, the quantity of labor supplied exceeds the quantity demanded. There is excess supply of labor, so some people looking for jobs won't find any. Competition among workers for the relatively small number of jobs will bid down the wage. As the wage decreases, we move downward along the demand curve as firms hire more workers. At the same time, we move downward along the supply curve, with some workers dropping out of the market. The wage continues to drop until excess supply is eliminated at point *i*.

3.4 Shifting the Curves

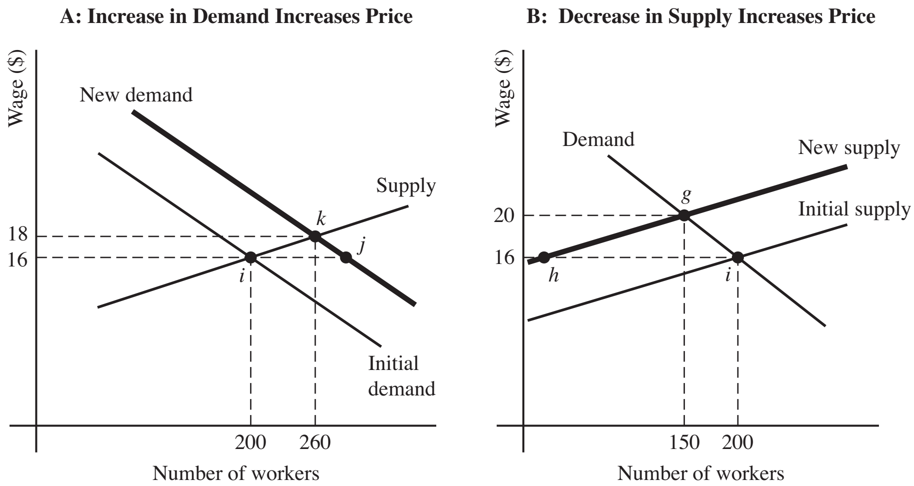
The supply and demand curves in this case show the relationship between the wage and the quantity of labor supplied or demanded, *ceteris paribus* (all other things fixed). What are the *cetera* (other things) that are *paria* (fixed in value) in drawing the curves? Once we've identified the other things that are fixed, we have a list of other (nonwage) variables whose values determine the position of the curve. When the value of one of these other variables changes, the curve shifts to a new position.

Recall that the demand curve is a marginal-benefit curve, showing the benefit of hiring workers. The following changes will increase the marginal benefit of hiring workers, shifting the demand curve upward:

- **Price of output.** If the price of the product produced by workers increases, each worker will generate more revenue for the firm.
- **Productivity.** If output per worker increases, each worker will generate more revenue for the firm. The possible sources of productivity gains include an increase in labor skills or an increase in capital (machines and equipment) per worker.

These changes also shift the demand curve to the right: At a given wage, a firm will want to hire more workers.

FIGURE A–8 Changes in Labor Supply and Demand



A: An increase in demand shifts the demand curve to the right, causing excess demand that increases the equilibrium wage from $16 (point *i*) to $18 (point *k*).

B. A decrease in supply shifts the curve to the left, increasing the equilibrium wage from $16 (point *i*) to $20 (point *g*).

Panel A of Figure A–8 shows the effects of an increase in demand. The demand curve shifts upward and to the right. At the original price of $16, there is excess demand: The quantity demanded (shown by point *j*) now exceeds the quantity supplied (point *i*). In the new equilibrium shown by point *k*, the price is $18 and the quantity is 260.

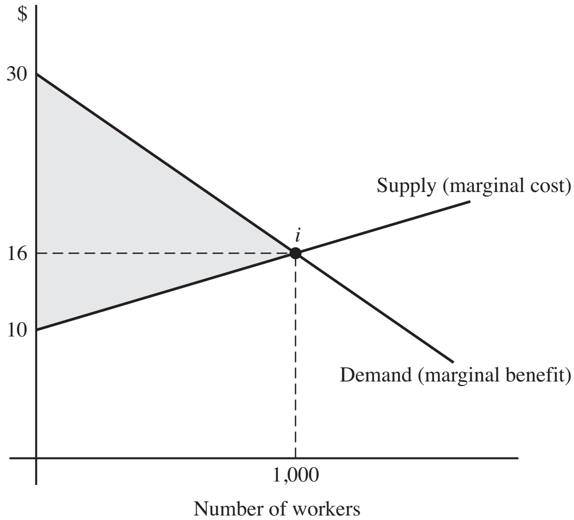
On the other side of the market, the labor supply curve shows how many workers participate in a labor market at each wage. In an urban context, anything that increases the relative attractiveness of a city (anything except the wage) will shift the entire curve to the right: More workers will be willing to work in the city. A city could become more attractive if it cuts pollution or improves public services. Conversely, anything that decreases the relative attractiveness of the city will shift the labor supply curve to the left.

In Panel B of Figure A–8, a decrease in labor supply shifts the supply curve up and to the left. At the original price of $16, there is excess demand: The quantity demanded (shown by point *i*) now exceeds the quantity supplied (point *h*). In the new equilibrium shown by point *g*, the price is $20 and the quantity is 150.

3.5 Market Surplus

We can use the supply and demand curves to measure the surplus or total value of the labor market. The demand curve shows the marginal benefit of labor for the firms that hire workers. Assuming there are no external benefits from labor, this is also the marginal social benefit of labor. The supply curve shows the marginal cost of labor to workers, and assuming there are no external costs, it also shows the marginal social cost of labor.

FIGURE A–9 The Labor Market Equilibrium Maximizes the Market Surplus



If there are no externalities, the market equilibrium satisfies the marginal principle, and the market equilibrium maximizes the market surplus, measured as the area between the demand (marginal benefit) curve and supply (marginal cost) curve.

The market equilibrium maximizes the surplus of the market because it satisfies the marginal principle. At point *i* in Figure A–9, the marginal-benefit curve (demand curve) intersects the marginal-cost curve (supply curve), and the market surplus is the shaded triangle between the demand curve and the supply curve. This is the best we can do. If we were to stop short of the equilibrium quantity, the marginal benefit of one more worker would exceed the marginal cost, so we could increase the surplus by moving toward the market equilibrium. If we were to go beyond the equilibrium quantity, the marginal cost would exceed the marginal benefit, so we could increase the surplus by moving back toward the market equilibrium.

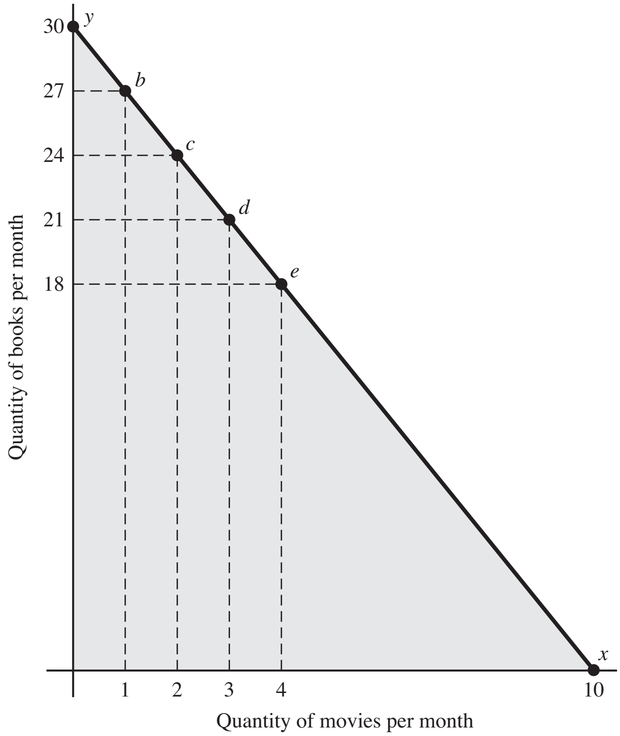
4. THE CONSUMER CHOICE MODEL

The consumer choice model shows how consumers make decisions about how much of a product to buy. The idea is that a consumer maximizes his or her utility, subject to the constraints imposed by product prices and the consumer's income. To illustrate, consider the decisions of Maxine, a consumer who must decide how many movies and paperback books to buy each month. Maxine has a fixed income per month to spend on the two goods, so her options are limited by her budget. To decide how to spend her money, Maxine takes two steps:

1. She figures out her menu of options, the list of alternative combinations of books and movies her budget allows.
2. She picks the combination of books and movies that generates the highest level of satisfaction.

We'll start with a discussion of Maxine's budget options, and then discuss her preferences.

FIGURE A–10 Budget Set and Budget Line



The budget set (shaded area) shows all the affordable combinations of books and movies, and the budget line (with endpoints *x* and *y*) shows the combinations that exhaust the budget.

4.1 Consumer Constraints: The Budget Line

Consider first the constraints faced by a consumer. Maxine's ability to purchase movies and other goods is limited by her income and the prices of movies and other products. Suppose Maxine has a fixed income of $30 per month, which she spends entirely on movies and used paperback books. The price of a movie is $3 and the price of a book is $1.

A budget line shows all the combinations of two goods that exhaust the consumer's budget. In Figure A–10, if Maxine spends her entire $30 budget on books, she gets 30 books and no movies (point *y*). At the other extreme, she can spend her entire budget on movies, getting 10 movies (point *x*). The points between these two extremes are possible too. For example, she could reach *b* (one movie and 27 books) by spending $3 on movies and $27 on books, or *c* (two movies and 24 books). A consumer's budget set is the set of all the affordable combinations of two goods. The budget set includes the budget line (combinations that exhaust the budget) as well as combinations that leave the consumer with extra money. In Figure A–10, Maxine's budget set is shown as a shaded triangle. She can afford any combination below the budget line but cannot afford combinations above it.

The budget line shows the market trade-off between books and movies. Starting from any point on the budget line, if Maxine buys one more movie, she diverts \$3 from book purchases, reducing the number of \$1 books she can purchase by three. The market trade-off equals the price ratio, the price of movies (\$3) divided by the price of books (\$1), or three books per movie. The market trade-off also equals the slope of the budget line, the "rise" (the change in books) divided by the "run" (the change in movies). If all consumers pay the same price for the two goods, they all have the same market trade-off of three books per movie.

4.2 Consumer Preferences: Indifference Curves

We've seen the consumer's budget set, which shows what the consumer can afford. The next step in our discussion of consumer choice is to look at what the consumer wants, what makes the consumer happy. Once we have a means of representing consumer preferences, we can show how a consumer makes her choice, picking the best of the combinations within the budget set.

We can represent the consumer's preferences with indifference curves. The idea behind an indifference curve is that there are different ways for a consumer to reach a particular level of satisfaction or utility. An indifference curve shows the different combinations of two goods that generate the same level of utility. In Figure A–11 (page 472), the indifference curve passing through points *b, z, m*, and *n* separates the combinations of books and movies into three groups:

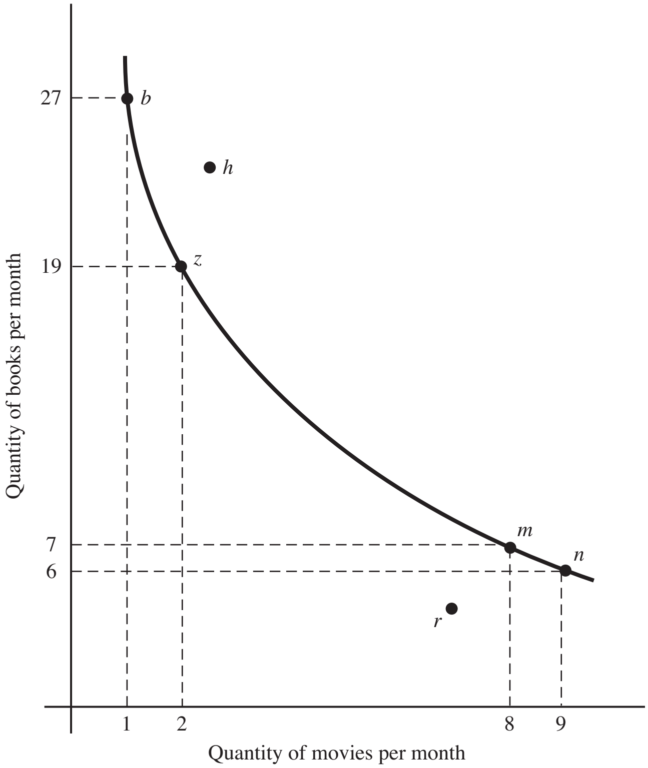
- **Superior combinations.** All the combinations above the indifference curve generate higher utility than combinations on the curve. Maxine would prefer point *h* to point *z* because she gets more of both goods at point *h*.
- **Inferior combinations.** All the combinations below the indifference curve generate lower utility than combinations on the curve. Maxine would prefer point *m* to point *r* because she gets more of both goods at point *m*.
- **Equivalent combinations.** All combinations along the indifference curve generate the same utility. Maxine is therefore indifferent between combinations *b, z, m*, and *n*.

An indifference curve shows the preferences of an individual consumer, so indifference curves vary from one consumer to another. Nonetheless, the indifference curves of all consumers share two characteristics: They are negatively sloped, and they become flatter as we move downward along a particular indifference curve.

Why is the indifference curve negatively sloped? If we increased Maxine's movie consumption by one unit without changing her book consumption, her utility would increase. To restore the original utility level, we must take away some books, and that's what happens along an indifference curve. To keep utility constant, there is a negative relationship between books and movies, so the indifference curve is negatively sloped. The slope of an indifference curve is the marginal rate of substitution (MRS) between the two goods, the rate at which a consumer is willing to substitute one good for another. In Figure A–11, if Maxine starts at point *b* and we give her one more movie, we take away eight books to keep her on the same indifference curve. Therefore, starting from point *b*, her marginal rate of substitution is eight books per movie. When she starts with many books and only one movie, she is willing to trade a lot of books to get one more movie.

The indifference curve becomes flatter as we move downward along the curve. This reflects the assumption that consumers prefer balanced consumption to extremes. As we move down Maxine's indifference curve, movie consumption increases while book consumption decreases.

FIGURE A–11 Indifference Curve and the Marginal Rate of Substitution

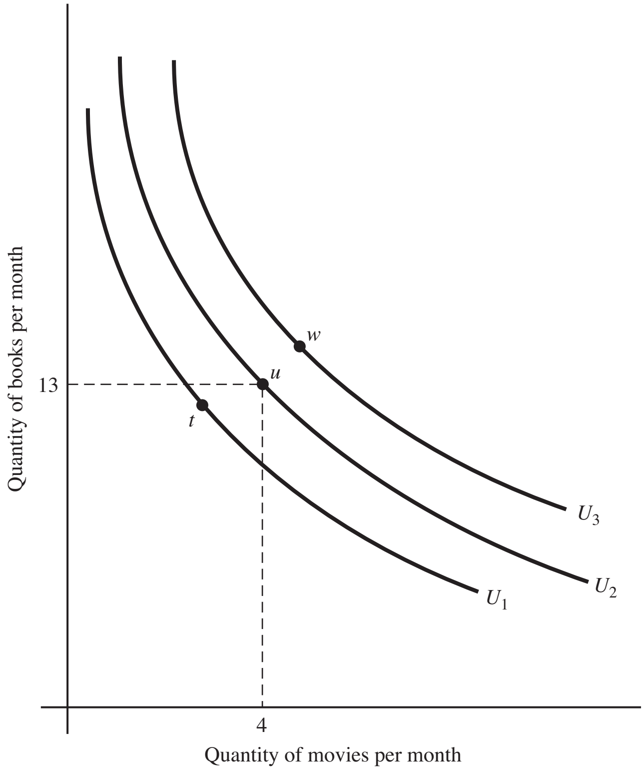


The indifference curve shows the different combinations of books and movies that generate the same utility level. The slope is the marginal rate of substitution (MRS) between the two products. The MRS is eight books per movie between points *b* and *z*, but only one book per movie between points *m* and *n*.

Starting from one extreme (few movies and many books), she is willing to sacrifice many books to get another movie: The MRS is large and the indifference curve is steep. For example, starting from point *b*, her MRS is eight books per movie. But as she gets more and more movies (and fewer and fewer books), she isn't willing to sacrifice as many books to get more movies. As a result, her MRS decreases, and the indifference curve becomes flatter. For example, between points *m* and *n*, the MRS is one book per movie.

An indifference map is a set of indifference curves, each with a different level of utility. Figure A–12 shows three indifference curves: U_1, U_2, and U_3. As Maxine moves from a point on indifference curve U_1 to any point on U_2, her utility increases. This is sensible because she can get more of both goods on U_2, so she will be better off. In general, Maxine's utility increases as she moves in the northeasterly direction to a higher indifference curve, from U_1 to U_2, and U_3, and so on.

FIGURE A–12 Indifference Map



An indifference map shows a set of indifference curves, with utility increasing as we move northeasterly to higher indifference curves (U_1 to U_2 to U_3).

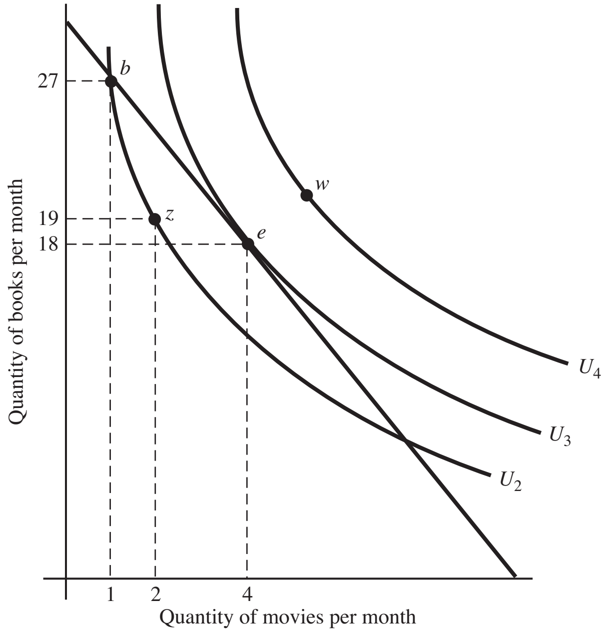
4.3 The Utility-Maximizing Rule

Maxine's objective is to maximize her utility, given her budget and the prices of movies and books. She can choose from many affordable combinations of books and movies, and she will pick the one that generates the highest level of utility. In graphical terms, Maxine will reach the highest indifference curve possible, given her budget set.

In Figure A–13 (page 474), Maxine maximizes her utility at point *e*, with four movies and 18 books. She achieves the utility level associated with indifference curve U_3. Why does she choose point *e* instead of other points such as *z, b*, or *w*?

- **Point *z*.** Maxine doesn't choose this point for two reasons. First, it is not on the budget line, so it does not exhaust her budget. Second, it is on a lower indifference curve than point *e*, so it generates less utility.
- **Point *b*.** Although point *b* exhausts Maxine's budget, it lies on a lower indifference curve than point *e*, so it generates less utility. Starting from point *b*, Maxine could reallocate her budget and buy more movies and fewer books. As she moves down her budget line, she

FIGURE A–13 Maximizing Utility: MRS = Price Ratio



To maximize utility, the consumer finds the combination of books and movies where an indifference curve is tangent to the budget line. At the utility-maximizing combination (point *e*), the marginal rate of substitution (the consumer's own trade-off, shown by the slope of the indifference curve) equals the price ratio (the market trade-off, shown by the slope of the budget line).

moves to progressively higher indifference curves, ultimately reaching point *e* on indifference curve U_3.

- **Point *w*.** Although point *w* is on a higher indifference curve and thus would generate a higher utility level than point *e*, it lies outside Maxine's budget set, so she cannot afford it.

At point *e*, Maxine reaches the highest indifference curve possible, given her budget set. Notice that at point *e*, the indifference curve touches—but does not pass through—the budget line. In other words, the indifference curve is tangent to the budget line.

What is the economic interpretation of the tangency condition? At the point of tangency, the slope of the indifference curve equals the slope of the budget line. The slope of the budget line equals the opportunity cost of movies, computed as the movie price ($3) divided by the book price ($1), or three books per movie. The slope of the indifference curve is the marginal rate of substitution (MRS), so if the two curves are tangent at point *e*, the MRS is also three books per movie. In other words, the consumer's trade-off between the two goods (the MRS) equals the market trade-off (the price ratio) between the two goods:

$$MRS = \frac{\text{price of movie}}{\text{price of book}}$$

To show why the tangent point is best, suppose Maxine tentatively chooses a point where the MRS is not equal to the price ratio. For example, starting at point *b*, the indifference curve is relatively steep, and the MRS is eight books per movie: She is willing to give up eight books to get a single movie. But given market trade-off, she can actually get that movie by sacrificing only three books, so she will move down her budget line and consume more movies. The same argument applies to any combination for which the MRS (the consumer's own trade-off) is not equal to the price ratio (the market trade-off). Anytime Maxine is willing to trade at a rate that is different from market trade-off, it will be in her best interest to do so. The benefits of moving along the budget line will be exhausted only when the MRS equals the price ratio. In Figure A–13, this happens at point *e*.

5. THE INPUT CHOICE MODEL

The input choice model shows how firms pick the best combination of inputs. There are many ways to produce a particular product, with different combinations of labor and capital (machines, buildings, and equipment). The idea behind the input choice model is that a firm will choose the input combination that minimizes the cost of producing a target quantity of output. To illustrate, Minnie produces catnip mice as cat toys, with a target production level of 100 toys per hour. She uses two inputs, capital and labor, and her objective is to minimize the cost of producing her target output level.

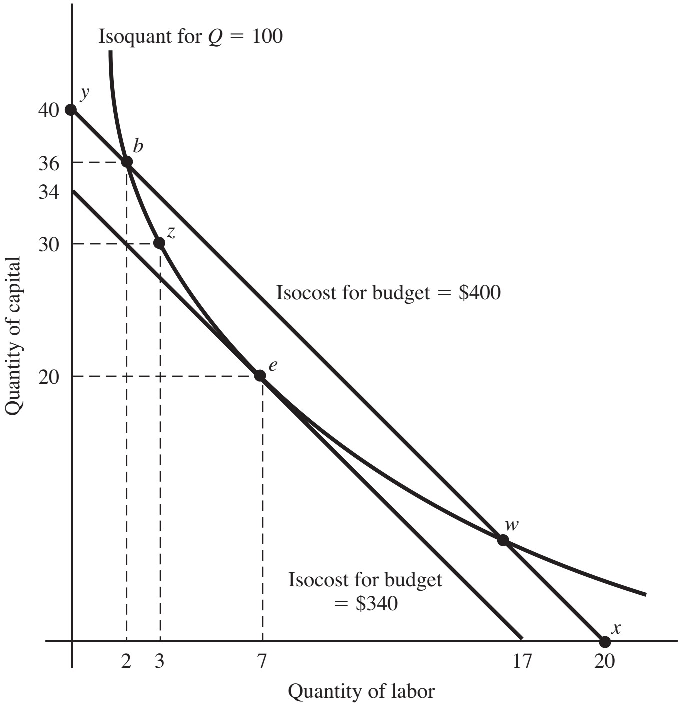
5.1 The Isoquant

An isoquant is the production analog of the consumer's indifference curve. It shows a set of production "recipes," different input mixtures that produce the same quantity of output (*iso* means equal in Greek). In Figure A–14 (page 476), the isoquant shows the different combinations of capital and labor that produce 100 toys. For example, Minnie can use 36 machines and two workers (point *b*) or 30 machines and three workers (point *z*). Further down the isoquant, points *e* and *w* show other input combinations that produce the target output.

The slope of the isoquant is the marginal rate of technical substitution (MRTS), the production analog of the marginal rate of substitution. The MRTS is the reduction in capital that offsets a one-unit increase in labor. For example, comparing point *b* to point *z*, if the firm adds one worker (going from two to three workers), it can reduce its capital by six units (from 36 to 30) and still produce the same quantity of output. As a firm adds more and more labor, the MRTS decreases. For example, the MRTS at point *e* is two units of capital per worker, and the MRTS at point *w* is less than one unit of capital per worker.

Why does the MRTS decrease as we move downward along the isoquant? A move down the isoquant increases the quantity of labor and decreases the quantity of capital. Starting from one extreme (many machines and few workers), adding a worker increases output by a large amount, so we can take away a large number of machines and produce just as much output: the MRTS is large and the isoquant is steep. For example, between points *b* and *z*, the MRTS is six machines per worker. But as we move downward along the isoquant to a point with more workers and fewer machines, there is less capital per worker, so adding another worker increases output by a smaller amount. Therefore, to keep output at the target level, we reduce capital by a smaller amount, for example two machines at point *e*.

FIGURE A–14 Minimizing Cost: MRTS = Input Price Ratio



The isoquant shows the input combinations that produce the target output quantity (Q = 100). The isocost shows the input combinations that exhaust a given budget. The firm's objective is to reach the lowest feasible isocost, the one tangent to the isoquant. At point *e*, the cost of producing the target output level is minimized at $340, compared to a budget of $400 at point *b* or point *w*. The MRTS (the slope of the isoquant) equals the input price ratio (the slope of the isocost).

5.2 Isocost Lines

An isocost is the production analog of the consumer budget line. It shows the combinations of two inputs that exhaust a given budget. Suppose Minnie can rent machines for $10 per hour and pays her workers $20 per hour. In Figure A–14, the higher of the two isocost lines shows the affordable input combinations for a budget of $400. At one extreme, Minnie could spend the entire $400 by getting 40 machines (point *y*); at the other extreme, she could spend it all on labor and hire 20 workers (point *x*). At point *b*, she can spend $360 on machines ($10 · 36 machines) and $40 on labor ($20 · 2 workers).

5.3 Minimizing Cost: MRTS = Input Price Ratio

The firm's objective is to minimize the cost of meeting its production target. Minnie's target output is 100 toys, so she wants to get on the lowest isocost line that makes contact with the

$Q = 100$ isoquant. Suppose she starts at point *b*, with a budget of \$400 spent on 36 machines and two workers. She could do better by moving from point *b* to point *z*. The MRTS between these two points is six machines per worker, so she can add one worker (+\$20) and get rid of six machines (−\$60), decreasing her total cost by \$40. If she moves from point *z* to point *e*, she can cut her cost another \$20, dropping it to \$340. At point *e*, she has reached the lowest isocost that makes contact with the isoquant, so she cannot do any better. The isocost is tangent to the isoquant, so she is minimizing her cost.

What is the economic interpretation of the tangency condition? At the point of tangency, the slope of the isoquant equals the slope of the isocost. As in the case of the consumer budget line, the slope of the isocost equals the price ratio. The price of labor is twice the price of capital, so the slope of the isocost is two machines per worker. The slope of the isoquant is the marginal rate of technical substitution (MRTS), so if the two curves are tangent at point *e*, the MRTS is two units of capital per labor. In other words, the production trade-off between the two inputs (the MRTS) equals the market trade-off (the input price ratio) between the two inputs:

$$MRTS = \frac{\text{price of labor}}{\text{price of capital}}$$

To show why the tangent point is best, suppose Minnie tentatively chooses point *b*. At this point, the isoquant is steeper than the isocost line, with MRTS = six machines per worker, compared to a market trade-off of two machines per worker. To keep output at the target level, she can eliminate six machines for each worker she hires (MRTS = 6), and because workers are only twice as expensive (price ratio = 2), substituting workers for machines cuts her cost. She will continue this input substitution until the production trade-off matches the budget trade-off of two machines per worker. This happens at point *e*, where the MRTS equals the input price ratio.

Subject and Author Index