City Economics

Brendan O'Flaherty

City Economics

City Economics

Brendan O'Flaherty

For Mary

Contents

Preface

If you could figure out how to cut traffic accidents in China by 2 percent, you would save more lives every year than were lost in the events of September 11. If you could reduce housing segregation in American cities, African American children would be healthier and would learn more in school. And when delayed buses bunch up, you sometimes save commuters more time by having more late buses than by having fewer.

This book is designed for people who want to examine issues like these. The book is about cities and the public policies that can make them better or worse places for living, working, and playing. This subject is traditionally called urban economics.

But this isn't a traditional urban economics book. The method of exposition is different, for one thing. Many people who want to think rigorously about cities are not accustomed to the way economists talk among themselves—the equations, proofs, complex graphs, and words that only the initiated understand. So I've tried to write for readers who have little or no background in economics or comfort with economists. There are no proofs, and only a few equations. The graphs are simple and less numerous than in most economics texts. I've minimized the number of technical terms, and those I do use are flagged in the text and defined again in a glossary. For instance, you won't see *utility, bid-rent curve, elasticity of substitution,* or *production function.* At the start of each chapter is a list of the economic concepts students need to know.

But while most of the trappings of academic economics are missing, the guts remain. Lazy or impatient readers will probably find parts of the text hard going, because the ideas are complex and subtle. Many of the end-of-chapter questions are challenging. My hope is that the book will appeal to students of urban studies, urban planning, public policy, and public administration, as well as to undergraduate economics students.

The choice of topics, and the emphasis I put on those topics, is also nontraditional. This book has two chapters on race, for instance, instead of the usual two pages. Such topics as traffic accidents, trolley conspiracies, water service, sewers, fire protection, garbage collection, police powers, street behavior, child care, housing codes, homelessness, incarceration, alcohol, illicit drugs, and tax abatements rarely if ever appear in other urban economics books, but they will be found here. Sprawl and technological change are also recurring themes throughout the book, and so receive more attention than in most economics textbooks.

In choosing topics, I asked myself what practical city actors would gain

from a careful analysis of each. By "practical city actors" I mean not just government officials but also leaders of neighborhood associations, labor unions, merchant associations, community groups, churches, and nonprofit organizations—anyone who can make something happen (or not happen) in cities. For almost every topic in the book, I can recall some discussion I've been a part of in an office, a storefront, a kitchen, a meeting room, or on a sidewalk.

(Most of my attention is focused on U.S. issues; I touch only lightly on cities in developing countries. The problems those cities face are of course incredibly important in determining how the twenty-first century will evolve. Fortunately, many principles are general enough that people who know those cities well will probably be able to translate some American lessons into useful ideas.)

Beyond choosing a coherent range of topics, I also looked for coherence among the subjects covered. As a result, many chapters reinforce each other. Buses tell us something about schools, for instance, optimal tolls tell us about land prices, and the arguments for the public provision of garbage collection are remarkably similar to the arguments for subsidizing safe occupational and recreational drugs.

The book is also consistent in its approach to policy. Here I am probably more traditional than the traditional books. I look at almost everything through the lens of Pareto optimality and potential Pareto optimality—how policies affect people and their well-being, not how they affect abstract entities like cities or the economy or growth or the environment. Such traditionalism leads to some radical questions, though. Should cities have police and fire departments? Should homeowners' tax preferences be repealed? Should whites be paid to move to minority neighborhoods or vice versa? Should public schools charge for their services?

Looking at issues from first principles also leads to serious consideration of policies that are not now part of the mainstream: pay-at-the-pump motor vehicle insurance, curb rights for buses, land taxes, marginal-cost water pricing, sidewalk zoning, housing partnerships, and FDA approval of safe recreational drugs.

People who regard themselves as practical may wonder why I pay attention to ideas outside common discourse. I do this because I want readers to think in new ways, to employ fresh perspectives. There would be no profit in a book devoted solely to policies we are all aware of. Every idea feels strange the first time you encounter it—that's how you can tell you're learning.

And learning, after all, is this book's goal. Cities enrich our lives so much that educated people can ill afford to be ignorant about them. Systematic thinking can make cities healthier, more productive, less wasteful, more fair, and more fun. This book shows how.

Acknowledgments

City Economics has been a long time in the making. My earliest memories are of the fresh smell of green palm cards outside the Roseville Avenue polling place during the 1954 municipal election in Newark, New Jersey, and of the wide stairway and ample crayon supply at the state CIO union headquarters on High Street. (Palm cards are cards that were given to voters as they approached the polls, asking them to vote a certain way, particularly for minor offices.) My parents, Annette and Charles O'Flaherty, started early on trying to make me understand how cities actually work and how they should work, and they never gave up.

Decades after Roseville Avenue, Harry Wheeler, John Caufield, Ken Gibson, and Tom Banker—among many others—continued my education in Newark City Hall. Since then, in many battles and skirmishes in Newark and in Essex County, I've been inside and outside, right and wrong, a winner and a loser (mainly the latter), but I've always come away with a better understanding (in retrospect) of urban economics and politics. I've been helped by both allies and adversaries.

I want especially to acknowledge Archie Williams's contribution. For two dozen years he has peppered me with questions, pushed for better answers, stood up for his neighborhood, and often saved me from falling into moral and intellectual sloth.

When I left Newark city employment in 1986, Columbia University welcomed me, and the urban economics tradition nurtured there by Bill Vickrey and Lowell Harriss gave me another home. Assisting Richard Arnott in putting Bill's papers together gave me a new appreciation of the depth and nuance of this tradition. Ken Jackson, though a historian and not by training an economist, has also contributed to my Columbia learning. I once aspired to make this the sort of book that Ken and Bill would have written if they had worked together, but realized that such a goal was too much to hope for.

More proximately, Mike Aronson of Harvard University Press has worked with me on this project since it began, kept it from going off track, revived it often, orchestrated the readers, and made sure the result was a book that someone other than myself would want to read. Ada Lee and Vanita Gowda worked as able research assistants on many chapters. My colleagues at Columbia University and Barnard College overlooked my distractions from more immediate concerns. The Economics Department staff, led by Angela Reid, rescued me whenever I panicked (which was often) about producing the actual manuscript. Nancy Degnan and Ann Doherty performed a similar role in the years I was working in the Masters in Public Administration program.

I have subjected many classes at Columbia (and one at Qinghua University in Beijing) to part or all of the material in the book. They were polite enough not to complain. By their comments—and sometimes expressions of incomprehension—they made the book better. My teaching assistants, especially Naoki Sudo and Ting Wu, resolved most of the students' problems and made contributions of their own.

Ingrid Ellen read several versions of the book, used it with her classes at New York University's Wagner School, and made voluminous comments. Her optimism buoyed me at key moments. Ken Jackson also commented extensively on drafts of the text, as did several anonymous reviewers. Seth Weissman gave me much helpful pedagogical advice.

Mary Gallagher, my wife, teased me incessantly about how long the book was taking, but always let me take a little more time to rewrite another chapter. Her humor, forbearance, and support made the book possible. Grete C. Gallagher went over the manuscript many times and provided invaluable assistance.

My final debt is to the many people who helped me but whose names I have inadvertently omitted.

City Economics

Chapter 1

Introduction

One summer afternoon we crashed. My wife and I were on vacation, headed from New Jersey to Niagara Falls. But on Interstate 81, two hundred miles from home, I lost control of my car and plowed into a guardrail. There were no injuries, but the car was totaled.

When we climbed out of the automobile, we knew we were facing a lot of problems: what to do with the wreck, how to handle the insurance, how to continue our trip, how to get food, where to sleep. The kind state trooper who arrived on the scene took us to Binghamton, New York, about eight miles away (after writing me a ticket). The remains of the car went there too.

Why Binghamton? Because it was the nearest substantial community, and everything we needed—garages, towing companies, insurance adjusters, car rental companies, motels, restaurants—was there. This would have been even more true if we had been injured and needed medical care. There was nothing on the rural stretch of road where we crashed.

Why were all the services we needed located within a few miles of each other instead of being spread evenly around the countryside?

In Binghamton, we still had problems getting from one place to another (the major car rental companies were all at the airport, eight miles from downtown), and there weren't any Chinese restaurants near our motel. Why was Binghamton so dispersed, and why weren't there any Chinese restaurants on the motel strip? Would it have been better if Binghamton had been more compact, and had a greater variety of businesses?

Much of the city itself was unattractive. The garage our car was towed to sat between a decrepit rail spur and a rundown house with a vicious 150-pound dog in the yard. Two nearby garbage containers were almost overflowing, and the puddles around our car were so wide we had to throw two-by-six planks into them so we could reach the car to empty it out. By contrast, the region itself had great natural beauty, with its forested hills surrounding the Susquehanna River Valley. Maybe Binghamton shouldn't have been there at all—maybe we should have had to go to Scranton or Syracuse, which would have been big enough to support the downtown car rental agencies and Chinese restaurants we needed.

Binghamton was different from the surrounding countryside in another way, too. Fewer than 3 percent of the population in the countryside was African American or Hispanic, but four times that proportion of the people living in Binghamton were. Why did minority people cluster there? Did this segregation make Hispanic and African American people worse off? Did it make European American people better off?

Binghamton is a city because it's a place people go when they're looking to make business deals or legal contracts, to buy clothing or marijuana, to watch a live baseball game or hang out at a bar, to attend a community college or university, or to work in an office or a factory or a motel or a store.

This book is about cities throughout the world. Urban agglomerations are a great idea. Concentrating a lot of activity in a small area saves on transportation; allows all sorts of convenience, sharing, and economies of scale; encourages the spread of new ideas; and simply satisfies the human desire to be social. Cities are such a good idea that most of the world's people now live in places that can reasonably be called "urban"—as opposed to only 16 percent in 1900 (Bairoch 1988, p. 502). Life without cities would be poorer than it is now—not just for city residents but also for anyone who consumes the products and services that are developed and produced in cities, and anyone whose wages are higher because he or she could migrate to a city. These categories include just about everyone in the world today.

But city life has drawbacks. Throughout most of human history sanitation has been a primary urban problem; living in a city was hazardous to your health. Life expectancy at birth in New York City in 1880 was only about twenty-nine, nearly the same as it was in imperial Rome. But by 1910 New York life expectancy had risen to forty-five (Meeker 1974, p. 392), and by 2000, death rates in cities were generally lower than in rural areas, even in developing countries. This improvement has been achieved only through large investments in sewers and plumbing, water systems, medicine, and nutrition (paid for by the higher wages that workers earn in cities).

Another drawback is that the more activity you have in a small area, the more opportunities people have to get in each others' way. Cities throughout the world are afflicted with traffic jams, airport congestion, crowded buses, and delayed subways. People fight over where schools will be built, how late bars will stay open, when and where loud music can be played, how tall apartment buildings can be, what provisions will be made for parking, and where fast-food restaurants, gasoline stations, and drug-treatment centers will be located. They kill, rob, mug, and swindle; their houses, factories, stores, and offices catch fire and burn more often than those in rural areas.

These difficulties arise out of the same geographical propinquity that makes cities work. You can't have one without at least encountering the other—just as you can't water ski without getting wet, or exercise hard without risking injury. The much lamented anonymity of city life, for instance, is

linked to the advantages of agglomeration: part of the lure is that in cities you can transact business with strangers, an advantage you can't enjoy without being surrounded by a bunch of strangers. If you want to avoid disease, congestion, noise, crime, and racial animosity, you can be a hermit on a mountaintop or hide in the desert, but then you will have to give up all the joys that come from physical encounters, all the help and advice that other people can give you, all the skills that they have, and all the advantages of intelligent organization of production. You will have to make your own clothing, grow your own food, smelt your own metal, build your own computer and write your own software to link it with the Internet, dispose of your own waste, craft your own musical instruments, diagnose and cure your own ailments. You will, in short, live crudely and poorly.

The more that different kinds of people with different talents and backgrounds come together more frequently in a smaller space, the better a city operates as a place to work and play—and the greater the possibilities for contagion, congestion, and hatred. Of course many different policies, some wise and some not, can affect the extent of these problems, just as careful stretching can (probably) reduce an athlete's probability of injury from a workout. But you can't enjoy the benefits of city life without doing *something* about the problems.

Any amelioration of urban problems produces a double dose of benefits. Less crime means not only fewer victims but also more opportunities for people to meet neighbors, go shopping, attend community meetings, and enroll in night school. Improving racial relations means not only fewer instances of discrimination but also more potential partners for everyone in every conceivable kind of transaction. This double dose is one good reason for studying urban economics.

I. Urban Economics

The advantages of getting people to cooperate with one another are greatest in cities, so if you think about economics as the study of how people cooperate, urban economics is economics at its most intense. What makes it more intense is that urban economics is the economics of everyday life—few people see bonds or debentures even occasionally, but almost everyone lives in a house or an apartment, flushes toilets, commutes to work or drives to recreation sites or a summer house, worries about safety, and does some shopping. In urban economics, you can't avoid the subject of your inquiry.

Urban economics is marked by a strong tension between looking primarily at history and ignoring it altogether. Some aspects of city life change almost by the minute—which neighborhoods have desirable real estate and what sort of people live in them, for instance—while others seem stable over many cen-

turies—the primacy of Paris among French cities, for instance. One of the great puzzles that urban economics confronts is how much of contemporary reality can be explained solely by reference to contemporary facts, and how much can be understood best as the product of the peculiar twists and turns of history. Are cities more like newspapers, which are printed anew every morning, or like family names, which are what they are only because of events in the long-distant past?

The past operates mainly through the force of conventions. Speaking English, using daylight saving time, calling the day after Monday "Tuesday" are practices that people follow not because they are the best possible way of doing things but because everyone else follows them; that's why we call them conventions (see Lewis 1969 for a rigorous definition). I have no feeling for the Norse god Tia, after whom Tuesday is named, but I use the name when making appointments because the people I want to see use the name "Tuesday" too, even though they don't venerate Tia either. If they called the third day of the week "Elvis-day" I would too, and I would be just as happy as I am now.

Money is conventional. The green pieces of paper bearing pictures of presidents and treasury secretaries could just as well be blue pieces of paper with pictures of football players and poets. But because other people accept the green pieces, I accept the green pieces too. Similarly with credit cards: businesses accept Visa cards because a lot of people carry them, and a lot of people carry them because businesses accept them. Money makes a lot of transactions easier *because* everybody accepts it and everybody knows that everyone else accepts it—that's how money does its job.

To a large extent, cities work the same way. Binghamton first grew because it was located at the confluence of the Chenango and Susquehanna Rivers, but this fact was of no more concern to us or to the state trooper or to the people who ran the motels than the relationship between Tuesday and Tia. The motels and garages are in Binghamton because people who crash go to Binghamton, and people who crash go to Binghamton because that's where the motels and garages are.

To be sure, cities are not entirely a product of this kind of circular logic. Neither is money: it's hard to imagine a society where chocolate fudge sundaes are the medium of exchange or where the major cities are located inside active volcanoes. But especially with large cities, convention plays a major role.

Consider, for instance, New York City and the city of Niagara Falls. Niagara Falls is clearly a more attractive location. No natural phenomenon in New York City is so beautiful as the falls, and New York City has no good sources of cheap electrical power. But New York City is the world's greatest and richest metropolis, while Niagara Falls is a small depressed city.

History matters. In the late 1700s and early 1800s, water transportation was the cheapest way of moving goods, and so the confluence of the

Chenango and Susquehanna Rivers was the obvious place to build a city, Binghamton. It's still there today.

The Erie Canal, which runs between Albany and Buffalo, New York, is a good example of how important history is in the development of cities. A shallow trench about ten meters wide, the canal is insignificant in the twenty-first century. But if the Erie Canal had not been constructed when it was (between 1817 and 1825), New York City would not today dominate the United States, and Chicago might not be the major city in the Midwest. Only by understanding history can we understand these two cities today.

Before the Erie Canal, New York was just one of four large East Coast cities; Boston, Philadelphia, and Baltimore were the others. The canal set Gotham apart: only New York could connect with the West by water. Farmers from upstate New York, western Pennsylvania, southern Ontario, Ohio, Indiana, and Illinois could all ship their produce to New York more easily than to any other Atlantic port. Grain going to Europe went through New York; so did manufactured goods coming from Europe. New York was the place where goods and grain changed between ocean-going vessels and river-going vessels.

As a result, insurers became concentrated in New York (transportation was risky and you had to insure your shipments on each leg of their journey). With insurance came banks and other financial services. The volume of commerce with Europe made a difference too. By the 1850s New York had become the first and only port in America with regularly scheduled shipping service to Europe. If you shipped through Philadelphia or Boston, your package had to wait until there was enough other cargo for a trip to be worthwhile; if you shipped through New York, you got a definite timetable—and faster service, because more people were shipping through New York.

Once the Erie Canal had made New York preeminent among American cities, preeminence spurred growth on its own. European firms that were going to locate only one office in America chose New York. Writers and artists and journalists gravitated there too. New York was the obvious destination for American railroads, and because the rail connections were superb, manufacturing firms found the New York area congenial too.

The Erie Canal also contributed to the growth of Chicago. Suppose it's the 1870s and you are a farmer in Iowa looking to ship grain to Europe or the East Coast. If you ship first to St. Louis or Kansas City, you will have to contend with the only railroad that heads east from there. That railroad company knows your grain will be stuck in St. Louis without its service, and so it can get away with charging very high prices. On the other hand, if you ship through Chicago the railroads can't charge too much, because they know you could put your grain on a boat instead and ship it across the Great Lakes and down the Erie Canal. So Chicago was a better place for shipping grain, and as more and more people realized this, Chicago became even more attractive because the heavy volume of freight made shipping cheaper and faster. And so it

is that Chicago and New York—rather than, say, St. Louis and Baltimore—are the major cities in the eastern half of the United States, more than a century after the Erie Canal, to which they owe their prominence, lost its economic role.

But to say that history is important is not to argue that cities never change. They are like newspapers, too. Cities are tools for working, living, and playing, and so cities change as the ways that people work, live, and play change. Houston, for instance, was almost uninhabitable in the summer before the invention of air conditioning. Before streetcars, in the early nineteenth century, cities could be only as large as the distance that someone could walk in about an hour; in the early twentieth century, skyscrapers became feasible only when elevators had been invented and powerful mass transit systems had been developed so that enough workers could be brought in to fill a skyscraper's offices. Since World War II, frozen food and television have diminished the traditional lure of cities: before frozen food you had to go shopping every day if you wanted fresh meat and produce, so living close to stores was attractive; before television, you had to live in or visit a major city if you wanted to see world-class athletes, musicians, or actors. Now you can live fifty miles from a supermarket, shop once a month, and still eat good food every day; you can watch the World Series or the World Cup or the Super Bowl from the living room on a farm.

There's little doubt that cheap information-processing capability and the Internet will also have a major impact on how cities operate and what they look like. Exactly what that impact will be is impossible to say now, because we don't know what future technological progress will occur. Chapter 2, however, includes some speculations about the implications of the information revolution.

Because history matters and because technology matters, cities differ. A gambling resort like Atlantic City is different from an oil town like Tulsa. Nevertheless, probably the major difference for this book is between cities in the United States and cities in developing countries. Many U.S. cities, in the East and Midwest particularly, were developed before cars, frozen food, and television; people worry that these cities are growing too slowly or shrinking too fast. Many cities in developing countries, on the other hand, have been adding population rapidly over the past several decades, often with hardly any water systems, schools, houses, and transportation in place. People worry that these cities are too big or growing too fast.

II. Defining Cities

Cities are places where a lot of human activity and interaction regularly occurs in a small space. Thus, for purposes of this book, Disney World; the

Englishtown, New Jersey, flea market; and the Watkins Glen, New York, race-track count as cities, as well as London, Shanghai, and Binghamton. Disney World, Englishtown, and Watkins Glen can be considered cities because they concentrate large numbers of people in the same space at the same time.

This definition is, of course, nonstandard. Most traditional definitions concern themselves with where people sleep on a regular basis. For instance, some definitions emphasize population density—the number of people sleeping per hectare, say. But hardly anyone sleeps regularly in downtown Manhattan or in Central Park, and these are precisely the kinds of spaces that most people think of when they think of cities. Similarly, defining cities as places that have a certain governmental apparatus is also inappropriate. Policing, firefighting, and street sweeping are important human activities, just like sleeping, but they are not the only human activities.

Using the term *city* to describe a political subdivision is, of course, common. In the United States, when areas of dense human activity began to emerge the states chartered corporations to provide some common services to these areas (chiefly policing and street lighting). These corporations were called cities, and so were the areas in which they were authorized to tax and provide services. At that time, these areas were the only areas of dense human activity—except for some large plantations—and so the name was appropriate.

As more and more different types of areas of dense human activity have evolved, the use of the word *city* to describe a particular type of political subdivision has become confusing. The Jamaica Bay National Wildlife Refuge, for instance, is located within the corporate boundaries of New York City, while the Willowbrook Mall is in the township of Wayne, New Jersey. But there is more human activity going on at Willowbrook than at Jamaica Bay, and so Willowbrook would be a more apt subject for this book than Jamaica Bay.

Of course, where people sleep and how activities are governed are still part of urban economics, even though I will not be using them to define the subject. Where people sleep is discussed in Chapter 6, on location theory, in Chapter 7, on sprawl, and in Chapters 13 through 15, on housing.

We will also have to be careful in using data that the Census Bureau and other government agencies have collected. The key term here is **metropolitan area,** which is defined as a core area with a large population nucleus, together with adjacent communities that have a high degree of economic and social cohesion with the core. There are a variety of refinements of this definition (metropolitan statistical areas, consolidated metropolitan statistical areas, and so forth), and in 2003 the federal government even stopped using the simple metropolitan area concept, but the basic idea is that a metropolitan area is a fairly small area in which a large number of people perform almost all the activities of their lives (sleeping, working, eating, playing). Thus when I refer to San Antonio, for instance, I'll be talking about Bexar, Comal, Guadalupe, and

Wilson counties, not just the city of San Antonio. When I talk about comparing cities, I mean comparing metropolitan areas.

III. Keeping Score

Urban economics differs from most of the rest of economics, then, in its preoccupation with everyday life, the extreme complexity of the coordination studied, the big differences little events can make, the importance of history and convention, and the indirectness of the consequences of technological change. But urban economics shares features with the rest of (neoclassical) economics. Most important is the way it keeps score—and this way of keeping score sets urban economics apart from most other ways of studying city life.

By "keeping score" I mean judging how well various policies and programs work. In economics, what matters is how well individual people are doing, and that means how well they are achieving their goals and satisfying their desires, regardless of what those goals and desires are.

This may sound obvious, innocuous, and uncontroversial, but it isn't. The restriction to people means that we don't care about dogs, cats, buildings, or the environment—except insofar as people care about them or they affect people. The restriction to goal achievement and desire satisfaction means consumer sovereignty. My judgments about what's good for people or what they should do don't matter (except insofar as people happen to agree with my judgments). No matter how disgusting I may think Barney, cocaine, and haggis are, if they are what some people want, I applaud when they get more of them.

This proposition implies that many of the ways people have of thinking and talking about cities have no place in urban economics. Take, for instance, statements like, "Binghamton has seen better days," "Newark is finally coming back," or "Singapore is booming." To an economist, these statements are as nonsensical as, "My kitchen table is sad," "This subway car is furious," and "The San Andreas Fault is doing well these days." Only people can be better off or worse off, only people are our concern, and cities aren't people. They are means, not ends. If all the land and physical structures in New York City were obliterated tomorrow night, but somehow no one became worse off and a few people became better off (in their own estimation), the world would be a better place, at least the way economists judge it.

Similarly, a lot of what journalists and government officials refer to as "economic concerns" and "economic development" really does not concern economics at all. The aims are usually something like increasing the number and size of structures in a particular area, or making that area look nice, or increasing the income of particular landowners. Perhaps these programs help people achieve their goals and satisfy their desires better; perhaps they don't.

In Chapter 18 I will investigate this issue in depth. Constructing more buildings and making places look better isn't necessarily the same as making people better off; economics as I will be using the term means the latter activity, not the former. The ethical system of neoclassical economics really isn't very materialistic and hard-boiled; it's closer to squishy soft.

Applying the usual economic criteria to many urban concerns, however, is not straightforward. The difficulty lies in deciding what people to worry about. Suppose we are thinking about building a new road in downtown Binghamton. What groups of people should we be concerned about: current drivers in downtown Binghamton, current residents of Binghamton city, current residents of the Binghamton area, people who might drive in Binghamton but won't if the road isn't built, people who might live in Binghamton but won't if the road isn't built, people who might live in Binghamton but won't if the road is built, other people who will never live in Binghamton but whose lives will be affected by what possible Binghamton-movers do, children who will never be born if the road is built because their parents will be too busy building it but who will be born if the road is not constructed?

In this book I will usually opt for considering the widest possible group; in trade and migration studies this is usually called the cosmopolitan view. Cosmopolitanism is another reason why the discussions in this book will be different from journalistic discussions of urban economic development. But I will also try to be clear about who gains and who loses, since aggregates don't always have much meaning.

IV. Plan of the Book

This book is about how the advantages of city life can best help people satisfy their desires and achieve their goals. The next chapter begins by outlining what the advantages are, and how economists have measured them. It shows what the world would be missing if cities were poorly managed.

Chapters 3 and 4 are about cars, and Chapter 5, mass transit. Understanding transportation lets us turn in Chapter 6 to the analysis of land use and land prices. How land gets used and how much it's worth depend in large measure on what sort of access it affords to places people want to go.

Chapter 7 introduces a recurrent theme, the question of whether U.S. cities are too sparsely settled and whether developing countries would be wise to avoid America's sprawl. The chapter examines externalities within cities and looks at "flight from blight"—the idea that bad conditions in central cities cause sprawl.

Chapter 8 is about the rules and regulations that suppress antisocial actions, and Chapter 9 is about providing alternatives to socially undesirable actions; here I discuss water, sewers, fire suppression, and garbage removal.

Zoning is a major topic in Chapter 8, and in both chapters I ask whether fail-ures in the policies under examination could cause flight from blight and sprawl.

Education is the subject of Chapter 10: the difficulties of developing good policies, the huge urban advantages, and the implications of changing technology. Transportation, housing, sprawl, and increasing returns to scale all figure in analyzing education, and all the types of policies discussed in the previous chapters are relevant.

Chapter 11 shows that racial and ethnic cleavages are probably responsi-ble for serious losses of the advantages of city life, and Chapter 12 looks at policies that might alleviate these problems. Understanding race is also nec-essary for understanding policies toward low-income housing, crime, and economic development, the topics that follow. So race becomes a recurrent theme, along with sprawl.

I devote three chapters to the major device that most people use to make urban existence safer and more pleasant—housing. I begin in Chapter 13 by looking at the policies that affect the vast majority of Americans: the income tax treatment of owner-occupied housing, tenancy laws and rent regulation, and the mortgage market. These topics have implications for sprawl, too.

The other two housing chapters are about the housing concerns of poor people. Chapter 14 discusses the policies that are designed to help poor peo-ple—ostensibly, at least—and Chapter 15 discusses homelessness.

The next two chapters take up crime, a major problem that cuts into the advantages of urban life and may contribute to both sprawl and racial ani-mosity. Chapter 16 is about crimes with victims; Chapter 17, crimes without victims (alcohol, drug, and gun policies).

Chapter 18 is about urban economic development. Once again I will make extensive use of the foregoing chapters, largely because most of the good policies for urban economic development have already been discussed some-where else.

V. Why Should You Care about Urban Economics?

You can't avoid urban economics. You have to worry about going to school, finding a place to play, flushing toilets, drinking water, getting mugged, being approached by panhandlers, getting to work, living someplace, buying food and clothing, enjoying yourself, staying healthy, getting sleep, and dealing with people who are different from you. You can't acquire a spouse without a mar-riage license, a bar without a liquor license, or a dog without a dog license. When you die, your remains are going to have to be disposed of in some sani-tary fashion.

Because you must deal with these things anyway, it makes sense to be

able to think of them more systematically and in a wider context. Studying urban economics won't teach you how to fix a leaky faucet or house-train a reluctant puppy. But it will let you understand the long-run consequences of those actions, why some people might take them, whether more or fewer should, and what sorts of policies would encourage them or discourage them.

References

Bairoch, Paul. 1988. *Cities and Economic Development: From the Dawn of History to the Present.* Chicago: University of Chicago Press.

Lewis, David K. 1969. *Convention: A Philosophical Study.* Cambridge, MA: Harvard University Press.

Meeker, Edward. 1974. "The Social Rate of Return on Investment in Public Health, 1880–1910." *Journal of Economic History* 34:392–419.

Chapter 2

Why Proximity Is Good

M*ost human activity* on this planet is concentrated on a very small amount of land. Less than 5 percent of the land area of the United States is developed. The largest twenty-five cities are home to 12 percent of the population, but those cities take up only 0.2 percent of the land. Mexico City covers less than 0.1 percent of Mexico's land area but accounts for almost 10 percent of that country's population. The entire population of the United States could fit into Maryland if it were packed as tightly as the population of New York City. About 18,000 students attend Hunter College in Manhattan, all of which is housed in buildings located within a couple hundred feet of the corner of Lexington Avenue and East 68th Street; only 30,000 attend college in all of Wyoming.

These concentrations are cities. The first step toward understanding cities, then, is examining why activity is concentrated in a few places. What advantages are there in putting so much activity close together? Anyone who has ever been stuck in traffic, caught a cold, or been annoyed by some variety of jerk knows that this pattern creates many problems. Cities could persist—as they have for thousands of years—only if their advantages offset the disadvantages.

The goal of this chapter is to identify those advantages—in other words, to figure out why cities exist. In economic terms the advantages are labeled **increasing returns to scale,** to emphasize that the same type of inputs are better at producing outputs when more of them are used together, or **economies of scale,** to emphasize that it costs less to produce additional outputs when a lot is being done nearby. The two terms are just two sides of the same coin. One says that you get more output for the same input, while the other says that the same output requires less input, so it doesn't matter which term you use.

You should interpret "inputs" and "outputs" broadly. Consider, for in-

To get the most out of this chapter, you should be familiar with these concepts: externalities and comparative advantage. You can find these terms in the Glossary. Also in the Glossary are terms that appear in bold the first time they are used in the text.

stance, one of the oldest reasons why cities were built: military protection. The inputs that went into protecting an ancient city were a wall, soldiers and weapons, a moat, alligators in the moat, and so on. The output was either the area protected, the number of people protected, or (in the most sophisticated terms) the value of what was produced in the protected region.

This example also shows what is meant by increasing returns to scale and economies of scale. Suppose the city is square and each hectare inside it has the same number of people and net value of production. Then output is proportional to the *square* of the length of a side. Symbolically, let O stand for the output—the area protected or the value produced in that area—and s stand for the length of a side. Then

(1) $O = s^2.$

On the other hand, the inputs depend on the length of the perimeter (how many stones you need for the wall, or alligators for the moat), and the perimeter is proportional to the length of a side (not the square of the length of a side). In symbols, where I stands for the quantity of inputs,

(2) $I = 4s.$

So there are increasing returns to scale. With twice the inputs, you produce quadruple the output (not just twice as much). Algebraically, combining (1) and (2),

$$O = I^2/16.$$

Obviously, there are also economies of scale. Combining (1) and (2) the other way gives

$$I = 4O^{(1/2)}.$$

For a city of 16 square kilometers, you will need the inputs to protect a perimeter of 16 (linear) kilometers. For a city of 9 square kilometers, you will need to protect a perimeter, not of 9 kilometers, but of 12. In a 16-square-kilometer city, adding another square kilometer to the area protected requires defending another 0.49 kilometers of perimeter; in a 9-square-kilometer city, adding another square kilometer requires defending another 0.65 kilometers of perimeter.

Cities, then, economize on protection, and so protection against marauding barbarian armies is one reason why people have come together to live in cities (though not an especially important one these days, and cities may be a bad way of avoiding terrorists, as opposed to marauding barbarian armies). Not all activities, though, are like protection. Many have constant or decreasing returns to scale. Consider, for instance, tooth brushing. It's difficult to imagine any reason why 100,000 people together in a stadium brushing their teeth would do any better job or use less toothpaste or enjoy any more of an

uplifting tooth-brushing experience than 100,000 people brushing separately, so tooth brushing probably has constant or decreasing returns to scale (because people would get in each other's way, especially when rinsing). Tooth brushing is not a reason to build a city.

Probably most activities, in fact, are not good reasons for building cities. The standard results of mathematical economics are built on the assumption that *all* production has constant or decreasing returns to scale. Economists look at increasing-returns-to-scale activities as special.

Which activities have increasing returns to scale, how those returns operate, and how important those activities are determines in part what cities exist, how big they are, and what goes on in them. Increasing returns to scale are the driving force behind the growth of cities. We want to know what makes the engine work before thinking about where the car might go.

I start this chapter with the most standard kind of increasing returns: the kind that usually occurs entirely within a firm. Next I look at reasons why some activities will be performed better in denser environments, even when those environments are not controlled by a single firm. After that, I investigate how cities affect innovation. Finally, I speculate about the impact of new technologies. Essentially, this chapter is a listing of sources of increasing returns to scale.

There is, however, a deeper motive that should be evident from the first chapter: the good things this chapter describes can be lost with too much deconcentration, racial animosity, poor public health, or stupid policies. This chapter is about why you should care about cities.

I. Increasing Returns within Firms

The increasing returns to scale that occur within a single firm are probably the least interesting for the study of urban economics, but they are the type most familiar to students of other parts of economics. Some activities just go better when they are done big-time.

Activities like brewing and chemical manufacturing are good examples. A large part of the work in these industries consists of putting liquids in various containers and doing things to those liquids while they are in the containers. The containers are a significant cost. Suppose you want to put eight times more liquid in a container. Since containers are three-dimensional objects, you have to double the length of each dimension (if it's a sphere, you have to double the radius; if it's a cube, you have to double the length of each side). That means that the surface area of the container has to increase only four times. Surface area is proportional to the amount of material you need to make the container. So costs go up only four times, although the amount of

liquid you can handle goes up eight times; there are increasing returns to scale.

This is an example of what engineers call the **two-thirds rule:** for activities where containers are the major input and the content of containers the major output, multiplying output by x multiplies cost by only $x^{(2/3)}$. Things like grain silos and warehouses also work according to the two-thirds rule, because they, too, are essentially containers. Ships are also containers, and so it should be no surprise that transport cost per unit volume decreases by 0.31 percent for a 1 percent increase in ship size, or that the shipping cost from Tokyo to Jakarta, a medium-size port, is 70 percent more than the shipping cost to Singapore, a large port, even though both are about the same distance from Tokyo (Mori and Nishikimi 2002).

Start-up costs are another reason for increasing returns to scale. Sometimes the first object you make is very expensive and the rest are much cheaper. Books and drugs are examples: writing a book and setting it in type is difficult, but once you've done that, additional copies of the same book require only paper and printing. Learning how to make a new kind of pill or writing some new software is expensive; copies of the pill or the software are cheap to produce.

Big firms also reduce customers' search costs, and so produce the same output at less cost to their customers. A Wal-Mart that sells books as well as diapers allows customers to make one shopping trip instead of two. A megaplex theater gives moviegoers a much wider array of movies to choose among and better fall-back alternatives if the movie they want to see is sold out. A brand name also makes search cheaper for customers because they don't have to learn again and again what a particular product is like. Wherever you find it, a McDonald's is going to be pretty much what you expect it to be.

But even though many different kinds of increasing returns to scale are available to firms, firms just don't grow very big, especially in market economies. Medium-size cities are larger than the biggest firms. The largest U.S. business in 1993, General Motors, had sales revenue of about $134 billion, but the people who lived in the Boston metropolitan area in that year made about $142 billion, and six U.S. metropolitan areas had more income than Boston. In the same year, eleven U.S. corporations had more than $33 billion in sales revenue, while twenty-five metropolitan areas had more than $33 billion in personal income (*World Almanac and Book of Facts* 1995; Slater and Hall 1996).

Why can't firms take more advantage of increasing returns to scale? We don't have a good or complete answer. One reason may be technological limitations on economies of scale. For instance, bigger and bigger containers require more supports and thicker skins; eventually the two-thirds rule gets overpowered by gravity. Another may be demand limitations; even the socially

optimal number of copies of a book entitled *Quartz Deposits in Western Alabama* is probably small.

Finally, another limitation on how large firms can grow may come from what firms are. Firms are groups of workers who interact not through the market or through completely specified contracts but through discretionary directives; they are islands of command economy in the sea of a market economy. When a boss asks her assistant to get coffee for a ten o'clock meeting, they do not have to negotiate about how much more or less he will get paid if he does so, nor has anything been written in his employment contract about this specific task for this specific meeting at this specific time.

Command economies have limitations. Because there is no market, the boss doesn't really know what it's costing her to have her assistant get coffee for this meeting; perhaps it's not worth the trouble. The boss's boss knows even less—perhaps he thinks the assistant is talking to customers rather than getting coffee—and so on up the hierarchy, a little bit more ignorance accumulating at each step. Big mistakes are possible. A system that works well between me and my cat (who tells me what to do) doesn't necessarily work well with hundreds of thousands of people. The accumulating inefficiencies of a command economy are one more reason why firms don't grow very big.

II. Agglomeration Economies

The most important economies of scale for urban economics are not captured within firms. They're called agglomeration economies. An **agglomeration economy** exists when production is cheaper because more activity is going on nearby. With agglomeration economies, economies of scale are "in the air": people can take advantage of them without joining big organizations. In this section, I'll look at several different sources of agglomeration economies.

Agglomeration economies come in two varieties. Those that arise from having many firms in the same industry are called **localization economies.** Cambridge, Massachusetts, supports high-end bookstores and specialized computer stores not because it's a particularly large city but because many universities and consulting firms are located nearby; this is an example of a localization economy. Agglomeration economies that arise from having many people located together, no matter what industry they work in, are called **urbanization economies.** New York City supports a store that sells wedding dresses for pregnant brides only. Because such a store couldn't survive in a smaller place, this is an urbanization economy.

The distinction between localization and urbanization economies is more than just an academic exercise designed to generate questions for multiple-choice exams. Suppose urbanization economies are the primary reason why cities are productive. Then there may be nothing at all wrong with very

big cities like Mexico City; it would make sense to have several different kinds of industry concentrated in the same city. On the other hand, if localization is the primary reason cities are productive, then every industry should have its own city. If steel plants only help other steel plants and soda companies only help other soda companies, then concentrating both steel and soda production in the same city is silly: soda workers are getting in the way of steel workers and bidding up the price of land; everyone could be made better off if steel and soda were produced in separate cities. If localization is the main source of economies of scale, then Mexico City and other similar cities in developing countries are too big.

When we look at the various kinds of agglomeration economies, it will be helpful to differentiate between localization and urbanization.

A. Specialization

One of the good things to do in New York City in December is to go to the Rockefeller Center tuba concert. There are about 300 tuba players, each in his own outfit (female tuba players are rare), and they stand in the plaza and play Christmas carols, as best they can.

If you organize big tuba concerts like this, New York City is the place to be. Where else could you find 300 tuba players to come out for an afternoon in December? People who organize big tuba concerts can do their work much better in New York than any place else; people who like to attend big tuba concerts can find their needs met in New York much better than any place else; people who like to play in big tuba concerts do better in New York than any place else.

The concerts are an example of specialization. According to Adam Smith, "The division of labour is limited by the extent of the market." Becoming a specialist entails paying a start-up cost (that's where increasing returns to scale come in): you have to set up a shop or learn a skill or establish an organization. If demand for a particular output is rare or only occasional, such investments are justified only with a lot of people around. Thus only the largest cities can support tuba concerts, ultramarathon running organizations, and marine chart stores; slightly smaller cities can support opera companies, quality book stores and professional wrestling; practically any place can support a 7-Eleven.

Specialization makes production cheaper and smoother. Consider, for instance, William Cronon's (1991) story of how the Chicago stockyards were able to lower the cost of beef by making sure they used every last little piece of cow. Cronon (p. 247) first quotes an account from 1893 by Julian Ralph:

> Everything—without particularizing too closely—every single thing that
> appertains to a slaughtered beef is sold and put to use. The horns become
> the horn of commerce; the straight lengths of the leg bone go to cutlery

> makers and others; the entrails become sausage-casings; their contents make fertilizing material; the livers, hearts, tongues, and tails, and the stomachs, that become tripe, all are sold over the butchers' counters of the nation; the knuckle-bones are ground up into bone-meal for various uses; the blood is dried and sold as a powder for commercial purposes; the bladders are dried and sold to druggists, tobacconists, and others; the fat goes into oleomargarine, and from the hoofs and feet and other parts come glue and oil and fertilizing ingredients.

And then he comments himself (p. 248):

> The portion of any single animal that might go into one of these by-products was very small. More than half of a steer's bodyweight became dressed beef, but less than 1 percent became glue or dried blood or neat's-foot oil. No ordinary butcher could afford the capital investment needed to deal in such small quantities, and so waste was inevitable when traditional methods were used.

Specialization is not something that happens exclusively or primarily within firms. That's why we call it an agglomeration economy. Marine chart stores can thrive in New York City and the users of marine charts can prosper not because they are part of a big company, but because New York is big.

Both localization and urbanization economies can arise from specialization. Cronon's stockyards are examples of localization economies: they wouldn't have worked so well if they were surrounded by boutiques, bookstores, and furniture makers. On the other hand, for setting up an organization of ultramarathon runners, for instance, the size of the city is much more important than the presence of other similar organizations. Thus specialization can also produce urbanization economies.

B. Demand Smoothing

Idle productive capacity is wasteful, but it's often unavoidable when the demands on that capacity at particular times can't be predicted perfectly. It would be nice to rent an umbrella only for the times when it's raining, or a fire truck only when a building is burning, or a can opener only for those few minutes a week when you want to open a can, but the world doesn't work like that. To have an umbrella, a fire truck, or a can opener when you need it, you must keep it during all the times when you don't need it, too.

This is a waste. If you could figure out some way to share your can opener with your neighbor, both of you would be better off. Together you would need just one can opener, not two, and since it's unlikely that both of you would ever want to open a can at the exact same instant, the only cost to this arrangement would be bringing the can opener back and forth. Here's where proximity matters: sharing a can opener with your neighbor is a lot eas-

ier when that neighbor is four feet across the hall than when that neighbor is eight miles down the road. This improved ability to share is another of the economies of scale that cities are set up to realize.

Not that New York City exists so that its residents can share can openers—they cost too little for sharing to be worthwhile (except within households). But New Yorkers do share umbrellas (willingly or unwillingly), fire trucks, road and sewer capacity, and a backyard (it's called Central Park). They also share automobiles (the mechanism here is called a taxi), and kitchen and dining room space (restaurants). But a lot of the sharing is accomplished more subtly. Queuing turns out to work better when it's centralized.

To understand how, think about teller lines at banks or lines at ticket windows. When there are several tellers or several ticket windows, two different kinds of lines are possible. Every window could have its own line (just as in a supermarket every checkout counter has its own line, or at a highway toll plaza, every toll booth has its own line); customers have to decide which window to head for when they first arrive on line. Call this the toll-booth system. Or there could be only one line, and whoever is at the front of that line goes to whichever window is first available. Call this the common-line system.

It's easy to see that the common-line system works better than the toll-booth system in the following sense. In any period of time, the same number of tellers will complete at least as many customer transactions under the common-line system as under the toll-booth system, and probably more. Why? Because tellers are idle less under the common-line system. Under the common-line system, a teller can be idle only when no one at all is on line, but under the toll-booth system, a teller can be idle when no one is on *her* line but there are still queues at the other windows.[1]

This implies increasing returns to scale in retail banking: the same number of tellers (or ATMs) can serve more customers per hour if they are arrayed in one bank than if they are arrayed in two. Similarly, two piano tuners in a city of 100,000 will tune more pianos in a year than one each in two cities of 50,000; and two open-heart surgeons in a city of a million will perform more bypass operations in a year than one each in two cities of half a million. The advantage of two heart surgeons in a big city is just like the advantage of a common-line system: neither will be idle as long as any customer is on line, but if they were in separate distant cities, one could be idle while patients were queuing for the other. This source of increasing returns to scale is called

1. This assumes that in the toll-booth system a customer can't change lines once she finds out she chose the wrong line; a toll-booth system with line hopping is clearly equivalent in this regard to a common-line system. In many toll-booth systems it is in fact very difficult to change lines, because of danger (switching lanes in heavy traffic), visibility (exit 16 on the New Jersey Turnpike is so big you can't see what's happening in the other lanes), special restrictions on transactions (exact-change-only or transponder toll lanes, express supermarket checkout lines, business-account-only tellers), or mechanical restrictions (customers waiting on hold in a telephone system).

demand smoothing. Cronon's (1991) description of the Great Lakes lumber trade between 1850 and 1880 illustrates some of the effects.

The Chicago lumberyards bought from small sawmills in northern Wisconsin and Michigan, and shipped to farms on the prairies. What was Chicago's appeal? For the sawmills, which often needed cash, the big appeal was that you could always sell your shipload and Chicago dealers always paid cash. Milwaukee was closer, but you couldn't always sell there. Wrote the *Northwest Lumberman* in 1883: "Sometimes after fruitless lingering on the meager market, a lumber vessel is forced to pull out for Chicago, the caresses of sharp-clawed friends being preferable to the supreme cold shoulder."

In the queuing example we have been using, you can think of sawmills as customers, arriving randomly, and cash and storage capacity as tellers. It was cheaper for the big Chicago dealers to keep these on hand, since it was less likely that they would have to stand idle, than it was for the small Milwaukee dealers, who stood a chance of seeing few shipments or none at all. The Chicago dealers, in other words, provided a pool of cash and storage capacity that all the sawmills could share in (for a fee, of course).

These advantages extended to the demand for lumber as well. Here the teller analogy works more directly. Because so many came to Chicago to buy, Chicago dealers did not have to worry as much about excess capacity and unsold lumber; thus it was cheaper for them to expand. There was another effect on the demand side as well: railroad rates for shipping lumber from Chicago were low because the lumber could be back hauled in cars that had carried grain into Chicago. Grain shippers and lumber shippers were in effect sharing the cost of railroad car capacity.

Finally, Chicago was also where the sawmills came to hire workers. They had sharp seasonal peaks in demand for labor, and so over the course of a year they shared workers with other employers. This arrangement helped employers, and it helped workers, too. They didn't have to worry so much about where the next job was going to come from, and they weren't so dependent on the fortunes of a particular industry or factory.

Demand smoothing works in many other ways as well. It allows stores to hold smaller relative inventories of everything from can openers to gourmet dog food to automobile parts. It works in the market for highly specialized labor—not only for piano tuners and open-heart surgeons but for washing machine repairers, actors and actresses, and software writers too. And it's what makes urban real property better collateral than rural real property (Helsley and Strange 1991): if the originally intended use of a building doesn't work out, finding an alternative use is much easier in a city than in the countryside.

Is demand smoothing an urbanization economy or a localization economy? Probably more often an urbanization economy—as in the bank, back-hauling, real-property, and sawmill-worker examples—but not always.

C. Intermediate Input Economies of Scale

Intermediate inputs are produced goods that are used in the production of other goods—things like screwdrivers, electricity, trucks, copying machines, legal reference books, accounting software, and railroad cars. Economies of scale are present in the production of many intermediate inputs. These economies arise from all of the different reasons we have already discussed—from the two-thirds rule (warehouses, grain elevators, and transportation), specialization (repair services), and demand smoothing (copying services, libraries, and movies). In that sense they are nothing we haven't discussed already.

They deserve special notice, though, because economies of scale in intermediate inputs can attract firms further "downstream" in production, even if downstream production has no economies of scale whatsoever. Thus firms that need fast delivery of their products cluster around interstate highway exits and airports even when their own production has constant or decreasing returns to scale; entertainment firms locate in Los Angeles, Mumbai, and New York because that's where their workforce is. Everything else being equal, if transportation is costly, it takes economies of scale at only one stage of production to draw all stages of production together.

The idea of intermediate inputs can be interpreted broadly. Labor is needed in almost all types of production, and workers have to live. So the necessities of workers' lives and things that make their lives more pleasant are intermediate inputs to production. Starbucks provides an intermediate input to many offices. Economies of scale in the medical care that workers receive, the sporting events they attend, the schools they or their children attend, the water they drink, the churches where they worship, or the bars where they seek partners are all reasons for the firms that employ them to locate together.

Amenities like these represent urbanization economies, but other intermediate inputs—grain elevators and movie equipment, for instance—are useful only within one industry. These are cases of localization economies.

D. Externalities

Sometimes the output of one firm affects the costs of another firm, even though no market transactions take place between the two firms. The traditional example is honeybees and flowers: the more honey the firm produces, and so the more bees it has flying around, the lower the costs of the flower-growing firm next door. A better example from the urban point of view is artists. Although there are many reasons that artists cluster together (stores specializing in art supplies, demand smoothing for gallery space, start-up costs because subsequent viewings of a work of art are much cheaper than the first one), one reason is that artists tend to appreciate each others' work. If you are

an artist, being in a place with a lot of other artists gives you a bigger audience, and a more discerning appreciation, for your work, as well as a wider range of other work for you to appreciate and enjoy. This is a localization economy: the more artists in a city, the more productive each of them is.

Perhaps the most important externality like this is knowledge. Being near other firms making the same kind of product you are making lets you see how they do it and then imitate what they do better than you (not every technique is patentable, and not all patents are ironclad). Firms located close together, then, will produce more in the short run because they use better techniques, and grow faster in the long run, because knowledge disseminates among them more quickly. This is one story told about Silicon Valley in California.

Knowledge also disseminates through learning. You learn techniques and ideas as you work—how to recognize a lying customer or a hole gnawed by a squirrel, where to find a good Thai restaurant or cheap source of hubcaps, when to file a bankruptcy petition or to stop repairing a broken carburetor— and no one can take what you've learned away from you. Workers in cities switch jobs more often than other workers, and they bring their knowledge with them to their new firms and spread it to their new co-workers.

E. Lower Search Costs

Finding what you're looking for is often easier in cities than in rural areas because it's easier to make observations. On Canal Street in lower Manhattan, sixty jewelry stores are located within a three-block radius, and so looking at rings is pretty easy; you can look in twenty stores in an afternoon without trying too hard. Visiting twenty jewelry stores in North Dakota, however, would be a prodigious undertaking.

This fact has several implications. The first is obvious: businesses where most of the activity involves physical searching for partners have increasing returns to scale and should be located in cities. Examples are stock and commodities markets, especially before the Internet, and entertainment management. The more people there are in a city trying to make deals of a particular kind, the easier it is to find someone to make a deal with you, and so the more you want to be in that city. This applies to personal relations as well as to business: cities are good places to look for companionship, sex, marriage partners, or any combination thereof. Lower search costs for commodity market traders are a localization economy; for life partners, an urbanization economy.

The second implication of lower search costs is better matches. You're more likely to settle for a passable ring at a price that's slightly too high in North Dakota than you are on Canal Street. Rejecting such a ring in North Dakota means you have to drive many miles to look at another one; rejecting it on Canal Street means you have to only walk next door or across the street. Better matches make production work better.

The final implication is that cities are a low-cost way of avoiding a problem called **ex post opportunism.** Ex post opportunism is a problem in contracts. It occurs when one party to a transaction has to make an investment that is worthless without the other party—build a steel mill next to an iron mine, for instance, learn the specific and idiosyncratic practices of a particular employer, or construct a telephone line to a specific house. This is called a **specific investment.** Once you've made a specific investment, you're stuck. The other party controls completely how much return you get on that investment and of course wants you to get as little return as possible. Ex post opportunism refers to the temptation that the other party has to take advantage of the party making a specific investment after that investment has been made.

(While cities provide a low-cost way of avoiding ex post opportunism problems, they aren't the only way. Contracts and firms are the most popular alternatives.)

Cities offer a good way to handle ex post opportunism because investments are subject to opportunism only if they're specific; cities and their low search costs make many investments less specific. A Lake Michigan lumber ship heading toward a port that has only one lumber dealer is making a specific investment and had better be prepared for the consequences when it gets there; a ship heading toward Chicago is not making a specific investment at all. If you don't like the first dealer you encounter in Chicago, it's easy to see another. It's similar for investments in skills: if there's only one bar in town and you are learning how to do its accounting, you're making a specific investment; if there are many, you're not. Low search costs mean that there are many alternatives to the person you are currently dealing with, and so you are less subject to their trying to take advantage of you; thus you can make more investments without worrying about contracts or long-term relationships. In cities, people are free to invest.

In this sense, cities are a substitute for firms. Companies in rural places, for instance, are likely to have their janitors and accountants on the payroll; in cities, contracting these services out is more prevalent.

F. Empirical Evidence

These, then, are several nice stories about why activities should cluster together. Are they true? This is an empirical question. How important are these economies of scale? What is the relative importance of localization and urbanization?

Crude empiricism answers these questions in part. There are lots of different cities in the world, most of which have no peculiar natural advantage, and so there must be some economies of scale operating. Wages in cities are higher than wages in rural areas, and wages in big cities are higher than wages in smaller cities. That means that recreation isn't the sole reason people come

to cities. In the United States in 1994, per capita personal income in metropolitan areas was 35 percent above per capita personal income in nonmetropolitan areas. The New York metropolitan area has more households with incomes over $100,000 than all of Canada put together (Glied 1997, pp. 147–148). Holding many other personal characteristics constant, workers in large metropolitan areas in the United States earn about 10 percent more than workers in small metropolitan areas, and 34 percent more than workers outside metropolitan areas (Glaeser and Mare 1994). In developing countries the differences are even bigger: Bairoch (1988, p. 445) estimates that in the 1970s the wages of unskilled industrial workers in cities were on average 80 to 100 percent higher than the wages of unskilled farm laborers.

(Yes, the cost of living is higher in cities, and so people who live in cities may not be better off than people who live in the countryside. But the point is not that people are happier in cities; the point is that they're more productive. Firms pay these wages; if people were not more productive in cities, firms would move to the countryside and pay lower wages.)

Crude empiricism also indicates that economies of urbanization aren't the whole story. If they were, every city would be the same size, or converging toward it. But since we see cities of many different sizes and can detect no trend toward homogeneity, localization must be playing a role. There is other evidence for localization, too: different industries are concentrated in different cities. Entertainment is concentrated in Los Angeles, advertising in New York, car manufacturing in Detroit, the computer industry in Silicon Valley. Localization economies cause these industry-specific concentrations; urbanization economies do not.

Many more detailed studies have also been undertaken. Henderson (1988), for instance, compares inputs and outputs for a variety of industries in different cities in the United States and Brazil. For many industries, particularly heavy industries, he finds strong localization economies. In Brazil, for instance, he estimates that a 10-percent increase in a city's steelmaking employment would decrease unit costs by about 1 percent. He finds smaller localization economies in light industries, and not much evidence at all of urbanization economies.

Wage evidence also supports the importance of localization economies in manufacturing. If you compare American workers in a specific industry and occupation, those in urban labor markets that have a larger national share of that particular type of employment earn considerably more (Wheaton and Lewis 2002). It's not the size of the city that raises manufacturing workers' wages, but the number of workers like them.

At least part of the urban wage premium comes from learning: the longer they stay in cities, the better at their jobs workers become. Glaeser and Mare (1994) find that the urban wage premium is much higher for older workers, who have probably had more time to learn in the city. Workers who come to

the city don't realize the urban wage premium immediately, and those who leave it don't lose it.

III. Innovation

Proximity also spurs innovation; it doesn't just make doing routine tasks easier. Cities are hotbeds of innovation. A disproportionate share of patents originate in large cities. About 45 percent of U.S. new product innovations occur in just four metropolitan areas—New York, San Francisco, Boston, and Los Angeles—and only about 4 percent occur outside metropolitan areas (Feldman and Audretsch 1999). A quarter of the winners of the Kennedy Center national lifetime achievement awards for contributions to American culture were born in New York City (Glaeser 1998). Industries that emphasize research and development tend to be more spatially concentrated than those that don't, and innovative activity is substantially more concentrated than production activity (Audretsch and Feldman 1996).

A. Why?

Several different mechanisms contribute to the advantages that cities have as locations for innovation.

We've already seen that people learn in cities; you would expect that the interaction that produces learning would also produce new ideas. Thus, for instance, when Irish and African cultures came together in the horrific New York slum called Five Points in the early nineteenth century, tap dancing was invented—at about the same time that the interplay of Dutch and English cultures in more respectable neighborhoods uptown gave birth to the Santa Claus we know today (and most of his reindeer).

Another reason for innovation to be concentrated in cities is that new knowledge is hard to communicate over long distances, even today. Fresh, strange, not-yet-fully-developed ideas seem to travel a lot better through face-to-face discussion (where feedback is immediate and richly expressed) than by telephone, e-mail, or publication. Thus when Jaffe, Trachtenberg, and Henderson (1993) examined patent applications in 1975 and 1980 to identify trends in what other patents were cited, they found that a patent from the same metropolitan area was five to ten times more likely to be cited in a new application than a patent from outside the metropolitan area, with many other influences controlled for statistically. The home-area advantage deteriorated as patents got older, which is what you would expect if new information travels slowly. The relationship was stronger in 1980, when information technology was better, than it was in 1975.

Two other explanations for urban technological prowess emphasize the process by which ideas become inventions, rather than the ideas themselves.

They argue that cities offer more fertile ground for new ideas—that seeds take root and bloom more easily there, not that more seeds fall on the ground. The explanations are closely related.

For Helsley and Strange (2001), the suppliers of intermediate inputs are responsible for this fertility. If you're putting together something novel, you don't know what you're going to need. So the more diverse and specialized the hardware stores, tool-and-die shops, copy shops, and software engineers you can find in your immediate vicinity, the faster you can make your invention and bring it to market.

For Jane Jacobs (1969), the key to fertility is the freedom that cities provide for start-up businesses, and the rich network of intermediate input suppliers is only part of this freedom.

The basic Jacobs story is Mrs. Ida Rosenthal and the invention of the brassiere. In the early 1920s, American women wore all sorts of different undergarments. Mrs. Rosenthal was a custom dressmaker in New York. She didn't like the way her dresses hung on her customers. To improve the fit, she started experimenting with underclothing, and the result was the brassiere. Customers liked her brassieres, and she started giving out custom-made brassieres with her dresses. After working like this for a while, she found a partner, raised some money, started the Maidenform company, and went into the brassiere business full-time. She left the dressmaking business.

In somewhat more abstract terms, the world Jacobs describes is a dynamic place where growth begets growth. Start with a growing city. Because growth overcomes fixed cost barriers, people start entering new specializations. As they specialize in activities no one had ever specialized in before, they make new discoveries, some of which turn out to be useful. The people who make these discoveries go off into business for themselves, and so growth continues.

It's important that people who make discoveries be able to go off into business for themselves. For Jacobs, cities are the correct institutional setting for implementing new discoveries; firms are not.

Firms aren't the correct institutional setting because they bind their workers too closely to their tasks. For people to start doing new and innovative things, they have to stop doing the things they were doing before. The people who were counting on them to do the old things will be disappointed if they stop. Firms are set up to avoid this kind of disappointment. They establish rules and incentives so that people can count on other people to continue what they're doing. If a firm is operating in the countryside, where search costs are high, the costs of someone leaving to do something new are high—finding a new accountant in Montana when your old accountant has left to join Maidenform is not easy—and so firms make it difficult for people to leave (or to do new things within the firm). The fetters may be comfortable (pen-

sions, for instance), but they are fetters nonetheless, and they slow the rate at which technological change occurs.

In cities, by contrast, it's much easier for people to stop doing the things they were doing before. The rate of job turnover is higher in metropolitan areas than outside them. Mrs. Rosenthal's dressmaking customers may have been a little unhappy to see her move on, but they didn't have to walk around in rags. New York was and is full of dressmakers (unlike Montana), and so replacing Mrs. Rosenthal was no big problem for her customers. Because they knew that replacing her would be easy, her customers, when they first started doing business with her, did not insist on guarantees that she would not go into another business. So when the time came to innovate, nothing stood in Mrs. Rosenthal's way. The same was true for the people who went to work for her. They could stop doing what they were doing without significant problems, because they were easy to replace. (Innovation, it should be noted, comes at a price. That price is rootlessness and anonymity. Mrs. Rosenthal can leave her dressmaking customers because she means little to them and they mean little to her. You can't water-ski without getting wet.)

Jacobs's story about Mrs. Rosenthal in the 1920s strongly resembles the narratives of the computer industry in California's Silicon Valley in the 1990s. Innovation in that case took the form of start-up businesses, which could thrive because people were not locked into big corporations, and intermediate suppliers were readily available. For AnnaLee Saxenian (1994), this was the major difference between Silicon Valley and the less dynamic computer industry along Route 128 in Massachusetts. She quotes a Silicon Valley entrepreneur who sounds a lot like Jane Jacobs (p. 116): "It would be very difficult for me to do what I'm doing here with DEC on the East Coast . . . Back East, I would have to rely on DEC's internal suppliers and functional groups for everything . . . The same job of bringing a new workstation to market takes two times as long in the East Coast and many more people than it does here. Here I can rely on the other companies in Silicon Valley."

B. Empirical Evidence

Except for the explanation emphasizing knowledge externalities, all of these explanations for why cities are seats of innovation portray the spur to innovation as an urbanization economy, not a localization economy: what matters are the size and density of the city, not the number of people engaged in some narrowly defined activity. Studies that have looked at the growth of cities and industries have generally been supportive of this picture.

For instance, one group of scholars (Glaeser et al. 1992), looked at how fast U.S. cities grew in the postwar period. More diverse cities, they found, expanded faster than more specialized cities. Henderson, Kuncoro, and Turner

(1995) looked at a shorter period and concluded that while diversity in a city didn't promote growth or better performance in mature industries, it did help in attracting high-tech industries and making them prosper.

Urbanization economies, therefore, seem more important in a dynamic context; localization economies, in a static context.

IV. The Future

Cities are thus clearly a good idea today. But will they continue to be a good idea? The cities we have now were created by the technology we have now (and had in the recent past); they are not unchanging products of the fundamental nature of the universe. As technology changes, they will change. Since we don't know how technology will change, we don't know what cities will look like 20 or 100 years from now. But it's a safe bet that they'll be different.

A popular school of thought, though, extrapolates from current trends in technology and concludes that the cities are going to disappear. Transmitting information electronically is getting cheaper because of inventions like fax machines, e-mail, and the World Wide Web; telecommuting, teleconferencing, and home shopping are becoming easier and more effective; and in a while, people will be able to do anything they want in the comfort of their own home. The advantages of city life we have been talking about in this chapter will disappear. When people can bring the New York experience into their bedroom in Tierra del Fuego, there will no longer be any need for New York. This view is found in the writings of such futurists and commentators on the information revolution as Toffler (1980), Naisbitt (1988), Negroponte (1995), and Knoke (1996).

Little empirical evidence supports this idea, however. Since the technological change the futurists envision is just an extrapolation of current trends, we should be seeing city advantages withering away already, especially in fields where technology is most advanced. Instead, advanced technologies, or at least rapidly changing technologies, seem to have significant localization economies. Silicon Valley, the center of the information revolution, is (in my terms) a city—one that would not exist if proximity did not count for a lot in the world of advanced technology. Software engineers are willing to pay a million dollars for a modest house in Mountain View, California, even though they could buy a much cheaper house in Tierra del Fuego and telecommute. This shows that physical presence in Silicon Valley contributes something to productivity. The information revolution is making cities look different and perform different functions, but it's not making them disappear.

Why won't cheap information transmission make cities obsolete? Because some transactions will probably continue to be best done with physical bodies present, and improved information technology makes those transac-

tions more important. To take a trivial example: marathon running has to be done physically, and the major marathons that have tens of thousands of participants are quintessentially urban events. But holding races of that size was completely out of the question until information technology made computerized scoring and course monitoring possible.

What kinds of transactions are likely to remain out of the information revolution? It's difficult to see how telecommunications devices will ever have a **comparative advantage** in eating, sex, singing hymns, changing diapers, teaching someone how to play baseball, or even transmitting detailed information about how to install and operate the most current software. Evolution has made humans corporeal beings who take delight in corporeal things, from intimate dinners to family barbecues to concerts in Central Park. For building a porch, performing open heart surgery, cutting someone's hair, installing a fiber optic network, or even conducting a tough negotiation on a complicated deal, being there on the scene is a big advantage and will probably remain so. This is not to say that computer systems can't be invented to do some of these things or to allow them to be done by someone far away or even to do them better than humans can do them. But remember that what counts in economics is comparative advantage, not absolute advantage. So long as a computer or an e-mail system has to sacrifice more calculations of satellite orbits than a human does in order to build a porch, computers will calculate satellite orbits and humans will build porches.

In what sense will the transactions that involve human contact become more important? Think about paper. When computers were first becoming popular, futurists thought they would replace paper, since computer records would be a good substitute for paper records. They take up less space, they're less susceptible to transcription errors, they don't give you paper cuts. As it turns out, though, paper is not disappearing. Total paper production in the United States rose by about 40 percent between 1980 and 1995, just about the same rate of increase as real gross domestic product.

Computers have stimulated as much paper demand as they have replaced. Before I had a computer, I had no reason to keep a ream of paper at home, because there was no way I could write fast enough to use up that much paper in less than a few years. With a personal computer, I can print a hundred pages a day and not even feel I'm wasting paper. I can print out a forty-page paper, decide I would rather have page numbers at the top than at the bottom, and print it out again. Computer records have proved to be a good substitute for some kinds of paper records, but by making documents cheaper to produce, they have encouraged other paper use.

Gaspar and Glaeser (1998) make a similar argument about telecommunications and the sort of face-to-face interactions that make cities valuable institutions. They point out that new technologies can be either a substitute for old technologies—replacing many kinds of paper records, for instance, or render-

---

Adultery

≡ It's interesting to contemplate how difficult it was for people in cities to arrange to meet before the first telecommunications revolution, the adoption of the telephone. If somebody lived two blocks away and you wanted to meet them for dinner, what did you do? One answer is that you sent a letter, but this was obviously expensive and problematic.

Theodore Dreiser's novel *Sister Carrie,* published in 1900, shows readers how hard it was for a couple to carry on an adulterous affair in late nineteenth-century Chicago (I am grateful to Ken Jackson for suggesting this book). Sometimes the characters send letters to each other when they want to meet, but this is often dangerous, and so mostly they rely on a standing appointment at a particular park bench at a particular time. Whenever one of them fails to show up, the other is totally perplexed and can go weeks at a time without understanding what has happened or where the other partner might be.

This is a good example of how the advent of the telephone increased the relative advantage of city life over rural life. Not only did people live closer together in cities, but now they could also contact each other quickly, directly, and discreetly.

---

ing slide rules obsolete—or a complement to them—inducing home computer users to buy paper by the ream.

While telecommunications will surely replace some face-to-face interactions, they'll also make it easier to contact and stay in communication with people (see Box 2A). People will have more communication partners and will want to have face-to-face interactions with some fraction of them. Cell phones, for instance, make it much easier to meet people at concerts and festivals, or just to hang out together. With better telecommunications, salespeople will have more contacts, lonely people will have more prospects for companionship, scholars will have more potential collaborators, businesspeople will have more people to try to make deals with.

Some fraction of these communications will lead to face-to-face interactions. If this fraction doesn't go down too much, the total amount of face-to-face interaction will increase, and cities—the natural venue for such interactions—will become more important. If I want to get a porch built and don't want to spend more than $10,000, the Internet makes it more likely that I will have the kind of face-to-face interaction that building a porch entails, since I am more likely to find a contractor willing to do the job cheaply. Comple-

mentarity works in the other temporal direction, too: making it easier for people in New York to communicate with people in London makes me more eager to go to a convention where I'll meet people from London, some of whom I'll want to stay in contact with.

The substitution and complementarity effects work in opposite directions, and so no one can say a priori whether the telecommunications revolution will increase or decrease the volume of face-to-face interactions. (In the early twentieth century, railroad executives thought trucks would be complements to rail cars—trucks would make it easier to bring freight to rail terminals—and so lobbied for more and better roads; they were wrong.) Gaspar and Glaeser (1998) point out that business travel appears to have been increasing faster than real gross domestic product during the past two decades, a period during which telecommunications were improving rapidly. The only reason for business travel is to engage in face-to-face interaction, and so there seems to be no evidence yet that telecommunications are mainly a substitute for cities.

Similarly, Ellen and Hempstead (2002) find that telecommuting has done nothing to lessen the allure of cities. Telecommuters, they find, are *more* likely than other workers to live in large metropolitan areas, even holding constant a large number of personal characteristics. Perhaps the story is that when you work at home, easy access to intermediate input suppliers like computer repair people, Kinko's, FedEx, Staples, and Starbucks grows in importance.

Of course, the information revolution is going to affect what cities look like and what purposes they serve. Some cities will grow and others will fade away—already San José, California, is almost twice as populous as Buffalo, New York. But proximity will still have advantages, and so there will still be cities.

V. Conclusion

Cities exist because there are advantages to concentrating a lot of activity on a small amount of land. These advantages are called either increasing returns to scale or economies of scale. Sometimes firms are organized to take advantage of scale economies, but often they are not. Cities are a substitute for firms because they reduce transaction costs and reduce the opportunities for ex post opportunism. They encourage specialization, smooth demand, lower search costs, and make innovation easier. Different cities have different advantages, and when technology changes, the reasons for cities change too. Cities are going to change because of improving telecommunications capabilities, but there is no evidence yet that they are likely to disappear, or even become less prominent in economic life.

Questions

1. Classify the following phenomena as contributing either to localization economies or urbanization economies:
   a. Lower freight rates for farmers shipping grain from Iowa to Chicago because lumber for fences and houses could be back hauled.
   b. Lots of good computer engineers in Silicon Valley.
   c. Frequent flights from the Newark airport to Los Angeles.
   d. A Kinko's located near Columbia University.

2. If you lived in nineteenth-century Chicago, before the telephone, and your stove broke, how would you arrange to have it repaired? If you repaired stoves, how would you find customers? How would you make sure your customers would be at home when you went to repair their stoves? Explain how the telephone affected economies of scale in cities.

3. Explain two ways in which the automobile and the truck have promoted increasing returns to scale in cities.

4. Consider water pipes. Suppose all water pipes of every size are made from a very strong material of effectively zero thickness (this is to make the math much easier). The output of a water pipe is the number of gallons of water that can flow through it in a month. This is proportional to the area of a cross section. Water pipes are cylinders. The cost of constructing a water pipe is proportional to the amount of material used.

   Suppose we're thinking about building a ten-mile-long water pipe from a specific reservoir to a specific town.
   a. Are there increasing returns to scale in the construction of this water pipe?
   b. Does this process obey the two-thirds rule?
   c. What is the mathematical relationship between cost and output?

References

Audretsch, B., and M. Feldman. 1996. "R&D Spillovers and the Geography of Innovation and Production." *American Economic Review* 86:630–640.

Bairoch, Paul. 1988. *Cities and Economic Development: From the Dawn of History to the Present.* Chicago: University of Chicago Press.

Cronon, William. 1991. *Nature's Metropolis: Chicago and the Great West.* New York: Norton.

Ellen, Ingrid Gould, and Katherine Hempstead. 2002. "Telecommuting and the Demand for Urban Living." *Urban Studies* 39 (4):749–766.

Feldman, M., and B. Audretsch. 1999. "Innovation in Cities: Service-Based Diversity, Specialization, and Localized Competition." *European Economic Review* 43(2):409–429.

Gaspar, J., and Edward Glaeser. 1998. "Information Technology and the Future of Cities." *Journal of Urban Economics* 43(1):136–156.

Glaeser, Edward. 1998. "Are Cities Dying?" *Journal of Economic Perspectives* 12(2):139–160.

Glaeser, Edward, and D. Mare. 1994. "Cities and Skills," Hoover Institution Working Paper E94-11.

Glaeser, Edward, et al. 1992. "Growth in Cities," *Journal of Political Economy* 100:1126–1152.

Glied, Sherry. 1997. *Chronic Condition: Why Health Reform Fails.* Cambridge, MA: Harvard University Press.

Helsley, Robert W., and William C. Strange. 1991. "Agglomeration Economies and Urban Capital Markets." *Journal of Urban Economics* 29(1):96–112.

———. 2001. "Innovations and Input Sharing." *Journal of Urban Economics* 51:25–45.

Henderson, J. Vernon. 1988. *Urban Development: Theory, Fact, and Illusion.* New York: Oxford University Press.

Henderson, J. Vernon, A. Kuncoro, and M. Turner. 1995. "Industrial Development in Cities." *Journal of Political Economy* 103(5):117–143.

Jacobs, Jane. 1969. *The Economy of Cities.* New York: Random House.

Jaffe, A., M. Trachtenberg, and R. Henderson. 1993. "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations." *Quarterly Journal of Economics* 108:577–598.

Knoke, William. 1996. *Bold New World: The Essential Road Map to the Twenty-first Century.* New York: Kodansha International.

Mori, Tomoya, and Koji Nishikima. 2002. "Economies of Transport Density and Industrial Agglomeration." *Regional Science and Urban Economics* 32(2):167–200.

Naisbitt, R. 1995. *The Global Paradox.* New York: Avon.

Negroponte, Nicholas. 1995. *Being Digital.* New York: Knopf.

Saxenian, AnnaLee. 1994. *Regional Advantage: Culture and Competition in Silicon Valley and Route 128.* Cambridge, MA: Harvard University Press.

Slater, Courtenay M., and George E. Hall. 1996. *1996 County and City Extra: Annual Metro, City, and County Data Book.* Lanham, MD: Bernan Press.

Toffler, Alvin. 1980. *Third Wave.* New York: Morrow.

Wheaton, William C., and Mark J. Lewis. 2002. "Urban Wages and Labor Market Agglomeration." *Journal of Urban Economics* 51:542–562.

The World Almanac and Book of Facts. 1995. Mahwah, NJ: Funk and Wagnall's.

Chapter 3

Cars, Pollution, and Accidents

The last chapter showed that city life has advantages and will continue to do so. Making those advantages real, however, takes transportation. That's because particular cities are great locations for doing particular things but not for doing everything. Wall Street is a great place to buy and sell stocks but only an average place for tooth brushing. So none of the land on Wall Street is set aside for tooth brushing. If Wall Street traders are going to brush their teeth, they have to go elsewhere. That's where transportation comes in.

Transportation is needed for production, too. Wall Street traders need paper and computers, neither of which are made on Wall Street. On the other end of the production process, localization economies imply that cities specializing in particular manufactured goods will produce a lot more of them than they can consume. So output has to be transported elsewhere. Airplanes would not be made very cheaply in Seattle if all the planes made in Seattle had to be used in Seattle.

If transporting goods and people were prohibitively expensive, cities wouldn't operate and no one could take advantage of increasing returns to scale. Everyone would have to be self-sufficient. The cost of transportation is one of the factors that limit how much cities can grow and how much increasing returns to scale can be realized.

At the same time, if transportation had no cost, cities wouldn't exist. It would be just as easy to travel between any two places as between any two other places, and so proximity would have no meaning: Moscow would be just as close to the White House as the Capitol is. If proximity has no meaning, it has no value, and if proximity has no value, neither do cities. Cities depend on transportation costs being neither too high nor too low.

People use many different modes of transportation: they walk, run, row, ride bicycles and skateboards, fly airplanes and helicopters; they load packages

---

To get the most from this chapter, you should be familiar with these concepts: consumer surplus, external cost, fixed cost, marginal cost, Pareto improvement, and potential Pareto improvement. You can find these terms in the Glossary.

on ships, mules, and trucks. But most trips in urban areas in the United States are made by car, and so this chapter and the next will focus on cars. Later, Chapter 5 will cover mass transit, and in Chapters 6 and 7 we'll examine the claim that U.S development has overemphasized automobile travel and that developing countries would be well advised to encourage more use of mass transit.

This chapter begins with an analysis of driving and public policy in the unrealistic case in which the driver bears all costs. This section also provides a review of some important microeconomic concepts that we will use extensively throughout the book. The next two sections show what can go wrong with that unrealistic model, and how the problems can be either fixed or mitigated. Thus section II is about road maintenance and pollution and section III, accidents and car insurance. In Chapter 4 I will discuss congestion.

I. The Simplest Case

Think about a car trip on an empty country road. Presumably this trip will confer some benefit on the person taking it. But there are also costs. Some of the costs are obvious: gas, wear and tear on the car, the danger of crashing. These are borne by the driver, and probably she'll think about them when she decides whether or not to make the trip.

Another cost borne by the driver is not so obvious but probably more important. That's the cost of the driver's time. For most people, driving is like working—something you do to accomplish some other goal, not something that's valuable and enjoyable in and of itself (if that weren't the case, drivers would try to go slowly on highways and would take the long way instead of looking for shortcuts). Driving takes people away from doing other things that would be either pleasant or useful. Time spent driving can't be spent working or minding children or playing with cats or exercising or watching television. To drive twenty minutes you must give up twenty minutes of one of those alternative pursuits. The cost of driving twenty minutes is the benefit you would have received from engaging in the most beneficial of the other activities you could have engaged in.

Economists have traditionally identified the cost of the time a person spends driving or traveling by some other mode with the amount of money she could have made in that time. This makes time by far the largest component of the cost of an automobile trip on a country road. Think about a twenty-minute trip at fifty miles per hour; the distance covered is about seventeen miles, and so it takes between half a gallon and a gallon of gas, which cost something between $0.50 and $1.50 in 2000. In contrast, the average hourly wage in 2000 for a production worker in private industry in the United States was $13.74 (U.S. Bureau of the Census 2002, table 616), and so the aver-

age wage cost of a twenty-minute trip would be $4.58, or three to nine times as much as the gas cost. For trips at lower average speeds, the proportion of time costs is even higher.

This conclusion, however, has to be modified in two ways. First, for many people driving isn't totally valueless, a simple subtraction from the time they have available to do the things they care about. They listen to music and the radio while they drive, see the sights, think about the problems that confront them, and daydream. Driving produces pleasures that working does not, and so is not so costly.

Second, in the short run not everyone can expand or contract at will the hours she works. People may not be working as much as they want to, and so traveling time would be impinging on less valuable pursuits. In the long run, though, people do have considerable flexibility—working second jobs, for instance, finding positions with good vacation packages, changing to more demanding or less demanding occupations—and most of the more interesting decisions we are concerned with, like where to live, are long-run decisions. So the error introduced by thinking that people can work as much or as little as they want may be small. But it's still there.

Economists have conducted many studies to try to tease out of people's behavior exactly how much they value their time. Someone who pays three dollars to save fifteen minutes by paying the toll on the Henry Hudson Bridge between the Bronx and Manhattan, instead of taking Broadway for free, places a value of at least a twelve dollars an hour on her driving time; someone who makes the opposite decision values his time at less than twelve dollars an hour. The consensus of studies in the 1970s and 1980s was that Americans valued driving time at about half of their wage—that is, they were willing to give up about half of their hourly wage to avoid an hour of driving (Small 1992). Those studies were done before the advent of high-quality car stereos and cellular phones. To the extent that such devices make driving time more productive and enjoyable, they reduce the cost of time even more.

Driving has other costs that are external to the motorist. But before enumerating and thinking about those **external costs,** it's useful to consider what would constitute good public policies if driving had only private costs.

Think about a particular stretch of road. Many potential trips could be taken on that stretch. For each potential trip, there is an expected benefit—how much you need to offer the driver not to make the trip if the driver could make the trip for free—and a cost, which we have already discussed.

Notice that I am accounting for both costs and benefits in terms of money, but money as perceived by the person making the choice. Are there some benefits that are too important to be accounted for by money—being at your child's school play, for instance, or with your romantic partner? Remember we're concentrating on only one particular stretch of road. If going to the school play is important, you could take a week off from work and camp out

in the school auditorium or leave work early and walk. So the trip's benefits are staying longer at work or attending the play without brambles on or creases in your clothing. Moreover, if a trip's benefits really were infinite, it would be made no matter what public policies were adopted—short of immediate capital punishment—and so we don't have to worry about how public policies will affect those trips.

Thus, for each potential trip we can think about net benefits—the difference between benefits and costs. For some potential trips, net benefits will be positive; for other potential trips, net benefits will be negative. According to almost any ethical criterion, those public policies are best that assure that all trips with positive net benefits are made, and none others. Suppose a policy kept some trips with positive net benefits from being made—a prohibition on driving on Thursday afternoons, for instance. Getting rid of the prohibition and letting drivers make trips with positive net benefits makes those drivers better off and makes no one worse off (since we are assuming for now that all costs are private). A change that makes some people better off and no one worse off is called a **Pareto improvement** (after Vilfredo Pareto, a Swiss-Italian economist and sociologist who flourished in the late nineteenth and early twentieth centuries). Being in favor of Pareto improvements is a weak position, one that it would seem only sadists would oppose, and so economists often use this criterion in thinking about policy recommendations. Getting people to stop making trips with negative net benefits would also be a Pareto improvement, since the drivers who would have made those trips would benefit and no one loses.

How can a benevolent government induce people to make the right trips—all the trips with positive net benefits and none of the others? Easy: leave them alone. Any other policy would make somebody worse off without making anybody else better off and so would be supported only by sadists.

A small diagram can illustrate this point. Array potential trips by the net benefits they provide, starting with the trip that has the greatest net benefit. For each trip, put its net benefit on the vertical axis. Then with many potential trips, you can produce a graph like Figure 3.1. Implicit in this diagram is the idea that if only a specific number of trips is being made, they will be those trips with the greatest net benefit.

Obviously, T is the number of trips that produces the greatest net benefit. If fewer than T trips were being made—say, only U—then all of the positive net benefits for trips between U and T would be lost—the triangle S in Figure 3.1. If more than T trips were being made—V trips, say—then all of the trips between T and V would be causing more harm than they were worth. Their drivers could gain the triangle W if these trips were not made.

Drivers left on their own, of course, will take precisely T trips.

All this is obvious, and I am writing it down only so that later we can understand more complicated and realistic situations.

Figure 3.1 Finding
the optimal number
of trips.


Figure 3.1 Finding the optimal number of trips.
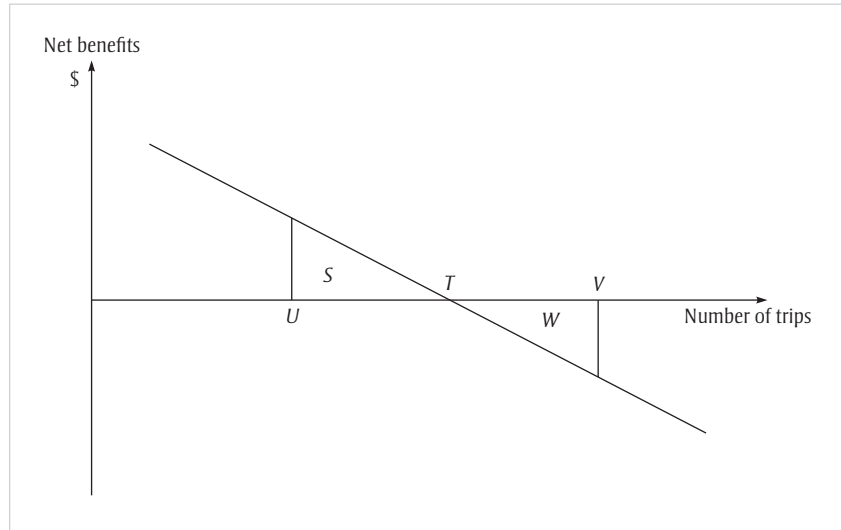
If all costs are private, tolls are always a bad idea, because they dissuade people from making some trips they should make. Suppose there's a dollar toll for every trip. Then only potential trips with net benefits greater than a dollar will be taken; trips with positive net benefits less than a dollar won't be taken even though they should be. Abolishing the toll and collecting from each person what he or she would have had to pay if there were a toll—but making that amount independent of the actual trips he or she makes—makes some people better off and no one worse off.

Figure 3.2 illustrates this conclusion. The toll reduces the net benefit that every driver gains from every potential trip, and so shifts the net benefit schedule down, as shown. Since after-toll net benefits are what matter to drivers, the number of trips made falls from T to U. We've already seen that U is the wrong number of trips; when only U trips are made, people in the aggregate lose the net benefits represented by triangle S. Triangle S is called the **deadweight loss** from the toll. Since the toll collector gets money only from the U trips that are actually made, the deadweight loss is a benefit to no one.

This sounds fine in the short run, when roads have already been built and paid for, but what about when deciding which new roads to build and how to pay for them?

First, think about which roads to build (or how wide to build them). The costs of building a road are obvious—land, concrete, labor, engineering fees. These are usually borne by the government. The benefits are the net benefits to all the individual users. Suppose that the sum of the net benefits is greater than the cost of construction. Then there is some way of financing the project

Figure 3.2 The
effects of a toll.
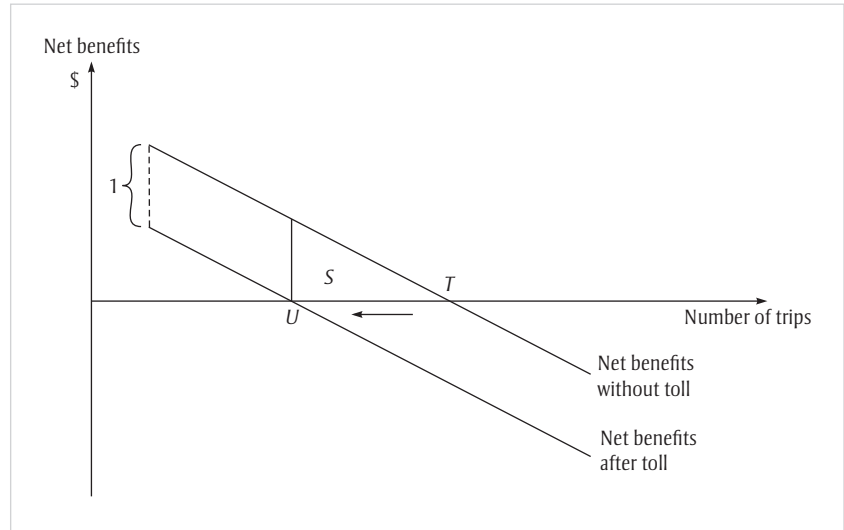
that makes some people better off than they would have been if the project
had not been done, and no one worse off. If the sum of the net benefits is less
than the cost of construction, any way of financing the project must make
someone worse off.

A project for which the sum of the net benefits is greater than the cost of
construction is called a **potential Pareto improvement.** Carrying out such a
project could be a Pareto improvement if the financing were set up right. The
ethical case for recommending projects that are potential Pareto improve-
ments is nowhere near so compelling as the case for recommending projects
that are Pareto improvements, since a project that is only a potential Pareto
improvement might make some people worse off. Of course there is nothing
automatically ethically wrong with a project that makes some people worse
off—most people think the defeat of Hitler in World War II was a good thing,
even though it made him worse off. But one needs to be more hesitant in rec-
ommending potential Pareto improvements.

Still, in this book I will usually content myself with just identifying po-
tential Pareto improvements; that's the way economics is usually done. How-
ever, it's important to note the financing system that could make them into ac-
tual Pareto improvements, and if that financing system would be impossible,
who the winners and losers would be.

So the basic question about building new roads is whether the net bene-
fits to the potential users would exceed the costs of construction. People
shouldn't have to pay for using a road, since once it's built there's no reason
(under our current assumptions) to discourage usage. Once you've bought a

cake you should eat it; you shouldn't charge yourself a fee per slice to pay yourself back and then end up throwing away half the cake because it wasn't good enough to warrant the price you were charging yourself. In the absence of external costs to driving, tolls are stupid.

II. Road Maintenance and Pollution

Of course the driver doesn't bear all the costs of driving. Driving pollutes the air and water, tears up roads, and causes accidents in which other people get hurt and their property damaged. Driving also causes traffic congestion.

When people don't bear all the costs of an activity, they usually do too much of it. Thus, restricting the activity is a potential Pareto improvement: if the driver were to drive less, and the people harmed by his driving paid him something, everyone would be better off. People wouldn't drive as much as they do if they had to compensate other people for the harm their driving costs.

To understand why this is so, consider our simple trip example and suppose that each trip generates external costs of one dollar. If left to their own devices, drivers will make all the trips that have net private benefits between zero and one dollar as well as all the trips that have net private benefits of more than a dollar. But with every trip they make that has net private benefits less than a dollar, they are hurting others more than they are helping themselves. Everybody would be better off if the drivers didn't make these trips and appropriate compensation were paid.

Figure 3.3 illustrates this argument. Suppose that the true net benefits of trips are represented by the usual line through T, but that drivers' private net benefits are greater. That's because they don't experience the external costs they impose. So the private net benefits line goes through V. The number of trips drivers make is V rather than T, since drivers care only about their private net benefits. The trips between T and V impose more costs than they generate benefits, and so on net people would be better off by the triangle W if the trips were not made. Thus W represents the deadweight loss from the externality.

Public policies to curtail trips that cause more harm than benefit are therefore not inappropriate, as they would be if all costs were private. There are, in fact, many such policies: gasoline taxes to deal with road maintenance, emissions standards and vehicle testing to deal with pollution, tort laws to deal with the harm accidents cause. The question is, do they curtail enough (or too many) of the right trips?

The answer, especially for the United States, is that the policies don't curtail enough trips; potential Pareto improvements can be realized by making driving more expensive per mile.

Figure 3.3 The
effects of external
cost.

Net benefits

$

Private net benefits
before external cost is
subtracted

Total net
benefits

T

V

W

Number of trips
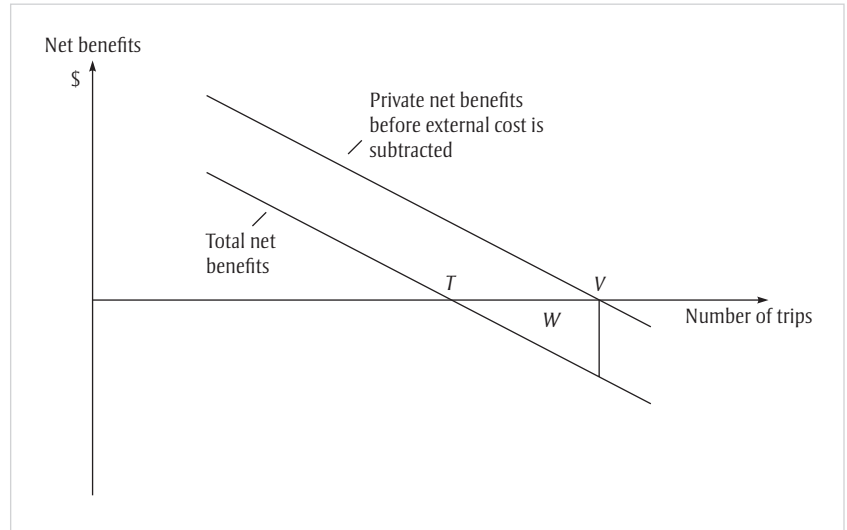
Consider road maintenance and repair first. Cars and trucks damage roads when they pass over them. The damage that any one vehicle causes on any one trip is small, almost invisible, but over time the accumulated effect of many trips is substantial. In contrast, parked vehicles don't damage roadways.

Pavement damage per mile depends on vehicle weight per axle, not on total vehicle weight. Damaging power rises with the cube of the load—so twice as much weight per axle translates into eight times as much damage per mile. The rear axle of a typical thirteen-ton van causes more than a thousand times as much damage per mile as a typical car does. As a result, trucks and buses cause almost all the pavement damage. A truck causes about three cents per mile in road damage, using late 1980s figures (Winston 1991). (Since federal taxes on trucks increase with the number of axles, the federal tax structure encourages the use of fewer axles for any load, and hence more road damage.)

Consider pollution. Road transport is a major source of air pollution. In some cities (Mexico City, São Paulo, London, Los Angeles, for instance), it accounts for more than 80 percent of air pollutants by weight (Brown et al. 1999). Cars and trucks emit carbon monoxide and dioxide, particulates, and the precursors to ozone (particularly hydrocarbons and nitrogen oxides) (Tietenberg 2003).

Particulates are the most costly component of air pollution from vehicles. They arise both from combustion of fuel and from road dust. In humans, particulates aggravate existing respiratory and cardiovascular diseases, alter the body's defense systems against foreign materials, damage lung tissues, and in some cases lead to cancers. Delucchi (2000) estimates that around half of the cost of vehicle pollution in the United States comes from the effects of

particulates on human health. Aside from affecting human health, air pollution damages forests and building materials, reduces visibility, and contributes to global warming. Driving also creates noise and causes runoffs of oil and gas that pollute water.

Assessing a single pollution cost per vehicle mile is impossible. Pollution costs vary greatly by type of car, location (urban driving exposes more people to pollution than rural driving), time (since high concentrations of pollutants are more dangerous than low ones, peak-hour driving is more harmful than off-hour driving), and type and speed of driving. But even though these costs are difficult to quantify, they are still clearly positive, even for driving a new subcompact car in Wyoming in the middle of the night. Small and Kazimi (1995) estimated pollution-damage costs per mile at around three cents for 1992-model cars, six cents for 1977-model cars. Those estimates did not include costs from global warming.

Though there is no direct charge for polluting, the United States has other policies designed to reduce transportation-caused air pollution. Those policies concentrate on reducing pollution per mile driven rather than reducing miles driven. The most successful policy in this line has been the virtual elimination of lead from gasoline through a series of federal regulations issued in the mid-1980s. From 1982 to 1991 lead emissions dropped 90 percent and ambient concentrations fell by 89 percent. Even low levels of atmospheric lead can harm children's health and development rather severely, and so the benefits of this policy have been great. This is particularly true for children growing up in the older, denser parts of American cities, because traffic is heaviest in those neighborhoods. Atmospheric lead also exacerbates high blood pressure in adults, which is linked to many serious conditions.

The other major federal policy directed against transportation-caused pollution has been forcing manufacturers to produce cleaner cars—cars that get more miles for every gallon of gas they burn, and that produce less pollution for every gallon burned. Between the 1970s and the 1990s, this policy made new cars a lot more friendly to the environment. The average car on the road got 13.5 miles per gallon in 1970 and 21.4 miles per gallon in 1998 (U.S. Bureau of Transportation Statistics 2002, table 4-11).

Two recent trends have partially offset this development, however: people are holding on to cars longer—the median age of an American car was 8.3 years in 1999, as opposed to 4.9 years in 1970 (U.S. Bureau of Transportation Statistics 2002, table 1-22)—and they are buying more light trucks and sport utility vehicles (SUVs). (In 1980, light trucks made up only 20 percent of new passenger vehicles; by the turn of the century they represented about half.) Older cars and SUVs cause substantially more pollution than new passenger cars do. While the average passenger car on the road in 1998 got 21.4 miles per gallon, the "other two-axle, four-wheel" vehicle—the category that includes

SUVs—got 17.1 miles per gallon (U.S. Bureau of Transportation Statistics 2002, table 4-12).

Both of these deleterious trends were at least partially caused by the way the federal government chose to fight air pollution. People held on to their old cars longer because pollution-abatement requirements made new ones more expensive, and the exemption of light trucks from the emissions standards rules made them cheaper in real terms for manufacturers to make than other luxury vehicles. Because the law required companies to achieve a certain average miles-per-gallon rate across their entire line of *passenger cars,* a company that sold a line of luxury cars—Lincoln Continentals, say—that got only twenty-one miles per gallon would have to find ways to engineer even cleaner engines for its other cars. This would be an expensive proposition—and more expensive the more Lincoln Continentals were sold—and so would increase the true cost to the company of each Lincoln Continental it sold. A company that instead sold a line of SUVs that got twenty-one miles per gallon would face no such cost.

Fuel efficiency standards also encourage people to drive more. If you use less gas to drive a mile, you'll drive more miles, because each mile is cheaper. Increased driving distance eats up some of the pollution reductions from greater fuel efficiency—estimates place the "rebound effect" at 10 to 20 percent of the initial fuel-use reduction—and it also has implications for accidents and congestion (Portney et al. 2003).

Still, despite these trends and readjustments, highway travel has become less polluting. Even though vehicle-miles traveled rose by 137 percent between 1970 and 1998, carbon monoxide emissions fell by 43 percent; volatile organic compounds (the precursors of ozone), 59 percent; particulates, 41 percent; and sulfur dioxide, 59 percent. Nitrogen oxide emissions rose, but only by 5 percent (U.S Bureau of Transportation Statistics 2002, tables 1-29, 4-40, 4-42, 4-43, 4-45, and 4-41).

This cleaner air has not been free. The price has primarily shown itself in higher prices for new automobiles. But while federal environmental policies restrict the vehicles you can buy and may raise prices sufficiently to discourage some purchases, once you have a car these policies do nothing to restrain you from driving it as much as you want. You can pollute all you want, so long as you do it either slowly or in an SUV.

This is not to argue that current policies have been worthless—far from it, although some economists have concluded that their costs have exceeded their benefits. The point is that current policies do nothing to discourage driving (by car owners), and so people make trips that harm other people through pollution more than they benefit those making the trips.

The kinds of expensive policies that the United States has used to reduce air pollution from motor vehicles also may not be the best policies for the cit-

ies that now have the world's worst air pollution problems. These are cities in developing countries, particularly in Asia and Latin America. Rich countries can afford to use needlessly expensive policies, poor ones cannot.

III. Accidents

The same sort of conclusion applies to accidents and insurance. If you have a car, almost nothing in the current system forces you to recognize the dangers that your driving poses for other people, and so you have an incentive to drive too much—and you get into too many accidents. (Notice that fuel efficiency standards unambiguously increase accidents. They make people drive more, and they don't make them any less likely to have an accident in any mile of driving.)

Of course a legal system that functioned well would be able to handle this problem. People would be held responsible for the full cost of the accidents they caused, and so they would have the right incentives to restrict their driving and to drive carefully. Unfortunately, our legal apparatus for handling accidents, called **tort law** because it deals with the wrongs (*torts* in French) that people commit, is woefully inadequate. Even without any car insurance at all—if, say, accidents were relatively cheap occurrences that people could pay for out of their own pockets—the incentives to avoid accidents would be too weak. Adding the layer of car insurance, a layer with many problems of its own, makes the system worse.

Tort law by itself has two major weaknesses. The first is that it usually holds wrongdoers (tortfeasors) responsible only for the harm that they do to people who can be compensated. Since dead people can't be compensated, tortfeasors are not held responsible for the harm they cause them. To be sure, the surviving relatives of people killed in traffic accidents receive compensation, but the amount is relatively small—generally about $500,000. While claims for wrongful death are permitted in rare cases, still, almost nobody wants to die.

The harm that a death causes is much greater than $500,000. Economists have tried to estimate this harm by figuring out how much more money people would have to receive to be willing to engage voluntarily in risky activities and risky jobs. They've found that modern Americans will voluntarily accept a 0.1 percent increase in the probability of death for an amount between $3,000 and $7,000 (Aldy and Viscusi 2003). To keep things simple, let's say $5,000. If a driver wants to engage in some risky kind of behavior that increases the probability of my death tomorrow by 0.1 percent, such as driving ninety miles per hour on Amsterdam Avenue and ignoring stop lights, then if he is willing to pay me $5,000 for the privilege and does so, we are both better off. But if he's not willing to pay me $5,000 for the privilege of endangering my life—

and a similar amount to everyone else similarly endangered—then stopping him from speeding is a potential Pareto improvement.

How can tort law force him to decide whether speeding is worth at least $5,000 to him? By making him pay $5 million (= 5000/.001) if he kills someone. That's why economists say that $5 million is the "value of a statistical life." It is what people who cause deaths should be made to pay, in order to deter them and everyone else from activities that endanger other people. Killing people is bad, and tort laws that let drivers get away with it for only $500,000 encourage excessive and excessively reckless driving.

(They also encourage the proliferation of SUVs. In crashes with standard vehicles, SUVs are more likely to kill the occupants of the other vehicle than are standard vehicles. Tort laws make death cheaper than many serious injuries. So the actual damage that SUVs do is much greater, the way economists calculate it, than the damage regular cars do, and no one has an incentive to reduce this damage.)

If $5 million is what drivers who kill someone should pay, to whom should they pay it? For the purpose of getting drivers to behave right, they could pay it to the decedent's heirs, although there is no good moral reason why the heirs should receive it; or to a designated charity; or to the government (which could then reduce taxes for everyone); or to me; or they could just burn it in a giant bonfire. It is more important to give than to receive.

This principle carries over to the second major failing of tort law: lawyers, in looking for people to be paid (they are the ones who initiate lawsuits) have ignored the economists' question of who should pay. Think about a two-car crash. Both parties typically suffer damages, and both parties could have acted in some way differently to avoid the accident. Tort law now says that a party should pay damages only if she was negligent and the other party was not; you need not take any responsibility for your recklessness so long as you collide with drivers more reckless than you. (Under the doctrine of contributory negligence, if both of you were very reckless, you won't be held fully harmless in an accident. And you will never be held responsible for the other driver's pain and suffering—and in many states, lost wages—if the driver you hit was more negligent than you.)

But in a two-car accident, even the safer driver is a "but-for" cause of the accident because there is some course of action she could have followed that would have kept the accident from happening. So both parties should each be responsible for the full cost of the accident; only that way will each have the proper incentives to stay home and to drive carefully. This is called double-strict liability. As Vickrey (1968, pp. 481–482) argues,

> The damage involved in the accident could have been totally avoided if any
> party had acted differently, whether by driving less recklessly in the case of
> the "guilty" party, or driving more defensively in the case of the "innocent"

> party, or by accomplishing the purpose in some way not involving the spe-
> cific activity at all, as by travelling by train rather than automobile, or by
> living closer to one's place of work, or even by giving up the object of trip
> entirely. Double (or triple or more) counting at this point is correct; none
> of the activities involved in the accident were economically justifiable unless
> they were able to bear the full costs of accidents which would be avoided if
> the activity were not undertaken . . . Economically speaking, it is just as im-
> portant to provide an adequate incentive for driving defensively rather than
> merely non-negligently as it is to provide an incentive for driving non-neg-
> ligently rather than recklessly.

Thus, even in a system without insurance, both major failings in tort law en-
courage people to drive too much (and too carelessly).

Car insurance serves two purposes: it reduces the risk borne by the per-
son who buys it, and it assures sufficient resources to compensate all victims
(and their lawyers). The damage an accident causes (even measured in tort law
fashion) can easily exceed the financial wealth of almost any household, and
so victims of uninsured motorists are not able to collect what is due them. A
more serious problem from my perspective is that a cap imposed by your
wealth on what you must pay in the event of an accident reduces inappropri-
ately your incentives to stay home and to be careful. If all that you can lose
from an accident is $100,000 (because that's all you have), even if you cause $5
million worth of damage, then you will drive (or not drive) as if the worst ac-
cident you could cause would do only $100,000 worth of harm. Ordinary peo-
ple are risk averse and **judgment-proof,** and so they are not good candidates
to bear well the responsibilities that driving and accidents impose.

Car insurance shifts those responsibilities to an entity—an insurance
company—that is neither risk averse (since the risk of any single motorist is
shared among thousands or millions of stockholders) nor judgment-proof
(since it is required by law to have significant assets). Most states (forty-five in
1997) make liability insurance mandatory. Shifting responsibility is a good
way to solve these two problems. But it also creates a new and perhaps bigger
problem.

That problem is separation of control from responsibility. The insured
motorist is still making the decisions that determine whether a car will have
an accident or not. If the insurer can exercise no influence over the motorist's
actions, the solution is worse than the problem—uninsured judgment-proof
drivers will exercise more care than insured drivers.

Insurance companies, however, use a number of devices to influence
drivers' behavior. They maintain deductibles. They increase the premiums for
people who have accidents. They charge higher premiums for people who say
they drive to work and who live farther from work. They give rebates for cars
that have airbags and antitheft devices. They charge extra for cars with multi-

ple drivers, cars likely to be driven by a teenager, and cars likely to be driven in congested city traffic.

But these devices create only weak incentives to exercise care or refrain from driving. Deductibles are small. After my own car accident, my insurance premiums went up by a few hundred dollars for a few years, even though the accident caused more than $10,000 in damage, not counting the delays that other drivers suffered. Motorists' statements about driving to work are almost never checked by insurance companies, and the differences in premiums for those who do drive are small. Teenagers cost a lot more to insure, but there isn't anything you can do about being a teenager. Since insurance companies exercise such little influence over the actions of the motorists they insure, it's not clear that mandatory insurance is better than no insurance at all. Prohibiting car insurance might be a potential Pareto improvement.

These are not idle theoretical speculations. Cohen and Dehejia (2004) show that mandatory car insurance increases accidents because it reduces the number of uninsured motorists. They found a 2-percent increase in traffic fatalities for every 1-percent increase in insured drivers.

If we kept car insurance, could it be made to work better? There is one obvious way to improve it: charge by the mile. Driving is the leading cause of accidents (collisions between two parked cars are rare), and insurance companies can influence how much people drive. Drivers now treat insurance as a **fixed cost**—you pay it and then you drive as much as you want. The more you drive, though, the more likely you are to be involved in an accident, and so the more you should pay in insurance. If insurance companies charged by the mile, they would be exercising significant influence over an important aspect of their drivers' behavior.

Is this feasible? Edlin (2003), Butler (1993), Litman (1997), and others have worked out a lot of the questions in some detail. Probably the most practical approach at the moment would be "pay at the pump"—adding a surcharge to the price of gasoline that would pay for automobile insurance. The surcharge should differ according to the amount of congestion the car will be driven in, the likely driver's skill and habitual level of care as revealed by her age and record, and the vehicle's weight and fuel efficiency. With information technology and purchases made chiefly by credit card (or a combination driver's license–insurance card), such a system should not be overwhelmingly difficult to manage. And the size of the potential Pareto improvement is large.

Perhaps even a better idea than pay-at-the-pump would be the use of global positioning systems by insurance companies to charge by actual miles driven. Some insurance companies in Texas are experimenting with this. The great advantage of global positioning systems is that they allow different rates to be charged for miles driven under different conditions. For instance, driving at night is considerably more dangerous than driving during the day, and driving at high speed is considerably more dangerous than sitting in the drive-

way burning gas and listening to the radio. Global positioning allows for such distinctions to be made.

If pay-at-the-pump fees and global positioning systems are such good ideas, why haven't insurance companies instituted them already? One reason, of course, is that until very recently the technology was a lot more expensive. A more important reason, as Edlin (2003) points out, is that since most of the accidents a car has are with cars insured by other companies, under current tort law, a company that reduced dangerous driving by its own cars would be primarily benefiting other companies. Only under double-strict liability would an insurance company have the proper incentive to reduce driving by the motorists it covers. Politics and regulation may also play a role: automobile insurance rates and rules are heavily regulated, and regulators look primarily at a company's rate of return; an insurance company that reduced costs might get very little additional profit—or even lose profit—once state regulators adjusted the rates they permitted it to charge.

The magnitudes involved with automobile accidents and insurance are considerable. Americans spend about as much money on the current insurance system as they do on gasoline (net of taxes)—4 cents a mile—but they do not pay on a per-mile basis. The costs of undercompensated deaths, the medical costs that are borne by other payers, and the pain and suffering and lost wages of at-fault drivers increases the cost of accidents by 50 to 100 percent. Double-strict liability increases still further the costs drivers impose on others for every mile they drive, probably to the range of 8 to 12 cents a mile, even higher in more densely populated states like New Jersey and Rhode Island. If people had to pay for the dangers they caused this way, driving would probably fall by about 10 percent, but the net benefit to drivers would be more than $10 billion a year, according to Edlin's estimates.

Motor vehicle accidents aren't a growing problem in the United States. Fatalities per billion miles driven fell from 4.7 in 1970 to 1.5 in 1999 (U.S. Bureau of Transportation Statistics 2002, table 2-17). This happy trend may be partly due to better cars and falling alcohol consumption (see Chapter 17), and it may also be due to better driving, better roads, and better emergency medical treatment. The number of pedestrians killed in accidents in which alcohol was not a factor fell by 24 percent between 1985 and 1999 (U.S. Bureau of Transportation Statistics 2002, table 2-30).

Accidents are a huge problem in developing countries, however, probably because their poorer infrastructure forces pedestrians, bicyclists, and animal carts to mix with car traffic in dangerous ways. In Kenya, for instance, where the rate of traffic deaths per motor vehicle is more than twenty-seven times that in the United States, one observer has described "donkeys, pushcarts, rickety bicycles loaded to the hilt, and barefoot children in full sprint" on the roadways: "Fast moving tractor-trailers surge straight at oncoming cars, forcing them to the shoulder. Minibuses dart back and forth in traffic, with the

limbs of passengers jutting out of open windows and doors. Roads are poorly paved, and people have stolen many road signs for firewood and grills" (Lacey 2004).

Worldwide, about 20 million people a year are injured in traffic accidents, and more than a million are killed. Of fatalities, 85 percent are now in developing countries, and 96 percent of the children killed are in developing countries. After HIV/AIDS, traffic accidents are the major cause of death worldwide in the fifteen- to forty-four-year-old age group (Natulya et al. 2002). As more and more inexperienced people start driving in these countries, the problem will worsen.

IV. Conclusion

U.S. motorists pay for only a small portion of the costs they impose with every mile of driving. They pay about two cents a mile in gas taxes but cause three to five cents in pollution costs and about ten cents in accident costs. Trucks cause an additional three cents or so of road damage costs. As a result, U.S. motorists drive too much—they make trips that cause more harm than they do good.

I'm not saying that there is a class of people—motorists—who are exploiting the rest of us. The argument is about efficiency, not justice. Much of the harm that driving causes falls on people who are also drivers—in other forms, such as the taxes they must pay to repair roads, the bad air they must breathe, the crashes they experience. The problem isn't evil people, it's stupid activities.

Questions

1. Some data show that people who drive more have fewer accidents. Why do you think this is? What does it imply for the arguments for pay-at-the-pump insurance?
2. The primary argument for mandatory car insurance is that it assures that victims of car accidents will be compensated. Evaluate this argument. Is compensation of these victims a reasonable public policy goal? Always? Ever? If so, are there better ways of achieving this goal?
3. If a car is stolen and the stolen vehicle crashes and causes harm to a third party, should the car's owner be responsible? Why or why not? If the stolen vehicle burns gas on which the owner has already paid state and federal gas

taxes, should he get a rebate? Why or why not? Reconcile your answers to the two parts of this question.

4. Should car insurance companies be allowed to charge premiums that differ according to the drivers' race? Why or why not? Should car insurance companies be allowed to charge premiums that differ by neighborhood? Why or why not? Neighborhood and race are highly correlated—in most of the United States, if you know where somebody lives you pretty much know what his race is. Reconcile your answers to the two parts of this question.

5. Some people have argued that the best way to regulate guns is to imitate the mandatory liability insurance system now in place for automobiles. If so, should gun insurance have different rates, like car insurance, based on the gun owner's demographic characteristics and neighborhood of residence? Is double-strict liability a good principle for torts involving guns? Should the insurance premium be collected from the sale of ammunition, like pay-at-the-pump insurance? Should gun owners be liable for damage caused by their stolen guns? Compare with your answer to question 3.

References

Aldy, Joseph, and W. Kip Viscusi. 2003. "The Value of a Statistical Life: A Critical Review of Market Estimates throughout the World." National Bureau of Economic Research Working Paper 10199.

Brown, Jeffrey, et al. 1999. "The Future of California Highway Finance." Berkeley: California Policy Research Center.

Butler, Patrick. 1993. "Cost-Based Pricing of Individual Automobile Risk Transfer: Car-Mile Exposure Unit Analysis." *Journal of Actuarial Practice* 1:51–84.

Cohen, Alma, and Rajeev Dehejia. 2004. "The Effect of Automobile Insurance and Accident Liability Laws on Traffic Fatalities." *Journal of Law and Economics* 47(2):357–394.

Delucchi, Mark A. 2000. "Environmental Externalities of Motor Vehicle Use in the United States." *Journal of Transportation Economics and Policy* 34(2):135–168.

Edlin, Aaron. 2003. "Per-Mile Premiums for Auto Insurance." In Richard Arnott et al., eds., *Economics for an Imperfect World: Essays in Honor of Joseph Stiglitz.* Cambridge, MA: MIT Press.

Litman, Todd. 1997. "Distance-Based Vehicle Insurance as a TDM Strategy." *Transportation Quarterly* 51(3): 119–138.

Marc, Lacey. 2004. "Driving in Kenya Can Be Hazardous to Your Health." *New York Times,* May 4, A3.

Nantulya, V. M., et al. 2002. "Report of the Road Traffic Injuries and Health Equity Conference." Cambridge, MA: Harvard Center for Population and Development Studies. Accessed on July 22, 2002, at *www.hsph.harvard.edu/traffic/conferencereport.*

Portney, Paul R., et al. 2003. "Policy Watch: The Economics of Fuel Economy Standards." *Journal of Economic Perspectives* 17(4):203–218.

Small, Kenneth A. 1992. *Urban Transportation Economics.* Philadelphia: Harwood.

Small, Kenneth A., and C. Kazimi. 1995. "On the Costs of Air Pollution from Motor Vehicles." *Journal of Transport Economics and Policy,* January.

Tietenberg, Thomas H. 2003. *Environmental and Natural Resource Economics.* Reading, MA: Addison-Wesley.

U.S. Bureau of the Census. 2002. *Statistical Abstract of the United States 2001.* Washington, DC: U.S. Government Printing Office.

U.S. Bureau of Transportation Statistics. 2002. "National Transportation Statistics 2001." Accessed at *www.bts.gov/publications/national_transportation_statistics/2001/index.html.*

Vickrey, William S. 1968. "Automobile Accidents, Tort Law, Externalities, and Insurance: An Economist's Critique." *Law and Contemporary Problems* 33(3):464–487. Reprinted in Herbert H. Mohring, ed., *The Economics of Transport,* vol. 1 (Brookfield, VT: Edward Elgar Publishing Co., 1994), 479–502.

Winston, Clifford. 1991. "Efficient Transportation Infrastructure Policy." *Journal of Economic Perspectives* 5(1):113–127.

Chapter 4

Congestion

*T*he final major kind of external cost that drivers impose on each other is congestion. On an otherwise deserted country road, or even on most city streets at four o'clock in the morning, it's difficult for one car to get in the way of another car or to impede its progress, and so congestion is not a problem. But when tens of thousands of cars an hour are converging on the Lincoln Tunnel during rush hour, they get in each others' way, and since the number of cars that can go through the tunnel in any minute is less than the number that want to go through, each car that enters the tunnel during rush hour is forcing all the cars behind it in the queue to wait a few seconds. Because drivers don't bear the cost of the congestion they cause, they cause too much of it.

I'm not saying that drivers don't bear the costs of congestion. Of course they do; being stuck in a traffic jam is unpleasant and time-consuming. But *I* don't bear the costs of the congestion *I* cause; instead I bear the costs of the congestion *you* cause, and so *you* have no incentive not to cause it. Or more starkly, think about a line of cars going through the Lincoln Tunnel. The first car may encounter no congestion, but if it were vaporized, every other car would get through the tunnel more quickly; and so the first car is causing a lot of congestion. The last car, by contrast, is encountering a lot of congestion, but causing none (if it were vaporized no one else would get through the tunnel sooner). What matters is that costs and benefits match on an individual basis, not a class basis.

Most discussions of congestion focus on one specific kind of trip: people going to and from their jobs. Most people don't live where they work, for several reasons. First, there are often greater increasing returns to scale in working activities than in home activities, and it's wasteful to put activities without

---

To get the most from this chapter, you should be familiar with these concepts: consumer surplus, efficiency, elasticity, equilibrium, external cost, marginal cost, Pareto improvement, potential Pareto improvement, progressive, regressive, and transfer. You can find these terms in the Glossary.

increasing returns to scale on land that could be generating increasing returns to scale. Second, there may be bad spillovers between home activities and work activities—for instance, it would not be nice to raise children next to a metal-shredding yard. As a result, many people have to get back and forth between where they work and where they live, and, for the reasons discussed in Chapter 2, many of them work in about the same location. These journeys are called commuting trips.

Not only are commuters' trips concentrated on a few locations, they are concentrated at a few times. A lot of jobs begin in the morning and end in the late afternoon. "Proximity" means "proximity at the same time"; interaction between people at work can take place only if they are at work at the same time. A salesman in an empty office will make no sales; a supervisor whose subordinates are at home can do no supervising.

This spatial and temporal concentration of commuting trips makes them the focus of urban transportation literature. A chain is only as strong as its weakest link. Transportation systems are like storm sewers: they have to be built to a fixed **capacity,** and most of that capacity goes unused most of the time. No one worries about the nice days, when storm sewer capacity is not being used. Storm sewer design discussions focus on the bad days when it rains a lot, and the money that is spent on storm sewers is spent to prevent bad things from happening on those days. "Average daily flow" through a storm sewer is a number that has no practical implications. Peak flow and peak demand are what matter.

The situation is similar for urban roadways. In 1995 only about 31 percent of vehicle miles overall were driven on work trips (so only about 15 percent were trips *to* work), and some of those trips were not made at peak times (U.S. Bureau of the Census 2002, table 1091). But 26 percent of the daily entries into Manhattan through the Lincoln Tunnel were made in the three-hour period starting at 7:00 AM; so were 29 percent of the entries through the Brooklyn Battery and Queens Midtown Tunnels (New York Metropolitan Transportation Council, 1995).

Congestion during peak periods is a major problem in most cities. One estimate for 2002 was that in eighty-five large urbanized areas of the United States, the cost of congestion was more than $63 billion (Texas Transportation Institute 2004). Most of this was the cost of the time that people lost while stuck in traffic. (In contrast, in 2002 all levels of government spent about $38 billion on Temporary Assistance to Needy Families and Food Stamps combined, the largest two welfare programs [U.S. Bureau of the Census 2004, table 519].) Unlike traffic accidents, congestion appears to be growing. The average peak-period driver lost about three times as much time to congestion in 2002 as he did in 1982.

Congestion is also a problem in developing countries. For instance, the Korea Transport Research Institute found that the average speed of cars on

Seoul's roads fell from 11.9 miles per hour in 1980 to 6.3 miles per hour in 1990.

Of course, one could argue that the losses due to congestion are not real losses because they are unavoidable concomitants of the good things of city life—the wetness that inevitably accompanies water-skiing, to use the analogy of Chapter 1. This turns out not to be the case, however; most of these losses are avoidable (although not all, and so the $63 billion estimate for U.S. congestion cost is somewhat overstated).

I. Queues, Reservations, and Tolls

Congestion arises when more people want to use a facility at a particular time than can comfortably use it then. Economists call a situation like this a **bottleneck.**

Bottlenecks are seen throughout life: when forty students want to borrow the same library book the night before an examination, for instance, or when hordes converge on the bathrooms during a concert intermission. Many thousand drivers may want to use the Lincoln Tunnel at 8:30 AM on a workday, but fewer than a hundred can actually emerge from the tunnel in that minute. It can't accommodate everybody. Economists try to think about how various institutional arrangements cause the losses to be allocated, and ask whether some arrangements result in smaller aggregate loss than others.

The most common institutional arrangement for handling bottlenecks is to do nothing about them—to let drivers decide for themselves when to use the tunnel, with no more guidance than perhaps the information from radio traffic reports on the size of the tie-ups. We call this the **come-as-you-please system.**

Let's think about what happens when a large but fixed number of identical people all want to go through the Lincoln Tunnel at the same time in the morning. Under a come-as-you-please system, they will bear two kinds of costs: waiting costs and what economists call **schedule delay costs.**

Time you spend in a traffic jam is time you could have spent doing something else. The cost of waiting is your inability to accomplish what you would have otherwise done in that time. Waiting costs also include the cost of the gasoline you waste while waiting. Waiting costs are proportional to the amount of time you wait.

Schedule delay cost is a more technical term. It is the cost you bear when you use a facility at a time that isn't optimal for you, in order to avoid the waiting—in this case, when you use the Lincoln Tunnel at some time other than 8:30 AM. There are two kinds of schedule delay costs: costs from using the facility late (after 8:30 AM) and costs from using it early (before 8:30 AM). The problems that arise from being late for work or for an appointment are

well known. What are the costs of being early? These are primarily the costs of giving up whatever you wanted to be doing before you left for work, or for the appointment. If your job starts at 9:00 AM but you arrive at 6:00 AM and then sit and read the newspapers, you lose several hours when you could have been home sleeping, getting the kids ready for school, watching television, or working out. That's a cost.

From the point of view of a driver thinking about when to use the Lincoln Tunnel, there is an obvious trade-off between waiting costs and schedule delay costs. If you plan to go through the tunnel early or late, you'll avoid traffic and so your waiting costs will be small, but your schedule delay costs will be big. If you plan to go through the tunnel very close to 8:30 AM, your schedule delay costs will be small but your waiting costs will be big.

Can an economist give a driver useful advice about handling this trade-off? Unfortunately there is no useful advice to give. As a driver facing a bottleneck in a come-as-you-please system, you're damned if you do and damned if you don't. If all drivers are identical and strive to reduce their costs and have been doing so for some time, then no time of arrival can be any better (or worse) than any other. If arriving at the tunnel at 8:05 AM, say, were better than any other time, then everyone would arrive at 8:05 AM. But then 8:05 AM would be a totally miserable time. Conversely, if 8:05 were worse than any other time, nobody would arrive at 8:05 AM and it would be a great time to use the tunnel. Since imagining that any time is better or worse than any other leads to a contradiction, every time has to be equally good (or miserable).

The trade-off between schedule delay cost and waiting cost is therefore a dollar-for-dollar one. Before 8:30 AM, as you think about arriving later, every dollar in lessened schedule delay cost is exactly offset by a dollar in increased waiting cost. After 8:30 AM, as you still think about arriving later, every dollar in lessened waiting cost is exactly offset by a dollar in increased schedule delay costs. Figure 4.1 shows how this works out.

On the horizontal axis in this diagram are the times that people arrive at the tunnel, from the earliest to the latest. The vertical axis shows the cost for arriving at a given time. For everyone, the total cost is the same: $32; that's what is needed to keep people from piling up at any one particular arrival time. For the earliest arriver and the latest, all costs are schedule delay costs; for the person who arrives at 8:30 AM, all costs are waiting time costs.

The total cost of going through the tunnel is the same for everyone in the group. The cost for each individual depends on the size of the group relative to the rate at which cars can go through the tunnel. Think about the person who comes earliest and the person who comes latest. Neither bears any waiting cost, but both bear big schedule delay costs. How big those schedule delay costs are depends on how long it takes the whole group to pass through the tunnel (since everyone has to pass through between the earliest person and the latest). But since everybody else's costs are the same as those of the earliest

Figure 4.1 Come-as-
you-please system.
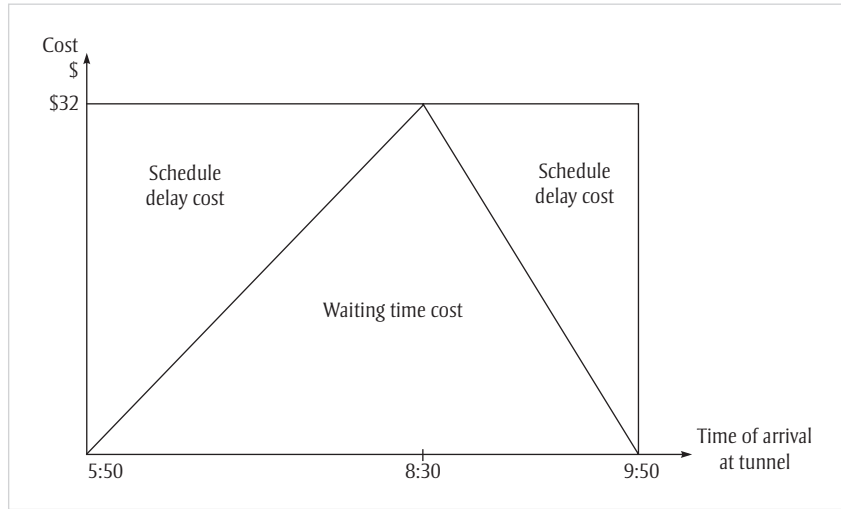
person and the latest, everyone's cost depends on the length of time it takes the whole group to go through the tunnel. The bigger the group, the worse for everyone, including the person who gets to go through the tunnel at precisely 8:30 AM.

Economists call a situation in which everyone is doing as well as possible, given correct beliefs about how everyone else will behave, an **equilibrium.** The situation depicted in Figure 4.1 is a come-as-you-please equilibrium; no one has any reason to change.

Notice that in the come-as-you-please system, every driver except the first and last is imposing an external cost on some other drivers. Suppose some driver (except the earliest) coming before 8:30 AM decided to stay home or use an alternative transportation mode. Then everyone entering the queue earlier than this driver could have left home a little bit later and still waited just as long—so they all would have reduced their schedule delay costs a little bit. Under a come-as-you-please system, the closer to the optimal time you arrive at the tunnel, and the more time you spend waiting in line, the greater the burden you impose on other people. You may not think of what you're doing as having that effect because you can't see the harmed parties writhing in agony—you may not be able to see them at all—but by taking something they would otherwise have had you are harming them just as surely as if you had stuck a gun in their ribs and extracted some money from their wallets. Nice people impose costs, too. (See Box 4A for another example.)

Is there a better way? A popular restaurant would never operate like this. You would have to make a reservation to eat at the most desirable times. People would be disappointed when they could not all get a reservation for the time they wanted most—so there would be schedule delay costs—but there

See for Yourself

☰ A numerical example can help you understand come-as-you-please equilibrium and later see how it compares with other ways of handling a bottleneck.

Suppose that the Lincoln Tunnel can accommodate 1,000 cars an hour, but that 4,000 drivers want to enter the tunnel at 8:30 AM so they can get to work at 9:00. Every one of these 4,000 drivers is identical: they all want to get to work at 9 AM; going in early costs each of them $0.20 per minute, going in late costs them $0.40 per minute, and waiting in line costs them $0.30 a minute. Find the come-as-you-please equilibrium, and describe how long the bottleneck lasts and what costs are imposed.

Step 1: How long does it take for everyone to get through the tunnel? It's going to take 4,000 cars/1,000 cars per hour, or 4 hours, no matter what system is used.

Step 2: When do the latest and earliest drivers arrive? It's twice as bad to be late by a minute as it is to be early by a minute ($0.40/$0.20), and so the earliest arrival must be twice as early as the latest arrival is late; otherwise the two arrival times would not be equally attractive. Since the two times are 4 hours (240 minutes) apart, the rest is simple: the earliest arrival time has to be 160 minutes before 8:30 AM, and the latest has to be 80 minutes after. The queue starts to form at 5:50 AM and dissipates at 9:50 AM.

Step 3: How much total cost does each driver bear? The cost that the earliest arriver bears is $32 ($0.20/minute × 160 minutes); this is the same as the cost that the latest arriver bears ($0.40/minute × 80 minutes). Everyone in between must bear this same cost. This is illustrated in Figure 4.1.

Step 4: What is the total cost? How much is schedule delay cost? How much is waiting cost? It's easy to calculate what the total cost is: $128,000 (4,000 × $32). From Figure 4.1 it's also obvious that half of the cost ($64,000) is waiting time cost (that comes from the formula for the area of a triangle, one-half base times height) and half is schedule delay cost. This is what happens when people are left alone to try to do as well as they can.

---

would be no waiting costs. No one would show up when a table wasn't available (barring unexpected delays and other mishaps). When people can make reservations, they never have to stand in line.

What would a system of reservations for the Lincoln Tunnel look like? The amount of time it takes to get everybody through the tunnel can't change, and so the interval between the earliest time and the latest time would have to be the same as under come-as-you-please. The schedule delay cost at the earliest time would still have to be the same as the schedule delay cost at the latest time: if the schedule delay cost at the earliest time were less than the schedule

Figure 4.2 Reservation system.
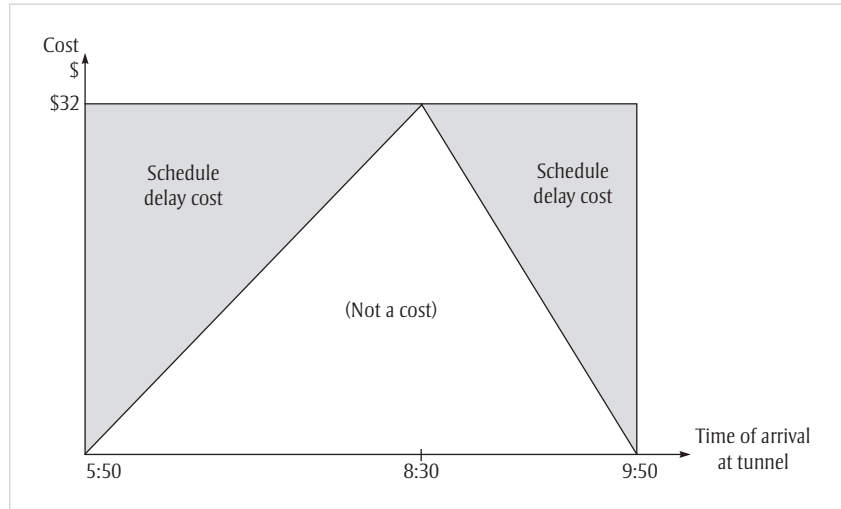
delay cost at the latest time, you could cancel the latest appointment and re-schedule it right before the earliest appointment, and make somebody better off without making anyone worse off; the reverse procedure works in the opposite case. So the arrivals at the tunnel would still run over the same interval as under come-as-you-please, and the schedule delay cost would be the same. But there wouldn't be any waiting cost. Thus a reservation system would cut total costs substantially.

The shaded area in Figure 4.2 represents the costs when a reservation system is in effect. Almost everyone is better off with a reservation system, and no one is worse off. In contrast, the come-as-you-please system, in which everyone is free to decide when to use the tunnel, hurts everyone.

Notice, however, that with a reservation system in effect, some people are better off than others; when people decide on their own, everyone is equally miserable. With reservations, the person who gets the most convenient time (total cost 0) is much better off than those who get the earliest or the latest and have to bear large schedule delay costs. In fact, the earliest person would be willing to pay the entire amount of her schedule delay cost to get the reservation at the most convenient time. This is another reflection of the external costs imposed by people who go through the tunnel at desirable times. (See Box 4B.)

With current technology, of course, setting up a reservation system for the Lincoln Tunnel is not a practical alternative. Thinking about a reservation system, though, suggests that a lot of congestion cost is unnecessary; if a simple idea like reservations could cut the cost of congestion in half, there should be some more sophisticated way of doing the same thing. Is there?

The sophisticated way of making reservations is to set up a system of **congestion pricing:** in this case, to impose a toll that varies by arrival times

See for Yourself

☰ *How much cost does a system of reservations eliminate in our example of a come-as-you-please equilibrium?* From Step 4 in Box 4A, we know the waiting time costs are $64,000, or half of the total cost. A reservation system eliminates these costs.

How much external cost does someone arriving at or before 8:30 AM cause? There are two ways of answering this question, depending on what you imagine happening if this particular person decided not to use the tunnel. Fortunately, both methods always give the same answer.

The easy method is to imagine that the driver you are interested in would be replaced by the person coming earliest—in other words, the person who had the worst reservation would take the reservation of the driver who decides not to use the tunnel. The earliest person had schedule delay costs of $32 originally, and so if the position being vacated had schedule delay costs of *s* dollars, the earliest person would gain

$$32 - s.$$

That's one way of calculating external costs.

Alternatively, if all the drivers coming earlier moved up a little bit, all would gain a little bit. The answer will be the same.

Why? In the first method of thinking about the problem, one driver moved in one long jump from the earliest position to the position being vacated. Instead of one long jump, we could have thought of this one driver as making a series of short hops, each hop a move to the next best time that a driver was arriving, and then added up the gains of those short hops. Since the series of short hops covers the same ground as the big jump, the gain has to be the same. Then, instead of imagining one driver making all of the short hops, we could think of a series of drivers making one short hop each—each driver moving up one place. These are the same short hops that the single driver made, and so the aggregate gain has to be the same. The series of short hops is the second method, and the long jump is the first method, and we get the same answer both times.

and that is equal to the cost a driver imposes on others, shown in Figures 4.3 and 4.4. The discussion of pollution in Chapter 3 argued that making people bear all the costs their decisions impose on other people is a good way to promote **efficiency,** and a time-varying toll does this. A tax that makes people bear all the costs of their decisions is called a **Pigouvian tax** (when the decisions help other people rather than harm them, the subsidy is called a **Pigouvian subsidy**). So congestion pricing is another example of a Pigouvian tax. The toll also makes every time equally attractive—the total cost is the

Figure 4.3 Social costs with no waiting.
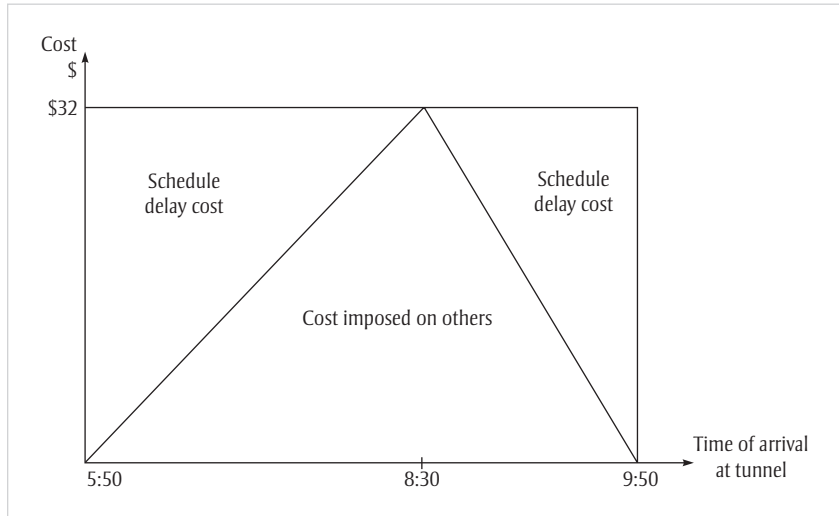
Figure 4.4 Optimal toll system.
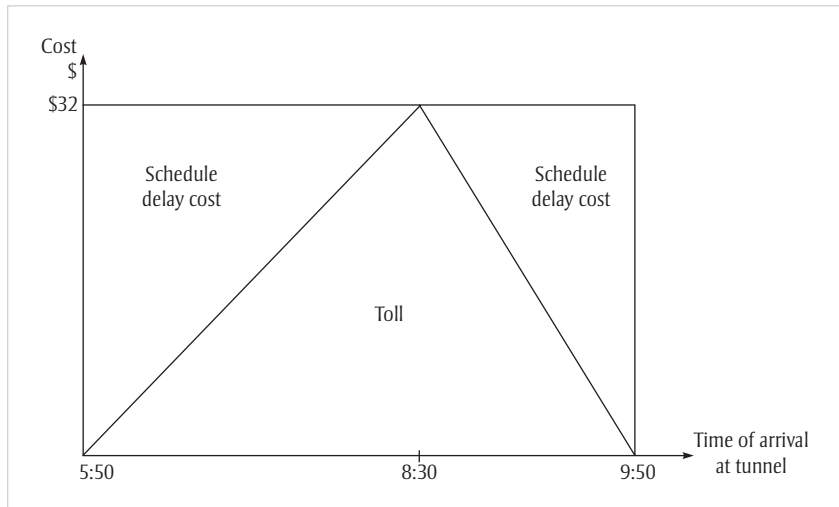
same, consisting of schedule delay cost and a toll cost—and so no one wants to change his or her arrival time.

Establishing a congestion toll is just like setting up a reservation system and charging enough for each reservation so that no one envies anyone else— that way people don't want other people's reservations and you don't have to do anything to enforce the system. And because whoever is administering the system can identify the person arriving at, say, 8:23 as the person with the 8:23 reservation, who should therefore pay the price of the 8:23 reservation, you can run a "reservation system" without actually going to the trouble of mak-

ing reservations. That's the beauty of time-varying congestion tolls as proposed by Vickrey (1969).

Of course, at first it doesn't look like any drivers are better off under a system of time-varying congestion tolls than they would be under the toll-free come-as-you-please system. Consider the numerical example. With congestion tolls, everyone bears a cost of $32—partly in tolls, partly in schedule delay costs. With toll-free come-as-you-please, everyone bears a cost of $32—partly in waiting costs, partly in schedule delay costs. From the point of view of a driver, congestion tolls merely substitute out-of-pocket toll costs for waiting costs.

But in the larger picture, there's a big difference between out-of-pocket costs and waiting costs. Waiting costs are pure waste, but tolls are a **transfer.** Nobody gains from waiting costs, but the toll authority gains from tolls (as does anyone who benefits from whatever the toll authority does with the revenue). The toll authority could, for instance, rebate $16 each to everyone, and everyone would be better off than under come-as-you-please. Or it could just rebate a dollar each and use the rest for prodigality and parties; everyone would still be better off than under come-as-you-please. Installing a congestion toll system is like making a large amount of money drop from the sky.

Notice also that we are assuming that drivers themselves bear all the waiting costs. Waiting in a traffic jam, however, burns gasoline, and gasoline causes air pollution—a cost that not just drivers bear. One estimate is that congestion pricing would reduce emissions by 10 percent in the aggregate in parts of Delaware, and by 30 percent in crowded areas; reduced emissions would amount to between 15 and 30 percent of the total benefits of congestion pricing (Daniel and Bekka 2000).

Many other kinds of congestion tolls would also work like a reservation system. For instance, the toll authority could pay drivers: $32 to the people who arrive earliest and latest, nothing to those arriving at the most convenient time. This would act just like a reservation system, making people indifferent about when they travel. This would cost the toll authority a lot of money, but it could raise the sum through a tax of $16 per capita, and everyone would be glad to pay it. Or the toll system could be revenue neutral: it could charge $16 for the driver at the most convenient time, and pay $16 to the drivers at the earliest and latest times.

II. Capacity

Congestion tolls are one way to get rid of congestion, and they do it for free. The more popular approach is to build more roads and add more lanes—to increase capacity. This approach, though, has several drawbacks, aside from the obvious problem of being expensive.

If the number of drivers is fixed, doubling capacity lets them get through the tunnel in half the time, no matter how their use is regulated. More capacity allows a fixed number of drivers to impose smaller external costs on each other.

But assuming that the number of drivers is fixed is not realistic. If the Lincoln Tunnel becomes a more convenient way to get into New York, it will draw commuters from the Holland Tunnel, the George Washington Bridge, and the Tappan Zee Bridge; some bus riders and rail passengers will switch to cars (or use cars more frequently if they occasionally drive to the city anyway); some people will make car trips into Manhattan that they would not have made (professors on days that they don't teach, shoppers and theatergoers); some people who live in Manhattan and work there will decide they can live in New Jersey and still get to work without too much hassle. Heretofore we have been explicitly assuming that, no matter the cost, only a fixed number of drivers wanted to use the Lincoln Tunnel during rush hour, but clearly the number of people who want to use the tunnel depends on the cost of using it. Demand is not totally inelastic.

To see the effects of additional drivers, let's first go to the other extreme and suppose that demand is totally elastic. That is, all potential drivers are identical, there are lots of them—way more than the fixed number we have been thinking about—and the benefit of the trip for each of them is $32. With the original tunnel capacity, the original number of drivers will go through the tunnel, under either the come-as-you-please system or the standard optimal congestion toll. If more than that tried to go through, the cost of a trip would rise above $32 and making the trip wouldn't be worthwhile for anyone anymore; if fewer than that tried, the cost would fall below $32 and everyone would want to go.

Now suppose capacity doubles under the come-as-you-please system, but with perfectly elastic demand. Once everybody adjusts, the cost for everybody still has to be $32. So the earliest and the latest drivers will arrive at the same times that they did before capacity doubled. In between those times, twice as many people will go through the tunnel, since capacity has doubled. But each of them will have to wait in line just as long, since waiting costs have to be the same. The only difference is that the queues will contain twice the number of cars, since they move twice as fast.

Who will gain from building the additional capacity? No one. Not the original drivers—they are still bearing a cost of $32 each. Not the new drivers either, for while they are now making a trip with a $32 benefit, but they are also now bearing a $32 cost. They're no better off. All the construction has been wasted because it just gave more drivers opportunities to impose costs on each other.

This tendency of additional drivers to overwhelm the effect of any physical attempt to relieve congestion has often been described. Anthony Downs

(1962) dubbed it the "law of peak-hour expressway congestion." Here is how Robert Caro (1975, p. 516) describes the evolution of traffic in eastern Long Island in the late 1930s:

> The Triborough Bridge opened on July 11, 1936 . . . City officialdom and press assured the public that the project would, as [Robert] Moses had promised, provide at a single stroke a solution to most of the traffic problems between Long Island and New York . . . Police Commissioner Valentine said that the opening of the bridge would sharply reduce traffic on all the East River spans, on the Queensborough "by from 40 to 50 percent."
>
> On August 17, 1936, a little more than a month after the Triborough Bridge opened, Long Island's parkways were the scene of what some observers called the greatest tie-up in the history of the metropolitan area . . . "The Interborough Parkway, barely opened, was already impassible," Brooklyn Borough President Ingersoll noted.

With congestion tolls, the situation is not nearly so dismal. The flood of new drivers assures that every driver's cost will be $32 after the new construction. Since there are no lines, the peak toll must be $32, as before. The earliest and latest drivers, who pay no tolls, must bear $32 in schedule delay costs, and so they must arrive at the same times that they did before the new capacity was added. But then twice as many drivers pass through the tunnel between the earliest and the latest drivers, and twice as much toll revenue is collected.

Notice that this increase in toll revenue is the total increase in benefits, since drivers are neither worse off nor better off. Unlike in the come-as-you-please case, here there are benefits to expansion, even when demand is totally elastic. When demand is totally elastic, congestion pricing gives the toll authority enough information to judge whether an expansion in capacity is worthwhile, and aligns its incentives properly. Adding the capacity is a potential Pareto improvement if and only if the construction cost is less than the added revenue, so a profit-maximizing toll authority will add capacity (or build a tunnel in the first place) if and only if doing so is a potential Pareto improvement.

This neat efficiency result for elastic demand in our model holds only for the toll schedule where the maximum toll is $32. The alternative toll schedules that collect less money—for instance, those with maximum tolls of $16 and of $0—allocate a fixed number of drivers efficiently among the various times of arrival at the Lincoln Tunnel, but they do not allocate people efficiently among such alternatives as staying home, taking a train, and using the Tappan Zee Bridge. A person who uses the Lincoln Tunnel at 8:30 AM is imposing $32 in costs on the rest of the world; unless anyone who does so pays $32 for the privilege, too many people will want to do so. Thus, for instance, if the toll authority rebates the toll revenue to drivers, the rebate should go to potential users of the tunnel, not just actual users; otherwise it will encourage tunnel use.

What happens when demand for tunnel crossing is neither totally elastic nor totally inelastic—when an increase in the cost of using the tunnel will discourage some users but not all? If we assume that people are identical in the costs they incur (both for waiting time and schedule delay) but differ in the benefit they derive from a trip through the Lincoln Tunnel, we can line people up in order of decreasing benefit (or we can line up their numerical benefits) and draw a demand curve like that in Figure 4.5. On the horizontal axis in Figure 4.5 is the number of cars using the tunnel. On the vertical, for any number X of cars using the tunnel is the Xth greatest benefit. For instance, if the four-thousandth greatest benefit was $32, then the 4,000 drivers with benefits of at least $32 would use the tunnel when it cost $32, and no one else. If the demand curve is very steep (inelastic demand) then we have a situation like the first one we studied, when 4,000 people used the tunnel no matter what. If the demand curve is very flat (elastic demand), then we have a situation like the second one we studied, where the benefit is going to be $32 no matter what.

The area under the demand curve has an important interpretation. Consider the shaded area in Figure 4.5. If the X' people with greatest benefits go through the tunnel, this area is the total benefit they receive. This is just like **consumer surplus** under a regular demand curve.

So the demand curve tells us for any cost of using the tunnel how many people will use it. The cost of using the tunnel, in turn, depends on how many people use it, and on its capacity. This allows us to derive a "supply curve." Suppose capacity is 1,000 cars an hour. With either a come-as-you-please system or congestion pricing, the cost to everyone who uses the tunnel will be the same. The earliest user will not encounter congestion under either system, and

Figure 4.5 Demand curve.
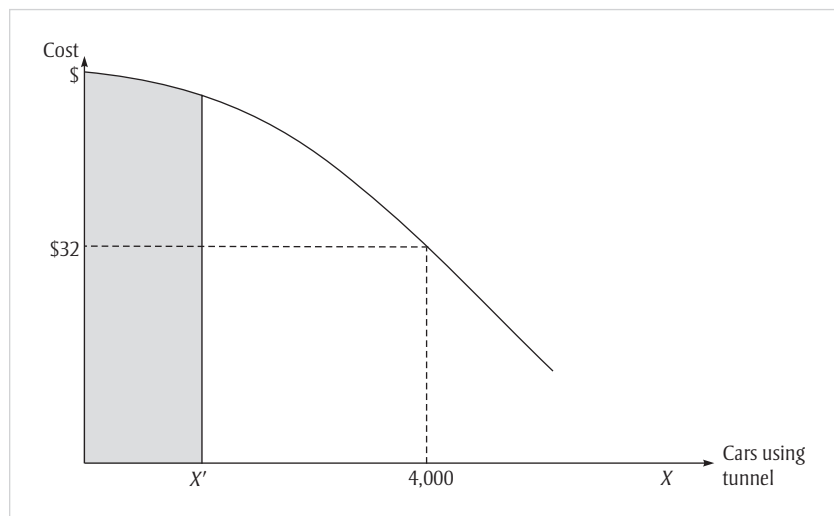
Figure 4.6 Supply curve.
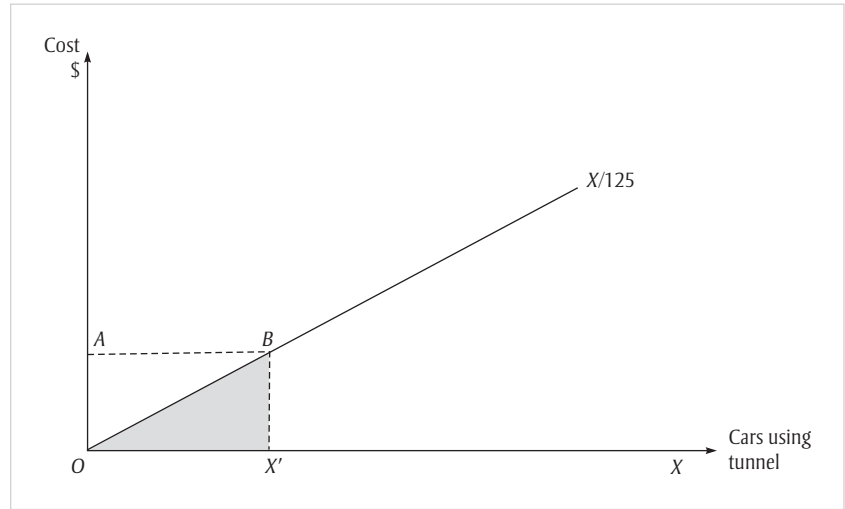
so it is easy to calculate her cost; everyone else's cost will be the same as hers. The more cars that use the tunnel, the earlier the earliest user will have to arrive and the greater the cost that everyone who uses the tunnel will bear. This is why the supply curve slopes up. The supply curve is shown in Figure 4.6.

The area under the supply curve has an interpretation, too. It's the total amount of schedule delay cost that drivers will experience if X' drivers use the tunnel. The quickest way to understand this is to realize that the rectangle $OABX'$ represents the total cost that drivers bear when X' of them use the tunnel—it is equal to the number of drivers (X') times the cost for each of them ($X'B$). Since schedule delay cost always equals half of total driver cost, and the shaded area under the supply curve represents half of rectangle $OABX'$, the shaded area equals the schedule delay cost.

Alternatively, the driver with the greatest amount of schedule delay has schedule delay costs of $X'B$, the driver with the least has schedule delay costs of zero, and in between, drivers with every intermediate amount of schedule delay cost are equally numerous. Adding up all the drivers gives you the shaded triangle once again.

The intersection of the demand and supply curves tells us how many people use the tunnel: the number of cars that generates the cost that precisely that many people are willing to endure, as Figure 4.7 illustrates. This is an equilibrium. Figure 4.7 shows us more, as well. The area under the supply curve represents the schedule delay cost that drivers bear. The area under the demand curve but above the line AB is the net benefits that accrue to drivers: drivers get the entire area under the demand curve as benefits but incur costs equal to the rectangle $OABX$. If the demand curve is flat, this consumer surplus area doesn't exist—everyone gets zero net benefits.

Figure 4.7 Equilibrium.
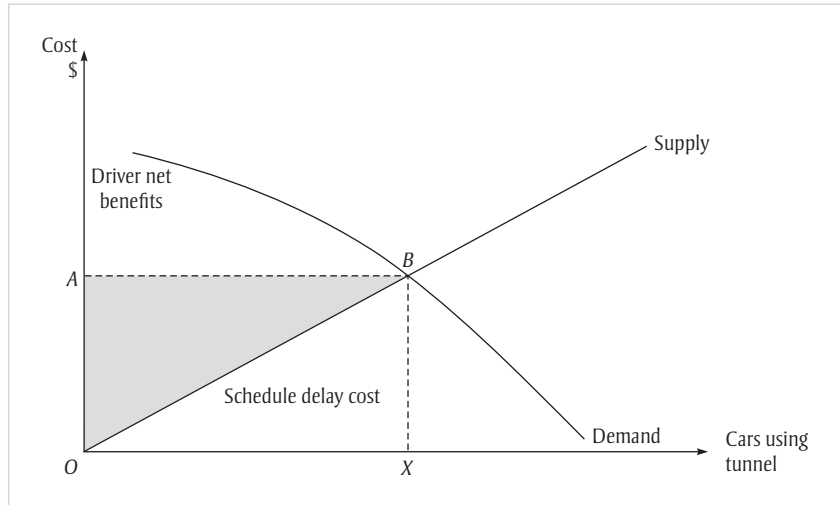
The shaded triangle *OAB* in Figure 4.7 is the only place where congestion pricing differs from a come-as-you-please system. Under a come-as-you-please system, this triangle represents waiting time costs; under congestion pricing, it represents revenue from the tolls.

Thus when we think about congestion pricing as a way of alleviating a classic bottleneck, we see that two common beliefs about congestion pricing appear to be wrong. First, congestion pricing in a classic bottleneck doesn't reduce the number of cars making a specific trip or the amount of driving. The number of cars going through the tunnel is the same under congestion pricing as it is under come-as-you-please. Second, congestion pricing is not even a way to get people to spread out their arrival times at work. People go through the tunnel at the same rate for the same period under congestion pricing as they do under come-as-you-please. They arrive at work at the same times. The only thing that congestion pricing does is transform the waste of waiting time into money that can be used for good purposes (or not).

These very stark, somewhat unusual conclusions arise from the very simple way we've been thinking about the physics of congestion. More sophisticated ways of thinking yield a slightly different picture in this area. The major difference between our approach and the more sophisticated traffic models is in the full array of costs that drivers impose on each other. In the way we've been thinking so far, driving is like eating dinner in a popular restaurant at a peak time: the only external cost is that you prevent someone else from eating dinner then; you don't interfere with other diners or make their meals longer or less pleasant. With traffic, however, there is often another externality as well: you slow down the people who are using the road at the same time you are. This additional externality means that sometimes optimal tolls should re-

duce the number of drivers at a particular time, not just induce them to avoid waiting time.

Such more sophisticated models are beyond the scope of this book, but their lesson shouldn't be ignored. What the simpler approach tells us should not be ignored either: congestion pricing has major advantages even if it doesn't reduce traffic flows or spread out the peak.

III. Reasons for Tolls in Popular Discourse

Notice that the reasons we are giving for tolls are almost totally different from the reasons generally given for tolls in popular discourse. Paying off construction is not a good argument for tolls (although if demand is elastic enough the right toll will pay for the right amount of construction); regulating traffic flow is. There's no good reason to have a toll on an empty stretch of new road that hasn't been paid off; there's plenty of reason to have one on a congested stretch of road that has been paid off. Since roads are generally uncongested right after they're built and become more congested as they become older (it takes time for people to find out about new roads, learn their advantages, and then adjust where they live and work), the idea that tolls should be used to pay off a road's construction cost leads to the worst possible allocation of resources: high tolls at first that discourage trips that should be made, and no or low tolls later that encourage waiting that should not be done. It's like taking a vile-tasting medicine when you're young and healthy, and stopping it when you get sick and it could help you.

Of course, the kind of tolls we are talking about here are the ones that only cost drivers money. Until recently, most tolls were not like this: they cost drivers time, too. You had to stop at a tollbooth, roll down your window, fumble with change, chat with the tollbooth attendant or aim for the coin basket, and roll your window up again as you tried to avoid being hit by other cars hurrying to make up for lost time. Often there were lines at the toll barriers, creating congestion rather than alleviating it. This system of collecting tolls is now technologically obsolete. Transponder systems like E-ZPass can identify cars traveling at speeds of up to eighty miles an hour, and collect tolls automatically by electronically debiting the appropriate accounts. When you picture toll collection, you should think about these systems, not tollbooths.

IV. Equity

Many objections to congestion pricing systems have been raised. Most we will discuss in the next section. One objection, though, requires a more detailed analytic discussion. It's the objection on equity grounds: that by charging a

price for something that was formerly free, congestion pricing will unfairly burden poor people and benefit rich people.

In the framework we have been using, we can't evaluate this objection. So long as every driver has the same time costs, every driver, rich or poor, is just as well off under congestion pricing as under come-as-you-please; the toll authority can make everyone better off and no one worse off. But since the value of time is tied to wages, and rich people generally have higher wages than poor people, rich people have different values of time than poor people do. So congestion pricing might have different effects on different classes of people.

Whether or not it actually does depends on the **elasticity** of demand and on the ways in which poor people differ from rich. If demand by both classes of people is infinitely elastic, congestion pricing by itself can't hurt either group. Both have zero net benefits with come-as-you-please; both have zero net benefits with congestion pricing. The toll authority gains from congestion pricing, and which group benefits more depends on what the toll authority does with the added toll revenue.

Similarly, even when demand is totally inelastic poor people don't lose with congestion pricing if rich people's time is uniformly more valuable than poor people's. But rich people can gain. The reason is that congestion pricing can induce different classes of drivers to sort themselves better into the different time slots. To minimize total costs, people who suffer the greatest per minute schedule delay costs should be using the tunnel at the most desirable times (this is the same principle that says ambulances and fire trucks, vehicles with huge schedule delay costs, should have priority on all streets). Better sorting is an advantage of congestion pricing that we couldn't examine earlier—you can't sort people who are identical. (This analysis follows Arnott, de Palma, and Lindsey 1994.)

Whenever rich people have higher per-minute schedule delay costs than poor people do, congestion pricing schemes will induce rich people to arrive close to 8:30 AM and poor people to arrive far from 8:30 AM. This intelligent sorting pattern won't always evolve in come-as-you-please systems. Under come-as-you-please, the people who have the highest ratio of schedule delay cost per minute to waiting cost per minute will occupy the most desirable times. Under come-as-you-please, you pay for convenient arrival times with waiting, while under congestion pricing you pay for convenient arrival times with money. To be able to arrive at a convenient time under come-as-you-please, you have to wait a very long time; under congestion pricing, you have to sacrifice a lot of other consumption.

With perfectly inelastic demand, the only way that congestion pricing can make poor people worse off before distribution of toll revenues is by forcing them out of the most convenient arrival times. Such a scenario is possible when schedule delay is relatively more expensive for poor people than waiting but not absolutely more expensive for them. This is not implausible—if poor

people work on assembly-line jobs where they will be severely penalized for being late (since one person's absence keeps the whole line from moving) while rich people have more flexibility, then getting to work late will be relatively more expensive for poor people than rich; if poor people have heavier family responsibilities before they head to work, leaving home early will be relatively more expensive for them than waiting time.

Under come-as-you-please, these relatively more expensive schedule delays will mean that poor people, not rich people, will arrive at the more convenient times around 8:30 AM. They are relatively good waiters, and come-as-you-please rewards waiting. With congestion pricing, poor people and rich will change places, since waiting abilities are no longer relevant: rich people will arrive at the convenient times, poor at the inconvenient times.

This switch is a potential Pareto improvement. Suppose a poor person who values early time at twenty cents a minute is arriving at 8:30 AM under come-as-you-please and a rich person who values early time at forty cents a minute is arriving 100 minutes earlier. If they switch positions—this is what congestion pricing would do—the poor person loses twenty dollars while the rich person gains forty dollars. The winner could compensate the loser.

Figures 4.8 and 4.9 illustrate a particular example in which congestion pricing makes poor people worse off. Demand is perfectly inelastic. The come-as-you-please equilibrium is shown in Figure 4.8. Rich people arrive very early and very late and bear enormous schedule delay costs. They are unwilling to arrive at more convenient times because the lines are too long, full of poor people with low waiting costs per minute. Poor people arrive at the convenient

Figure 4.8 Come-as-you-please equilibrium with poor and rich drivers.
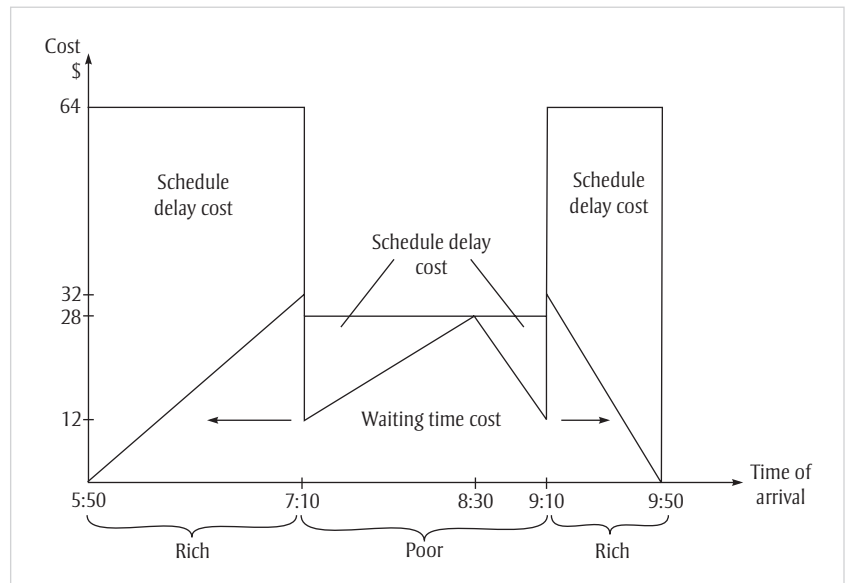
times and spend an awful lot of time waiting, because the schedule delay costs per minute are too great for them to move to times with shorter lines.

In Figure 4.9, the positions in the equilibrium are reversed with optimal congestion pricing. At the less convenient times, when poor people arrive, the toll rises more slowly as schedule delay decreases. This makes poor people indifferent but would make any rich people who found themselves in these time slots want to cut their schedule delay costs. The toll rises more quickly at the more convenient times, when rich people are arriving. Rich people are indifferent, but any poor people who found themselves in these time slots would want to cut the toll they were paying. The dotted lines in Figure 4.9 show what rich people would be willing to pay to arrive at the poor people's times and what poor people would be willing to pay to arrive at the rich people's times. Since what rich people are willing to pay to arrive at the poor people's times is less than the toll being charged, and since the same applies to what poor people are willing to pay to arrive at the rich people's times, no one is tempted to leave his or her time slot.

Figures 4.8 and 4.9 are derived from a numerical example that is presented in the Appendix at the end of this chapter.

This example assumes that demand by both rich and poor people is perfectly inelastic. If elasticities are different, the results can be different. We have already seen that neither group wins or loses when both have perfectly elastic demand, as may be the case over long enough time horizons. In general, higher elasticity of demand reduces the amount a group can gain from the adoption of congestion pricing, but it also restricts the amount it can lose.

Figure 4.9 Optimal toll equilibrium with poor and rich drivers.
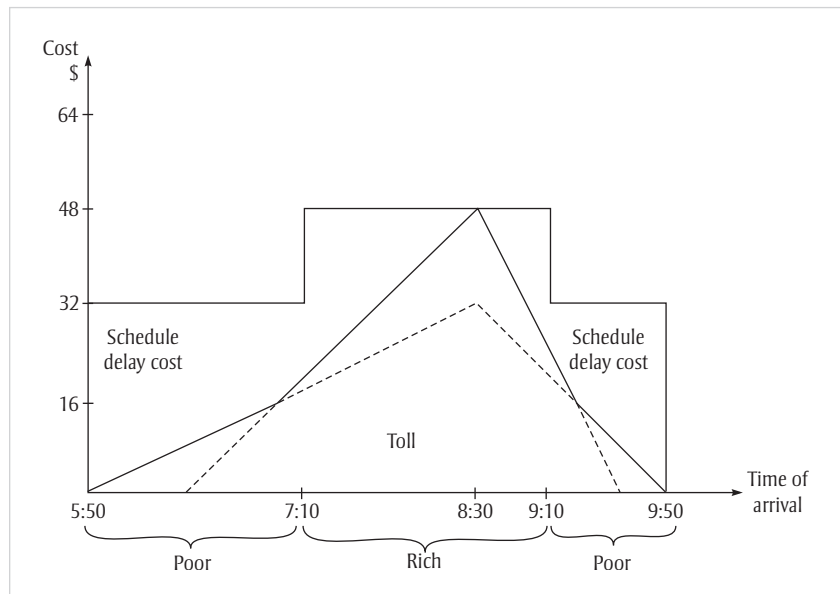
High elasticity of demand is like insurance. Thus in the example here, if poor people's demand for travel were highly elastic, they would not lose much from the adoption of congestion pricing; instead, fewer of them would use the tunnel.

On the other hand, high elasticity of demand in the *other* group amplifies any gains or losses that a group realizes from congestion pricing. In the example, if rich people had highly elastic demand, the gains that congestion pricing gives them would induce more rich people to use the tunnel. The increased congestion would make poor people even worse off. If poor people had highly elastic demand, the losses from congestion pricing would shrink their number. Since fewer people would be using the tunnel, rich people would gain even more.

What are we to make from all this about whether congestion pricing is **regressive** or **progressive?** It seems clear that under many circumstances congestion pricing would be regressive, assuming the toll authority squandered all the revenue. But there are also many circumstances under which it would be progressive or neutral. The issue hangs on the ratio of per-minute schedule delay costs to per-minute waiting costs. This parameter is not likely to be the same always and everywhere, and so the regressivity or progressivity of congestion pricing is not likely to be the same either.

Most empirical studies of modern U.S. cities have reached the conclusion that congestion pricing would be mildly regressive (see Segal and Steinmeier 1980, Small 1983, Anderson 1996; all these studies assumed, as usual, that the toll authority would squander all the revenue). But we should remember that some of the facts about life in U.S. cities in the 1970s and 1980s that led to this conclusion may not always hold. In particular:

1. *Rich people have higher relative waiting costs.* The more that rich people can do in their cars, the less likely this premise is to hold. Most empirical studies predate the wide availability of cellular phones. Better stereos, CD players, in-car faxes, and wireless Internet connections could all reverse this relationship. If it were to reverse, congestion pricing would not change how times are allocated and so would not push poor people out of the times with least schedule delays; rich people would be using these times even with come-as-you-please.

2. *One person to a vehicle.* If poor people are sharing rides or taking mass transportation while rich people are not, then congestion pricing won't push them out of the times with least schedule delays. Congestion pricing allocated the times closest to 8:30 AM to the vehicles with the greatest absolute schedule delay cost. A carpool with three twenty-cents-a-minute poor people incurs more schedule delay cost per minute than a car with a single forty-cents-a-minute rich person; a busload of poor people represents more schedule delay cost than a

very expensive executive. Poor people are more likely to carpool or use buses than rich people are, and so their losses, if any, from congestion pricing are likely to be somewhat less than we have calculated.

3. *Peak-hour drivers are a cross section of the population.* Although the poorest people don't have cars and many other poor people work unconventional shifts, this premise may not be a bad approximation to reality in the United States outside, perhaps, New York. But in many developing countries, including some of those with the worst traffic jams, the people with cars are in the upper strata of society; even the poorest drivers are relatively wealthy. To the extent that toll revenues allow governments to provide general services and reduce general taxes, congestion pricing in these countries can be progressive.

Aside from these (important) technical considerations, the question of whether congestion pricing is ultimately regressive or progressive also hinges on the reactions of two other types of actors. The first is employers: they can, after all, change the times that workdays begin and end and alter their policies on how they treat workers who are early or late. If congestion pricing makes it more difficult for low-wage workers to get to work on time, employers will have to rethink their job requirements. Requiring a secretary to be at work at 9:00 AM is less costly for his employer when he doesn't have to pay for displacing a stockbroker in the Lincoln Tunnel; the employer may act differently when the full costs of job requirements come back on the firm and its workers. Alternatively, employers of low-wage workers who really have to get to work on time may have to pay them more—in which case it is the employers and their customers who lose from congestion pricing, not workers.

The other big question is how governments—executives, legislators, voters, toll authorities—react to congestion pricing. What will a government do when it finds itself with a lot of new money and rich people much better off, but some poor people worse off? Will it use some of the money to reduce taxes for poor people or to improve their schools or add police in poor neighborhoods? Or will it do nothing at all different and save the money for a rainy day?

Clearly governments today do not totally exploit poor people—they provide some services, they do not tax them at confiscatory rates, and they even engage in some explicitly pro-poor actions (progressive income tax schedules and welfare programs, for instance). There is some reason why they do these things, and why they choose the extent to which they do these things. Those reasons are beyond the scope of this book. However, almost every plausible theory of government behavior implies that if poor people become worse off and the government gets more money and rich people become better off, then the government will use at least some of that additional money to help poor people. So it seems likely that governments as well as labor markets will

react by offsetting some of the regressive effects congestion pricing might have.

Governments today have many regressive programs that almost everybody supports: they fund libraries and subsidize state universities; they tax cigarettes and run tremendously unfair lotteries; they keep mobile homes out of swanky neighborhoods and ban motorcycles in Central Park. The ensemble of government programs and taxes is what matters, not any single particular program.

V. Other Objections to Congestion Pricing

Equity is not the only objection that has been raised to congestion pricing. In this section I consider some of the others.

It's impractical. For many years it probably was. But that was before electronic toll technology developed; every year this technology is getting better. (It's ironic that the U.S. military paid for most of the inventions that have made implementing Vickrey's ideas practical [Vickrey 1969]; he himself was a pacifist and conscientious objector.)

Today, many cities around the world are using some form of congestion pricing, or experimenting with it. Singapore has had a low-tech form of congestion pricing since 1975—to drive into the center of the city in the morning peak hour you had to display a license that cost $2.00 per trip; the price was $1.30 a trip in off-hours. In 1998 Singapore switched to automatic tolling and more finely graduated tolls. In Norway, the cities of Trondheim, Bergen, and Oslo have operated peak-hour pricing schemes since the early 1990s; originally most of the revenue was collected at old-fashioned tollbooths. The A1 motorway in France, between Paris and Lille, is operated by a private company that increases toll rates by 25 percent on Sunday afternoon. In the United States, a private company built new lanes in the middle of heavily congested State Route 91 in Orange County, California, and since 1996 has been charging a variable toll (collected electronically) for access. In San Diego and in Houston, high-occupancy vehicle (HOV) lanes—lanes reserved for cars carrying two or more passengers—have been converted to high-occupancy or toll (HOT) lanes that are open only to paying customers and HOVs during peak periods. In 2001 various weak forms of congestion pricing were instituted on several of the Hudson River crossings in New York. In 2003 London implemented a central-area congestion fee.

So far none of these projects has used the smoothly varying toll that this chapter has focused on, and so in that very important sense congestion pricing has not yet been implemented. Collecting a smoothly varying toll would not be particularly difficult with current technology, but this alone would not be sufficient. Drivers have to know what tolls they are facing in order to make

decisions about what they will do, and so some mechanism has to be in place to inform drivers early enough about what their options are. That mechanism hasn't been developed yet. But finding a way to inform drivers about a smoothly varying toll isn't a tremendously tough technological challenge; at the very least, flashing road signs (like the ones that now say "Fog Ahead") could do a substantial piece of the job.

It invades people's privacy. The government could keep track of you virtually all the time if you had to keep a transponder in your car and were tolled for all your trips. This objection derailed proposals for congestion pricing in Hong Kong a few years before it rejoined China. But congestion pricing doesn't require "Big Brother" technology anymore. People can prepay for anonymous "smart cards" that can be debited every time they use a toll road. And even without smart cards it would be pretty easy to evade a government that was using congestion pricing—all you would have to do is avoid driving a car on congested roads at congested times. You could walk, bike, take the subway anytime, anywhere, or drive at all but a few times and in a few places.

It will make untolled alternative routes worse. Drivers will clog side streets and overrun neighborhood roads as they try to avoid high tolls on the major thoroughfares. Congestion will just be transferred to streets not designed to handle it. There are several ways to respond to this objection. First, some cities, New York in particular, have only a few approaches; there are no side roads that cross the Hudson River. Second, when there are side roads that can't be tolled, substantial benefits can still be achieved from congestion tolls on the main road. Braid (1996) and Bernstein and El Sanhouri (1994) have studied how to get maximum benefit from tolls on one road when another road is a good substitute. If demand for the two roads together is perfectly inelastic and if the capacity of both roads is the same, then the optimal tolling system is the one that breaks even—it starts and ends with negative tolls and is positive at 8:30 AM. This is the toll we discussed earlier that started at ($-\$16$) at 5:50 AM, rose to \$16 at 8:30 AM, and fell back to ($-\$16$) at 9:50 AM. The more elastic total demand for the two roads is, and the greater the relative capacity of the tolled road, the higher the average toll on the toll road should be.

Roads just ought to be free. It is easy to see where this idea comes from: the beginning section of this chapter showed that it was a good idea under some conditions, and those conditions are not uncommonly met; this was especially true in the early history of automobiles. Indeed, agitation for good free roads in the United States predates the invention of the automobile: the League of American Wheelmen was organized by bicyclists in the 1880s and heavily subsidized by bicycle manufacturer Albert Pope (Gutfreund 1998, p. 10). Government-funded rural roads were part of the Populist Party platform beginning in 1885, and in the 1920s the Federal Bureau of Public Roads sponsored essay contests with topics like, "How Good Roads Help the Religious Life of My Community." The prize was a college scholarship funded by tire maker Harvey Firestone (Gutfreund 1998, pp. 11, 25).

But what's free and what has to be paid for are questions contingent on supply and demand; with relatively few exceptions they are not questions of principle. Land was free in North America before the Europeans arrived because there was an awful lot of land and very few people; it stopped being free when enough people came. Oilmen considered natural gas free and burned it off before the technology was developed to gather and use it. People today throw away (and so treat as free) used furniture and clothing that they would have preserved, repaired, and sold a century ago. What was free in America in 1910 has no bearing on what should be free in America or in India in 2010 or 2020.

It's not politically feasible. Nothing is ever politically feasible until it happens. Otherwise it would have already happened.

VI. Other Ways to Handle Congestion

Outside of economics, congestion pricing receives very little attention as a public policy to alleviate congestion. Capacity expansion, which I have discussed, receives more attention than any other policy, but several other notions have been seriously considered. This section looks at some of those other options.

Uniform tolls during rush hours. The examples of congestion pricing just discussed did not have the smoothly varying tolls whose theoretical properties we examined earlier in the chapter. Instead, they were marked by a high toll for several peak hours, and a low toll during off-hours, or no toll at all. This system—sometimes called **static congestion pricing,** as opposed to the **dynamic congestion pricing** we have concentrated on—is easier to administer, and may be somewhat more popular with drivers because it gives them some assurance about what they will be charged for each trip (although it's not clear how important such assurances are, since most people have to worry about risks that are a lot bigger than this, and since they have no assurance about the size of the traffic jams they will encounter).

How good is static congestion pricing at promoting efficiency? That depends on the elasticity of demand for peak-hour travel. The more elastic the demand, the better the job that static congestion pricing does. Static congestion pricing operates by reducing the number of trips, not arranging them better. Reducing the number of trips is a good idea when people are getting in each others' way—as they are without reservations or dynamic congestion pricing. Dynamic congestion pricing eliminates the root problem and so always does at least as well as static does, but static congestion pricing is better than nothing.

To see why, recall Figure 4.7, reproduced here as Figure 4.10a. In a come-as-you-please situation without tolls, triangle *C* represents the schedule delay costs drivers bear, triangle *B* represents the waiting time costs they bear, and
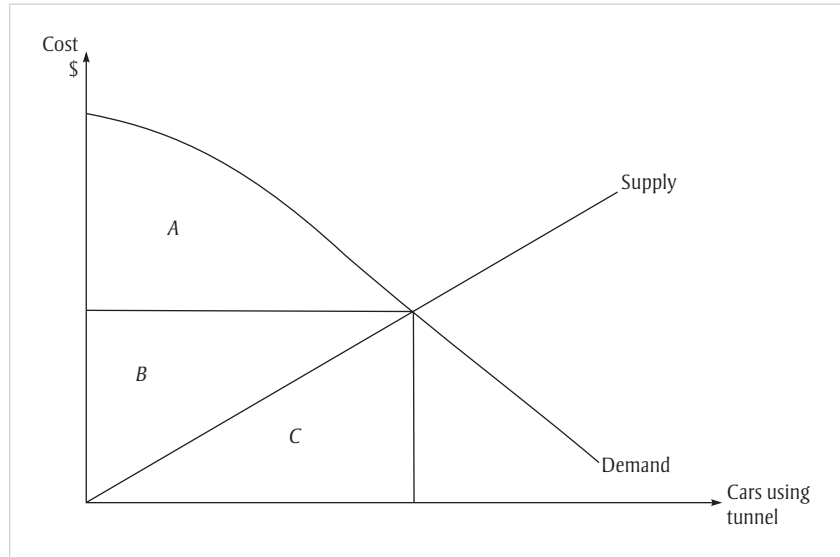
area *A* represents net benefits to drivers. Dynamic congestion pricing operates by transforming triangle *B* from the waste of waiting time into potentially productive revenue in the hands of the toll authority. Static congestion pricing converts some of triangle *B* into toll authority revenue, but to do this it must discourage some trips that would otherwise have net benefits, and it must also transfer some benefits from drivers who are not dissuaded.

Figure 4.10b illustrates how this happens. The uniform toll raises the supply curve—for any number of cars, it costs more to use the road—and so supply and demand intersect with fewer and more costly trips. Because there are fewer trips, drivers lose net benefits in triangle *L*. The toll authority gains revenue equal to the sum of rectangles *T* and *G,* but rectangle *T* comes purely at the expense of drivers; only rectangle *G* represents the type of social gain that can count for a potential Pareto improvement. Since the gain in rectangle *G* is bigger than the loss in triangle *L,* imposing static congestion pricing in this case is a potential Pareto improvement. But since rectangle *G* is smaller than triangle *B* in Figure 4.10a, and dynamic congestion pricing doesn't cause any losses like triangle *L,* dynamic congestion pricing is better.

Of course, if the problem you're concerned about is pollution or road maintenance rather than congestion, then the reduction in the number of trips is an advantage, not a disadvantage. But in that case tolls should be high all the time, not just in peak hours, and dynamic congestion pricing to allevi-ate congestion could be implemented in addition to a uniform time-invariant toll, to alleviate pollution.

Gasoline taxes and car taxes. Gasoline taxes and car taxes are essentially remedies for pollution and road deterioration. They also address in part the problem of free parking. But since gas is gas, no matter when and where it's
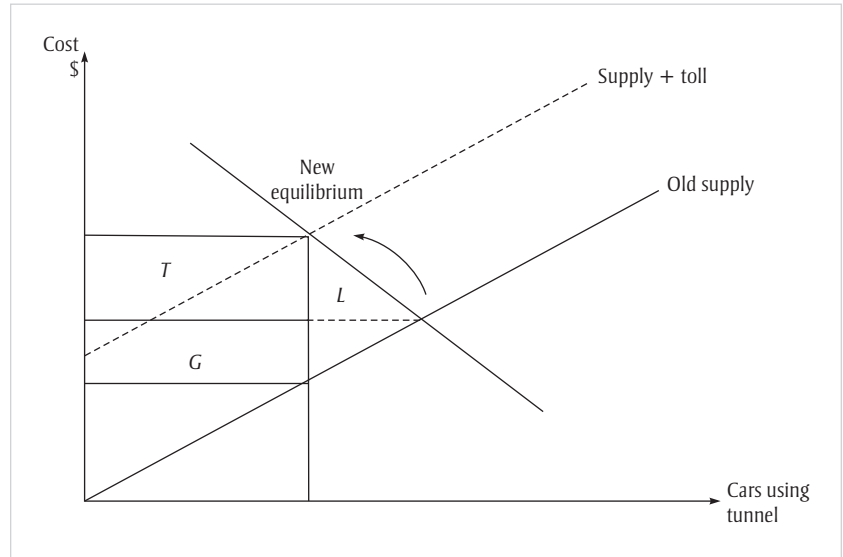
used, and cars are cars, gasoline and car taxes are very blunt, poor instruments
for dealing with congestion. Taxing gas to reduce congestion is like taxing pa-
per to reduce pornography.

   Parking policies. Hitting a stationary target is easier than hitting a moving
one, and so it is probably easier to collect fees from cars when they're parked
than when they're moving. An alternative way to discourage peak-hour com-
muting, therefore, is to charge parking fees for cars that park in central loca-
tions at peak times, say between 6 AM and 10 AM.

   Such parking fees would be very similar to static congestion pricing,
since they would not be able to eliminate congestion; they could only reduce
trips. Even relative to static congestion pricing, parking fees have disadvan-
tages, too. They would not affect travel at the most congested times and in the
most congested places by drivers not intending to park at the destination most
people are going to—long-distance truckers, for instance, people running er-
rands. To some extent, parking fees would merely substitute less-valued trips
by such sojourners for more-valued trips by commuters, for no net reduction
in pollution or congestion.

   The U.S. tax treatment of the free parking that many employers provide
their workers is a different matter. For many workers, this is a major fringe
benefit; 95 percent of commuters who drive to work get free parking (Shoup
1997). As with other fringe benefits, employers deduct as a business expense
the costs they incur in providing free parking—construction, cleaning, secu-
rity—and employees don't have to pay taxes on the benefit they receive. This
encourages people to drive to work—and to drive to work too much. An em-
ployer who gives a worker in the 30-percent income-tax bracket a dollar raise
in pay allows that worker to enjoy only seventy cents' worth of more yogurt;

an employer who devotes that same dollar to free parking allows the worker to enjoy a full dollars' worth of parking. So workers end up consuming too much parking and too little yogurt: everyone could be made better off if workers got the cash value of the free parking they received and could decide on their own how much they wanted to spend on yogurt, without any tax consequences or government interference.

When California in 1992 required some employers to offer workers a choice between free parking and cash, the number of solo drivers fell by 13 percent, and the number taking mass transit rose 50 percent. Women and minorities gained disproportionately from the parking cash-out (Shoup 1997). A cash-out, however, doesn't remove all the distortions that the income tax code causes, because those who decide to take the cash pay income taxes on it, but those who decide to take the parking space don't.

Longer workdays. If people worked, say, four days a week for ten hours a day instead of five days a week for eight hours a day, there would be only eight rush hours a week instead of ten. Each rush hour would be no better than it is now but no worse either, and so the total weekly cost of commuting would fall by 20 percent. Three-day weekends might also reduce congestion at malls and beaches.

If the four-day week is such a good thing, why haven't employers adopted it on their own? Perhaps fatigue and boredom are such big problems that the productivity losses from ten-hour workdays would offset the gains from less commuting. If that's why employers have chosen not to adopt the four-day week, then government efforts to move in that direction would probably be misguided.

But there's an alternative explanation for why employers embrace the five-day week: it is a convention. In many types of jobs, people are most productive when they can interact with people at work in other firms, and many people most like being off work at the times that other people—friends and family members, especially—are also off work. So while changing to a four-day week would be disastrous for an individual firm to do by itself, it might be no problem at all if all firms (or almost all firms) did so together. Coordination itself might be valuable, not any particular convention.

More important, employers don't now bear the full costs of their decision to have workers make ten trips a week instead of eight because they don't bear the cost of the congestion that their workers impose on other drivers.

VII. Conclusion

We could not maintain our health, our productivity, our sense of freedom and psychological adjustment without automobiles. Sometimes they're fun, too. In some circumstances, the government should just build roads and let motorists loose on them.

But cars also cause problems. Road damage, pollution, accidents, and congestion are among them. Numerous policies have been designed to address these problems—gas taxes for road damage; emission standards for pollution; tort law and car insurance for accidents; and a wide variety of ideas for congestion. Economists have devised a number of programs, such as double-strict liability, pay-at-the-pump car insurance, and dynamic congestion pricing, that could take advantage of advances in information technology. They might work better than existing approaches.

The policies in place in most places probably don't make drivers fully realize the costs in road damage, pollution, and accidents that they cause. Reducing the amount of driving and the concomitant levels of these problems would probably be a potential Pareto improvement. Greater driver caution would also be a potential Pareto improvement. Better congestion policies would save resources and make drivers (or somebody) better off.

Why should you care about reducing pollution and accidents and congestion while not losing worthwhile trips? The direct costs—tens of billions of dollars in the United States alone, probably much more in Europe and Japan, and rapidly growing amounts in developing countries—are enough by themselves. But there is a more subtle reason, too: to take advantage of city life and productivity. Cities can't realize the increasing returns to scale discussed in Chapter 2 unless people can get around in them easily and arrive at the right places at close to the right times. People stuck in traffic are a burden not only to themselves but also to those they were planning to see. The more costly it is for me to get through the Lincoln Tunnel, the less likely it is that I will visit the marine chart store on Manhattan's East Side or play in the Rockefeller Center tuba concert or enter the Joe Kleinerman twelve-hour race. If I don't visit the store or play in the concert or run in the race, I make it more likely that the store will close, the concert will fold, or the race will be cancelled. Then all the people who wanted to visit the store, attend the concert, run the race will be losers, too. The more valuable the outputs of city life, the greater the benefits that flow to everyone from intelligent public policies on urban transportation.

Appendix: The Numerical Example of Rich and Poor Drivers

Suppose demand is perfectly inelastic, with 2,000 poor people and 2,000 rich people wanting to drive through the Lincoln Tunnel in the morning. Per-minute costs, in cents, are

	Poor	Rich
Waiting	22.5	60
Early	20	40
Late	40	80

Find the come-as-you-please equilibrium and the optimal congestion pricing equilibrium. How do costs change, and how much revenue does the toll authority get?

A. Come-as-You-Please Equilibrium

Step 1: Who uses the tunnel at the more convenient times?
Think about before 8:30 AM first. A rich person is willing to wait an extra two-thirds of a minute to get entry to the tunnel a minute later ($0.40/$0.60), but a poor person is willing to wait eight-ninths of a minute more in order to get in the tunnel a minute later ($0.20/$0.225). So it will be the poor people who get in the tunnel around 8:30. Similar calculations yield the same conclusion for time after 8:30.

Step 2: Find total costs for rich people. The earliest rich person arrives at 5:50 and encounters no line. Her costs are all schedule delay costs, and they amount to $64. So every rich person must bear a cost of $64.

Step 3: What times are the dividing lines between poor and rich drivers? I claim that the schedule delay costs borne by the earliest poor driver have to be the same as the schedule delay costs borne by the latest poor driver.

To see why, suppose the schedule delay costs of the earliest poor driver—call her driver poor E—are less than those of the latest—call her driver poor L. Call the difference in schedule delay costs S. Since schedule delay costs per minute of rich people are twice those of poor, this would mean that the schedule delay costs of the rich person right before driver E—call her E's predecessor—are less than those of the rich person right after poor L—call her L's successor. The difference is $2S$. Since rich people all have the same total costs, the waiting costs of E's predecessor must be greater than the waiting costs of L's successor by the amount $2S$. That means the line is longer when poor E starts waiting than it is when poor L starts waiting, and poor E must wait longer than poor L. Since per-minute waiting costs of poor people are 37.5 percent ($= 22.5/60$) as great as those of rich people, the waiting costs of poor E must be 37.5 percent as large as the waiting costs of her predecessor, and the waiting costs of poor L must be 37.5 percent of the waiting costs of her successor. So the difference between poor E's waiting costs and poor L's waiting costs must be 37.5 percent of the difference in waiting costs between poor E's predecessor and poor L's successor—which is $2S$. Thus poor E's waiting costs are greater than poor L's by the amount

$$.375 \times 2S = .75S.$$

Thus, so long as $S > 0$, poor E is better off than poor L. Her schedule delay costs are lower by S and her waiting costs are higher by only $.75S$. So poor L would want to switch her time of arrival to about the same time as poor E.

Similarly, we can show that at equilibrium the earliest poor driver cannot

have schedule delay costs greater than those of the latest poor arrival; the earliest and latest poor drivers have the same schedule delay costs. That means that two-thirds of poor drivers arrive before 8:30 and one-third arrive after. So the dividing lines are 7:10 and 9:10. Half of the time before 8:30 is for each class; so is half of the time after 8:30.

Step 4: What are the waiting times of rich people? Schedule delay cost decreases at the rate of $0.40 per minute between 5:50 and 7:10, and so waiting cost must increase at the rate of $0.40 per minute. Thus waiting time must grow at the rate of two-thirds ($= 40/60$) of a minute per minute. The rich person who arrives at 7:10 (poor E's predecessor) has a schedule delay cost of $32 and a waiting time cost of $32. So does the rich person who arrives at 9:10 (poor L's successor). The wait at 7:10 and at 9:10 must be 53.33 minutes.

Step 5: What are the waiting times of poor people? The poor person who arrives at the tunnel at 7:30 (poor L) also waited 53.33 minutes. Since her schedule delay costs are $16, her total costs are $28,

$$16 + 53.33 \times .225 = 28,$$

and so every poor person must bear $28 in costs. The poor person who arrives at 8:30 has no schedule delay costs and so waits in line for 124.4 minutes ($28/.225$).

Thus in the come-as-you-please equilibrium, the cost for every poor person is $28 and the cost for every rich person is $64.

B. Optimal Congestion Pricing Equilibrium

Step 1: Who uses the tunnel at the more convenient times? Since rich people have higher schedule delay costs per minute (both early and late), they use the tunnel at the more convenient times.

Step 2: Find the total costs for poor people. The earliest poor person arrives at 5:50 AM and encounters no toll or line. Her costs are all schedule delay costs, and they amount to $32. So every poor person must bear a cost of $32.

Step 3: What times are the dividing lines between poor and rich drivers? Suppose the latest rich driver—call her rich L—had greater schedule delay costs than the earliest rich driver—call her rich E. Then total schedule delay costs for rich people could be reduced by moving rich L to before 8:30 AM. This would also entail moving rich E's predecessor in the opposite direction, from before 8:30 to after 8:30. The move would increase schedule delay costs for poor drivers, but the loss to rich E's predecessor would be approximately half the gain to rich L, since rich people's per-minute costs are higher. So exchanging places between rich L and rich E's predecessor would be a potential Pareto improvement.

Therefore with optimal congestion pricing the schedule delay cost for rich E has to equal the schedule delay cost for rich L. This implies, as usual,

that two-thirds of rich drivers arrive before 8:30, and the same for poor drivers. So the dividing lines are the same: 7:10 and 9:10.

Step 4: What are the tolls that poor people pay? Between 5:50 and 7:10, schedule delay cost falls at the rate of $0.20 a minute, and so tolls must rise at $0.20 a minute. The opposite happens between 9:10 and 9:50, but at a rate of $0.40 per minute. At 7:10 and 9:10 the toll is $16. The average toll that poor people pay is half that amount, and so the tolls that poor people pay amount to $16,000 (= 2,000 people × $8 average each).

Step 5: What are the costs rich people bear? Consider rich *E,* the earliest rich arrival, at 7:10. If she pays more than the $16 a poor person coming then would pay, she would want to arrive later, at 7:29, and people would not be distributed efficiently. If she paid less, poor people would want to come later, and once again people would distribute themselves inefficiently. So the toll a rich person who arrives at 7:10 has to pay is $16. In addition, schedule delay costs at that time amount to $32 for a rich person, and so the total cost for this rich person amounts to $48. So every rich person bears a cost of $48.

Step 6: What are the tolls rich people pay? To keep rich people indifferent about when they come, the toll has to rise at a rate of $0.40 a minute from 7:10 to 8:30, and so the maximum toll is $48. Then it falls at the rate of $0.80 a minute until it reaches $16 again at 9:10, when the poor people take over again. The average toll is halfway between the lowest ($16) and the highest ($48); and so it is $32. The tolls that rich people pay equal $64,000. Total toll revenue from all drivers is $80,000.

≡

Thus converting from come-as-you-please to optimal congestion tolls reduces the cost of travel for rich people from $64 to $48 and increases the cost for poor people from $28 to $32. Poor and rich people are equally numerous, and the gain by the rich is four times as great as the loss to the poor. In addition, the toll authority gains $80,000, or $20 per capita, which could be rebated to either group or both. If poor people, half the population, received only 15 percent of the toll revenue, they would be as well off as they were under come-as-you-please.

Questions

1. Anthony Downs concludes his book on peak-hour congestion by writing: "Therefore, my advice to American drivers stuck in peak-hour traffic is not merely to get politically involved, but also to learn to enjoy congestion. Get a comfortable, air-conditioned car with a stereo radio, a tape player, a telephone, and perhaps a fax machine. . . Then regard the moments stuck in traffic simply [as] an addition to leisure time" (Downs 1992, p. 164).

   Suppose people followed Downs's advice. This would be expensive for them, of course, but it would reduce the per-minute cost of waiting time. How would the length of the equilibrium queue change if people followed Downs's advice? How would the total cost of the time spent waiting change? In the aggregate, would all of the investments be worthwhile? Suppose everyone but you had made these investments. Should you make them too?

2. Wasting gas is part of the cost of waiting time. Suppose the price of gas goes up. Assume all drivers are identical and face a bottleneck. Demand is perfectly inelastic.

   a. In the come-as-you-please situation, how does the arrival time of the earliest driver change? How does the waiting cost of the driver who arrives at the most convenient time change? How does that driver's waiting time change? How does the physical length of the lines change? How much are drivers hurt by the increase in the price of gas?

   b. In the situation with optimal congestion prices, what changes as a result of the higher price of gas? How much are drivers hurt by the increase in the price of gas?

3. In 1998 New York City erected barriers on Fifth Avenue to keep pedestrians from crossing on one side at 49th and 50th Streets. The object was to make it easier for cars to make turns from these streets onto Fifth Avenue and thus to speed cross-town traffic flow. Manhattan has many cross streets parallel to 49th and 50th Streets, and vehicular traffic is usually very congested for many miles around these corners. People also have many alternative means of travel in Manhattan (foot, bus, subway, cab). Pedestrians and pedestrian-advocacy groups protested loudly about the barriers, but the mayor maintained that speeding vehicular traffic in midtown Manhattan was an important public-policy objective. Analyze this initiative to determine its impact on vehicular speed and congestion in midtown Manhattan.

4. In our standard tunnel example, suppose that everyone has the same time costs but that 2,000 drivers have regular cars and 2,000 have sport utility vehicles. SUVs are bigger than regular cars; in an hour, either 1,000 cars or 750 SUVs can go through the tunnel.

   a. What happens in the come-as-you-please scenario?

b. What happens with optimal tolls? (Assume that the tolling technology can recognize an SUV and, if required, charge a different toll.)

c. Suppose that SUV drivers are richer than drivers of regular cars and so have different time costs. Under what circumstances would this affect your answer to question b?

5. In the text, I compared the gains that rich people might realize from congestion pricing with the gains or losses that poor people might realize. Would congestion pricing affect men and women differently? If so, how?

6. Instead of just two different groups (rich and poor), suppose that there are hundreds of different groups, each with different per-minute costs of schedule delay. Sketch the optimal toll schedule.

7. Suppose everyone has a car with a good heater but only rich people's cars have air conditioners that work. Summers are very hot and winters are very cold (there is no fall and no spring). Schedule delay costs are higher for rich people than for poor and do not vary with the season. There are regular traffic jams at a bottleneck where there is no toll. The group of drivers who pass through the bottleneck at the most convenient times is different in the summer than it is in the winter. If congestion pricing is implemented, none of the revenue will be used to benefit poor people. You want to make sure that congestion pricing doesn't harm poor people. Can you implement congestion pricing in either season or in both with this assurance? Explain.

References

Anderson, David Richard. 1996. "Congestion Pricing on a Transportation Network: Efficiency Gains and Distributional Consequences." PhD diss., University of Minnesota.

Arnott, Richard J., Andre de Palma, and Robin Lindsey. 1994. "The Welfare Effects of Congestion Tolls with Heterogeneous Commuters." *Journal of Transport Economics and Policy:* 139–161.

Bernstein, D., and I. El Sanhouri. 1994. "Congestion Pricing with an Untolled Alternative." Unpublished manuscript.

Braid, Ralph M. 1996. "Peak-Load Pricing of a Transportation Route with an Unpriced Substitute." *Journal of Urban Economics* 40:179–197.

Caro, Robert. 1975. *The Power Broker: Robert Moses and the Fall of New York.* New York: Vintage Books.

Daniel, Joseph I., and Khalid Bekka. 2000. "The Environmental Impact of Congestion Pricing." *Journal of Urban Economics* 47(2):180–215.

Downs, Anthony. 1992. *Stuck in Traffic: Coping with Peak-Hour Traffic Congestion.* Washington, DC, and Cambridge, MA: The Brookings Institution and the Lincoln Institute of Land Policy.

———. 1962. "The Law of Peak-Hour Expressway Congestion." *Traffic Quarterly* 16:393–409.

Gutfreund, Owen D. 1998. "20th-Century Sprawl: Accommodating the Automobile

and the Decentralization of the United States." PhD diss., Columbia University, Department of History.

New York Metropolitan Transportation Council. 1995. *Hub-bound Travel.*

Segal, David, and Thomas L. Steinmeier. 1980. "The Incidence of Congestion and Congestion Tolls." *Journal of Urban Economics* 7(1):42-62.

Shoup, Donald C. 1997. "Evaluating the Effects of Cashing Out Employer-Paid Parking: Eight Case Studies." *Transport Policy* 4(4):201–216.

Small, Kenneth A. 1983. "The Incidence of Congestion Tolls on Urban Highways." *Journal of Urban Economics* 13:90–111.

Texas Transportation Institute. 2004. "2004 Urban Mobility Study." Accessed at *www.mobility.tamu.edu/ums/report/congestion_cost.pdf.*

U.S. Bureau of the Census. 2002. *Statistical Abstract of the United States 2001.* Washington, DC: U.S. Government Printing Office.

———. 2004. *Statistical Abstract of the United States 2003.* Washington, DC: U.S. Government Printing Office.

Vickrey, William S. 1969. "Congestion Theory and Transport Investment." *American Economic Review,* Papers and Proceedings, 251–261.

Mass Transit

There's something ridiculous about commuting by car. You use two tons or more of metal, rubber, and plastic to move a couple hundred pounds of human. Most of the energy and effort goes into moving the car, not the human; the car takes up road space and causes congestion; and once you get wherever you're going, you have to figure out what to do with the tons of vehicle that got you there.

Aside from walking, buses and trains are the major alternatives to everyday car travel in the United States. Like cars, buses and trains weigh a lot more than the people they transport, and they take up even more space. But the material-to-flesh ratio is a lot lower: a standard New York City bus weighs 36,800 pounds, about six to twelve times as much as a passenger vehicle, but it can carry more than forty people.

The reason for this savings is the two-thirds rule. Cars, buses, and trains are essentially containers for people. The bigger the container, the greater the ratio of volume to surface area, and the more people you can transport (that's what uses volume) for every pound of weight (that's what surface area produces) and every gallon of gas (that's what moves the weight). There are economies of scale in moving people (and goods).

Mass transit also reduces accidents. People on a bus can't collide with each other, and a bus, being one container, has less chance of hitting another vehicle than do forty separate cars on the road.

It's not surprising, then, that many people think mass transit is a good way to reduce expense, pollution, and congestion. Despite these advantages, however, few American commuters use mass transit, which accounted for only 4.9 percent of work trips in the U.S. metropolitan areas in 1999 (U.S. Bureau

---

To get the most from this chapter, you should be familiar with these concepts: antitrust, consumer surplus, deadweight loss, elasticity, inefficient taxation, monopoly, second-best, and sunk costs. You can find these terms in the Glossary. The concepts of congestion, schedule delay costs, and the two-thirds rule from previous chapters are also used.

of Transportation Statistics 2002, table 1-33). The preponderance of public transit trips are made in a few cities, especially New York, which alone accounted for a third of U.S. public transit riders in 1999. About twice as many people carpooled as used mass transit, and almost as many walked to work. In Walnut Creek, California, fewer than 4 percent of the people working in office buildings right next to Bay Area Rapid Transit (BART) stations commuted on BART. Mass transit was used in an even smaller share of nonwork trips.

Overall in the United States, the number of passenger-miles traveled in personal vehicles is about twenty-five times greater than the number of passenger-miles traveled in buses and on transit rail systems (this excludes intercity rail systems). In other countries, the picture is different. Personal vehicle passenger-miles are only three times greater than mass transit passenger-miles in Japan, ten times greater in Germany and Italy, and fourteen times greater in France and the United Kingdom (U.S. Bureau of Transportation 1999, table 5).

Shouldn't everyone use mass transit? Would developing countries do better to follow the U.S. model, or should they use Japan as their guide? This chapter will address such questions by thinking about how mass transit systems should operate. The first part of the chapter will examine the characteristics and problems of good transit systems, no matter who operates them. The second part will look at who operates transit systems, and what the advantages are of different kinds of operation and ownership.

I. Designing Good Mass Transit

A. Scheduling Trains and Buses

People differ, and their time is valuable. The arguments I have presented for mass transit have ignored these facts, but if you want to think about good schedules for trains and buses, you have to pay them a great deal of attention. And if you want to think about how mass transit should operate, you have to think about schedules. Once you start thinking about schedules, you can begin to see why Americans avoid mass transit. You can also see the crucial role of density.

Here's the basic idea: look at the people around you right now, or think about all the people you've met today. The odds are low (unless you're in prison or the military) that more than two or three of them were dressed the same way; perhaps every single person was dressed differently. Almost everybody wears different clothing, even though it would cost less—in terms of design effort and production—if everybody wore a uniform, and clothing would be even cheaper to produce if all uniforms were made the same size. They wouldn't fit too well, and most people wouldn't like them, but they would be serviceable, and the standardization would save a lot of money.

Americans don't go around wearing uniforms, because clothing that doesn't fit well is also costly, and so is wearing something you think is ugly. People come in different sizes and have different ideas about what's boring or interesting. They do different things, too—they fight fires, perform striptease acts, play tennis, slaughter animals, entertain the British ambassador—and different kinds of clothes go best with each of these different activities. The great variety of clothing that Americans wear is both a cause and an effect of living in a prosperous modern economy. People produce more because they wear appropriate clothing for the tasks at hand, and they use their enhanced income to consume things closer to what they want, even though they cost more.

In the same way, demand for variety makes mass transit scheduling difficult and mass transit itself unpopular. Just as you are unlikely to find an E-width ballet shoe in a discount shoe store, you are unlikely to find a non-stop bus that travels from Salter Place in Maplewood, New Jersey, to the VFW post in Carmel, New York, at six o'clock on a Saturday morning. The "mass" in "mass transit" means something—namely, that unless you are making a trip that a lot of other people are making at the same time, it's unlikely to help you much.

We can think of trips as differing along three dimensions: place of origin, place of destination, and time of arrival (just as shoes differ along such dimensions as size, width, mode of fastening, color, and style). Cars can often fashion trips that almost perfectly match what people want in all three dimensions, but at a price. Mass transit produces cheaper rides only because people are forced to compromise on one or more dimensions. If the bus stop is closer to my house, it's farther from yours, and in any event it's not likely to be as close as our garages. The train can come too late for me or too early for you; the bus terminal is not perfectly located for either of our jobs. The goal in scheduling mass transit is to force people to make as few bad compromises as possible.

Think about a situation in which people differ only in their desired times of arrival; there are a lot of people going from the same origin to the same destination, but they want to get there at different times. If a lot of people all want to arrive at the destination at the *same* time, then scheduling is either no problem—you put them all on the same train and schedule it to arrive at the desired moment—or a simple problem of the type we have already studied: if congestion makes it physically impossible for everybody to arrive at the same time, you pack trains or buses as tightly as possible around the desired moment. When everybody wants to arrive at the same time, the advantage of mass transit over cars is clear.

When people want to arrive at different times, somebody has to be disappointed.

Notice the difference with car congestion, which we examined in Chapter

4: there the inevitable disappointments arose because everybody wanted to come at the same time.

If the trains operate according to a published schedule that everybody believes, nobody waits. They show up when the train they want is due to depart and get on. The only time costs (aside from traveling, which is always the same no matter how trains are scheduled, and so can be ignored) are schedule delay costs. Commuters decide which train to catch by comparing schedule delay costs. They will take the early train (the train that gets them in before their desired time) if and only if the cost of being early that it imposes on them is less than the cost of being late that the later train would impose on them.

The more closely together trains are scheduled, the smaller the average schedule delay costs for commuters. But if trains are scheduled more closely together, they will each have fewer riders and thus will realize fewer of the economies of scale that make mass transit attractive in the first place. This is the fundamental trade-off involved in scheduling: mass transit can either take maximum advantage of economies of scale and impose lots of schedule delay costs, or it can minimize schedule delays and look little different from a taxi service.

The same trade-off occurs in the other dimensions of transit service. Suppose that people start their trips from a bunch of different origins. If there is only one train and it picks up passengers at only one spot, then it will be a very full train and realize lots of economies of scale, but commuters will have to bear on average very large costs to get to the station. If there are many different trains starting at many different stations, then commuters have short trips to the station but trains carry few passengers, and many trains have to run. And if there are only a few train lines but the trains make lots of stops, either the trains will be mainly empty for most of their runs; or commuters on board will have long waits while the trains slow down, stop, and allow new passengers to board; or both (if most of the passengers board early in the run, the boarding delays will be small but many people will experience them; in the opposite case, the delays will be large but few people will experience them).

The same sort of problems present themselves when people have different destinations, and when people differ—as they do—in all three dimensions, the problems compound. If only a few trains run, the average person finds that even the best train goes from an inconvenient origin at the wrong time to an inconvenient destination; if a lot of trains run, mass transit isn't "mass" any more.

These considerations tell us the circumstances under which mass transit is likely to be successful. Mass transit works well when riders are poor (so that schedule delay costs are small relative to savings in gasoline, vehicle operation, and so forth), and when lots of people are going from about the same origin to about the same destination at about the same time. With twice as many

people, you can run an express from each of two stations instead of a single local. If people live closer together, more can reach the station in a ten-minute walk, and if everyone works in about the same place, the average person's job can be very convenient to the train terminal.

Mass transit, then, is like marine chart stores and tuba concerts: it's something that only dense agglomerations of population can support. The availability of mass transit is another of the advantages of city life that I discussed in Chapter 2: only in dense cities can people realize the savings in gasoline costs, pollution, vehicle weight, and road maintenance that mass transit brings. Elsewhere the cost of assembling enough people to reasonably fill a bus or train exceeds the savings in direct transportation costs.

How dense is "dense" when it comes to supporting mass transit? No single answer is going to be right for all times and all places. The answer depends in part on the value of people's time, which may be different in Lagos than it is in New York, and different in Brownsville, New York, than it is in Manhattan's Upper East Side. The answer depends also on what the alternative modes of transportation are like—how polluting cars are, how expensive gasoline is, how often transmissions break down, how congested roadways are. Finally, the answer depends on the technology of mass transit: how expensive it is to build and operate a system, how many people it can carry, how reliably it can keep to a schedule.

The most comprehensive study of this question was made by economists at the University of California, Berkeley, in the early 1970s (Keeler et al. 1975). They were trying to determine whether BART, the new subway in the San Francisco Bay Area, made sense. They compared three modes of transportation: private car, bus (meaning standard city buses with about forty seats each), and rail (meaning four- to eight-car trains or subways with several hundred seats each).

The Berkeley researchers tried to estimate all costs, both private—like gasoline, accidents, schedule delay costs, operating costs—and social—like pollution, uncompensated accidents, congestion, and road damage. Since they were concerned with systems not yet built, they considered both the cost of building systems and the costs that would be incurred after they were built. The work they did is respected, and no comprehensive update has been done (Romilly 1999 updates the comparison between buses and cars for the United Kingdom).

They found that for a trip that fewer than about 1,100 people were making in an hour, cars were cheaper than any form of mass transit. Even when you count the cost of pollution, the schedule delay costs involved in getting enough people together to take a bus or train were so great that cars did better—replacing a bus with cars would be a potential Pareto improvement. Using buses was cheaper than cars if at least 1,100 people an hour were making a trip. Rail transportation started getting cheaper than cars at around

22,000 trips an hour, and approached the cost of buses at around 50,000 trips an hour. Rail was probably the cheapest mode in commuting corridors that saw 50,000 trips an hour or more—if you were starting from scratch. If a rail system had already been built, continuing to operate it was probably cheaper than buses at considerably lower densities—perhaps 30,000 trips an hour or so. Since operating a bus costs less than operating a train, and because a bus fills up with fewer people, lower densities can support buses than can support trains (and even lower densities can support smaller buses).

The Berkeley study predates the policy popularity of "light rail"—trolleys and streetcars—and so it didn't try to estimate their actual cost at various trip densities. A study by Gómez-Ibáñez (1985), however, does examine the economics of light rail. He found no obvious advantage for light rail over buses, especially with the trip densities at which buses are the cheapest transportation mode. Light rail puts more riders in a single vehicle and so saves money on operators' salaries, but this saving is offset by greater schedule delay costs for riders and greater capital costs for building the rails and power systems that trolleys need.

How do mass transit densities studied compare with actual densities in the United States today? Only Manhattan experiences densities of 40,000 trips an hour or more. Older big cities—Chicago, Boston, and Philadelphia—have a few corridors that see 10,000 to 30,000 people pass through them in an hour. The most heavily traveled links of BART in San Francisco carry about 15,000 people an hour—enough for a good bus system, but not for a subway. Other than these, there are almost no corridors with densities of more than 10,000 trips an hour.

The Berkeley study led economists to believe that constructing new subways in the United States would be foolish, and when the federal government nevertheless subsidized new subway and light rail systems during the late twentieth century, the consensus among economists turned out to be right (this doesn't always happen). The new systems almost always cost more than their proponents estimated and drew fewer riders.

Atlanta's Metropolitan Atlanta Rapid Transit Authority (MARTA) subway, for instance, is widely hailed as the great success story of this generation of subway building, but when Kain (1997) analyzed it carefully, he found that if the same amount of money had been spent for buses, lowering fares and raising frequency, twice as many trips would have been made on public transit. (Kain probably overstates the cheapness of buses; he doesn't consider road damage or particulate pollution from diesel engines.) When construction and all operating expenses were included, Kain figured each trip on MARTA cost about $8.

What's wrong with this? The trips are almost certainly not worth $8 on average. Few of the riders would have been willing to pay $8 for each trip, considering that they often have good alternatives, such as driving, carpooling, bi-

cycling, taking the bus, and walking. Nor do the external benefits of these rides plausibly approach anything like $8. Absent congestion, a ten-mile auto trip generates external costs of only one or two dollars, and only a fraction of MARTA's trips were substitutes for auto trips. Congestion pricing is a much cheaper way of dealing with congestion. MARTA's rides also generate external costs themselves. If MARTA had not been built and the money had been given to the prospective beneficiaries instead, everyone would be better off.

Such transit improvements haven't helped poor people either. Studying several of the new subway and light rail systems, Baum-Snow and Kahn (1998) found that the systems have disproportionately served white people and benefited landowners. They found no evidence of disproportionate service for poor people.

Remember, though, that these findings about subways apply to particular U.S. cities at particular times. They don't necessarily hold for cities in developing countries, which are considerably larger and where schedule delays are less expensive. Nor would they necessarily hold in the same American cities if the cities had been built differently, with houses closer together and jobs more concentrated in the central business districts. The transportation system that was available at the time had a big influence on the way these big cities were built, and so to say that subways are not a good idea now is not to imply that they were not a good idea fifty years ago. It's silly to start watching *Sesame Street* when you're fifty years old, but it may be very helpful to your development if you do so when you're three. This is an issue I will tackle in Chapters 6 and 7.

B. Gregariousness

Thus far this discussion of scheduling has presumed that the people running the transit system could control everything that went on. If they decided a train should arrive at 96th Street at 8:02 AM, a train would arrive at 96th Street at 8:02 AM. Actual transit systems, of course, are plagued by mishaps that keep trains from running precisely on schedule: trains break down, passengers get sick, leaves or snow impede progress, drawbridges open, passengers board too slowly, fires start, water mains break. Buses face additional problems of traffic congestion and red lights. The schedule may say a train or bus will arrive at 96th Street at 8:02 AM, but you can't really count on it.

As a result, in some transit systems riders don't pay any attention to schedules. This is likely to be the case when the wait between trains or buses is short and the probability that a train or bus will be exactly on schedule is small. Elevator systems don't work on schedules, for example, and while the New York subways theoretically have schedules, nobody knows what they are during peak hours. You show up and expect a train soon.

Under these circumstances random delays can cause problems for buses and trains not directly affected by the delay. Suppose a bus heading downtown loses a minute because a passenger asks for directions at 110th Street. When it gets to the next stop, at 109th Street, more passengers than usual will be waiting to board because some arrived at the bus stop during the delay, and so the passengers at 109th Street will take more time to board than usual. The bus will pull away from 109th Street more than a minute late and arrive at 108th Street more than a minute late and so find even more passengers than it found at 109th Street. The bus will become later and later as it encounters more and more passengers at each subsequent stop.

Meanwhile, the bus that was supposed to be following the bus that encountered the original delay is moving faster and faster. Suppose it was originally ten minutes behind the lead bus. At 109th Street it will be only nine minutes behind and so encounter fewer passengers than it normally would; the passengers who came nine minutes before were already picked up by the lead bus. So the following bus takes less time to board passengers at 109th Street and gets further ahead of schedule. At 108th Street, this bus finds even fewer passengers, and so on.

At every stop, the lead bus gets slower and the following bus gets faster, and in a little while they converge and start running together. Vickrey (1987) called this phenomenon **gregariousness**—the tendency, well known to everyone who has ever spent much time bus riding, for buses to travel in herds (especially on cold and wet days). The question that gregariousness raises is whether the following bus should slow down, and if so, how much. Vickrey's analysis is a good example of how economics can help with practical problems.

Consider the trailing bus at 110th Street. Slowing down will delay all the passengers now on board. It will also delay all the passengers further down the route who would be picked up if this bus stayed nine minutes behind the lead bus. The delay for these people, both on the bus and downstream, is the cost of slowing.

On the other hand, the delay will allow this bus to pick up more people at each future stop and save these people the wait for the next bus. That's the benefit of delay. If the next (third) bus is now about eleven minutes behind the bus we're thinking about slowing down, the slowdown will save these people about ten and a half minutes each. These people may not be very numerous—they're people who arrive at the bus stop in the span of a minute or so—but the gain for each of them is big.

If enough people arrive at future stops in this small window of time, then it's quite possible for the benefits of slowing down to exceed the costs. Notice that this means the second bus will be behind schedule too. If you evaluate a transit system by its on-time percentage, this system will look worse than a

system that doesn't delay the bus, even though it treats its customers better. So one lesson from this example is that on-time percentage is not a foolproof way to judge how well a transit system operates.

The second lesson from this example is that no simple hard-and-fast rule guarantees the right answer. We have just seen that an always-maintain-schedule policy doesn't work. If there had been more passengers on the bus, or more people waiting downstream, the best policy at 110th Street would have been to proceed without waiting. So always maintain spacing is not an infallible rule either. Generally, maintaining spacing is more likely to be optimal early in a run (because the number of people on the bus is small and the number who will be forced to a later bus is large) or when the second bus is really close to the first (because then the number of people who will get on the second bus is small and the wait for those who miss it is long), or when the weather is really lousy (because then additional outdoor waiting time is so much more unpleasant than waiting inside a nice warm bus).

These two lessons are important to bear in mind when you think about who should own and operate transit systems.

C. The Right Fares

The other very practical problem that any transit operator must grapple with (aside from figuring out the schedule and how to deal with delays) is how to charge for the service. Transit systems have adopted a wide variety of fare systems: most elevators and airport monorails are free; airplanes are expensive, and the rules under which they are priced are too complicated for any mortal to understand; the New York City subway system until very recently charged a single flat fee for any ride, any distance, any time; Washington Metro has peak and off-peak fares that vary by distance.

When we looked at car trips, we saw that an attractive principle was to charge for each trip exactly the costs that that trip imposed on the rest of the world. Then people would make only those trips and all those trips for which benefits exceeded costs. These are the right trips to be made in the sense that if any other set of trips were made, there would be some way to make someone better off without making anyone worse off. This principle is a good place to start when thinking about what transit fares should be like.

First let's think about a bus or train that's not crowded. It's hard to see any costs that my riding this train imposes on anyone—almost no one would gain were I to be vaporized after I boarded. My weight wears down the seat some and hastens the day when it needs to be replaced; the engine burns a little more fuel because of my weight; but I perturb the world little by my riding. So I should pay little or nothing for riding.

Boarding is more likely to be costly. On a bus, especially, boarding takes time, and if I clamber up the stairs and fumble with my change on a bus with

a lot of people on it, I delay them. I also delay people waiting for the bus downstream. Similarly, there are costs when I exit the bus if I do so before the end of the line (and if we all file off at the end of the line, I may be holding up people behind me who could get to their real destinations faster if I weren't there).

Even this account of boarding and disembarking costs may be somewhat exaggerated. Most subways and some trains have many doors, and so boarding goes quickly; indeed if I get on board while people are still getting on at another door I impose no cost whatsoever. If I delay a bus with an unknown schedule, as in the gregariousness problem, I help people who would otherwise just miss it, and if I delay a bus that is running ahead of schedule, I may be reducing the time it will sit somewhere waiting to get back on schedule.

So on uncrowded vehicles, fares should be minimal, and they shouldn't vary with distance (nobody gets on or off more than once). They should vary with the number of riders using the rest of a route—not necessarily the number boarding now. If fares are difficult to collect—or, as on buses, collecting them adds to boarding or disembarking delays—free transit has much to commend it. (And free travel makes a lot more sense than free transfers, since transfers encourage disembarking and boarding, which are the expensive components of travel during uncongested times.)

Notice that this argument for free transit is almost exactly the same as the argument for free roads. The answer to the question, Who pays for the bus driver and the gas? is the same as the answer to the question, Who pays for the road? It doesn't particularly matter who pays for the road, the bus driver, the gas—so long as it doesn't distort too much people's decisions about when to use the road or the bus. The road should be built or the bus route should be scheduled if and only if the benefits that people derive from it when it is optimally priced exceed the costs of providing it, but good roads and good bus routes don't necessarily show a profit, or even break even. That's what taxes are for.

Whether or not an uncongested road is worth building or an uncongested bus route is worth scheduling then depends in part on the elasticity of demand. Obviously an uncongested road or route can't have infinitely elastic demand, because by definition even if you cut the price all the way to zero you would still get so few people using it that it would be uncongested. Something is uncongested only if very few people get any benefit at all from using it. The route is worth scheduling or the road is worth building if some of the people who use it get substantial benefits—that is, if some people are a lot different from others and demand is fairly inelastic. If everybody is pretty much the same, the project is not worth doing.

Where mass transit varies in some ways from roads is in how mistakes are treated. If you build a road and it turns out that almost nobody uses it, there's not much you can do. It should be free, it should be maintained if the

cost of maintenance is less than the meager benefits the road provides, and it should sit there essentially forever as a monument to your stupidity. **Sunk costs** are sunk, and any other strategy only makes a bad situation worse. To a large extent the same is true of mass transit systems. If you build a subway line or buy buses and almost nobody wants to use them, you are going to have to live with your mistake for a long time. But it's easier to adjust mass transit service than it is to reduce the size of a road, and easier to adjust buses than rail routes. Subway schedules can be changed and buses can be rerouted when they turn out to be insufficiently popular. Buses usually wear out more quickly than roads or rails, and it's also easier to sell them in the secondhand market. So you should expect to see fewer mistaken bus routes than mistaken roads.

"Mistakes" here also include occurrences and changes that nobody could have anticipated.

There is no reason to expect every bus route and road to make money. Most elevators, parks, and public restrooms, for example, are free to the individual user. Nature is bountiful: once certain investments have been made, they bear fruit at very little additional cost; people should be able to enjoy as much of this fruit as possible. (Thus although Kain [1997] argued that the "cost" of a trip on MARTA was $8, there is no reason to charge $8 for a ride on MARTA once it has been built. The question Kain was addressing was whether MARTA should have been built, not how it should operate. When he says the cost is $8, he means that MARTA should have been built only if the benefits of the average trip were greater than $8.)

Congestion changes the story. The kind of congestion that matters for mass transit is not the bottleneck we examined in Chapter 4, where one person's use of a facility precluded another person's use at that time. Under congested conditions with mass transit, one person's use of a bus or train may degrade every other rider's experience. If a train is crowded, the number of people I delay while boarding and disembarking is much greater. If I sit, I deprive someone else of a seat and squeeze everyone else in even tighter. My legs create an obstacle for people standing, boarding, and disembarking. If I stand in one of the good places—next to a pole with a moderately interesting ad to look at—I force someone else to search for something to hold or to stand next to the door. If I stand next to the door, I force people to negotiate their way around me and to claw at the standees next to me for support, and I slow down the flow of people on and off the train and delay everyone (except people downstream who don't miss this train) by making these station stops take longer. I annoy the people around me because I can't help but observe what they're reading and overhear their conversations. If I sneeze, a score of people inherit my germs. If I get sick and need an ambulance, thousands of riders are delayed. And once I get off the train, I get in the way again in the crowd surging for the stairs, slowing down both those on the way out and those on the

way in. The important point is not that I am miserable under these circumstances but that I contribute to the misery of everyone else.

Thus fares should be much higher for congested trains and buses than for uncongested ones (even though the congested trains are more uncomfortable, too). Because the longer I stay on a congested train the more costs I impose, fares should also vary with distance. But the only kind of distance that matters is congested distance: the fare for a five-mile trip should be just about the same as the fare for a ten-mile trip if both trips are congested for only four miles. And like optimal tolls, optimal fares should vary by the minute.

Will fares calculated in this way allow trains and buses to break even? Under some circumstances the answer is yes. This sounds like a remarkable coincidence, because optimal peak-hour fares depend for the most part on the costs riders impose on other riders, not on the costs they impose on the transit system. What's even more remarkable is that under these circumstances optimal peak-hour fares will cover not only the cost of peak-hour operations but also the off-peak deficit, and they will do so exactly.

What are these conditions that make everything work out so tidily? There are four:

1. Optimal capacity. The system has been built to the right capacity.
2. Constant cost, flexible capacity. In designing the system, planners could think in terms of very small increments and decrements to capacity (capacity doesn't come in big lumps), and each of those small increments and decrements cost the same amount to build. When you think of capacity as the maximum number of riders who can arrive at a central destination in a five-minute period and exit, then this condition says that when you design the system, this number can be anything—it doesn't have to be evenly divisible by a hundred, say—and thus the cost of building a system for any number of riders is proportional to the number of riders.
3. Utilization ratio. The cost of congestion that each individual rider experiences depends only on the ratio of actual riders to capacity, which I call the **utilization ratio.** Each of forty riders in a train car designed for fifty is just as uncomfortable as each of eighty in a car designed for a hundred.
4. No other distortions. Everything else is priced at **marginal cost,** so what people are willing to pay for mass transit trips—the demand curve—reflects perfectly the social benefit of those trips.

Together these conditions imply that prices based on congestion alone will exactly cover the cost of building a subway or bus system. Why?

To keep things simple, assume no operating costs. Think about a certain number of people using the system, and concentrate first on finding the opti-

mal capacity to accommodate this number of riders. By optimal capacity I mean the capacity that minimizes the cost of moving these riders. This cost has two components: the cost of building the capacity (call this construction cost) and the cost of congestion borne by the riders (call this congestion cost). Increasing capacity raises construction cost but reduces congestion cost. Optimal capacity for any population of riders minimizes the sum of these two costs.

When capacity is optimal, the cost of building a tiny amount of additional capacity is going to be the same as the reduction in congestion cost that that tiny amount of additional capacity causes. (Condition two, above, allows you to think about tiny bits of capacity.) Consider a change in capacity of 1 percent from the optimal. If the construction cost of 1 percent more capacity were greater than the reduction in congestion cost it caused, then you could decrease total cost by reducing capacity—which is impossible if capacity is optimal. Similarly, if construction cost were less than the congestion-cost decrease, you could decrease total cost by increasing capacity. Since the construction cost of 1 percent more capacity can't be either greater or less than the reduction in congestion cost that a 1-percent change in capacity brings, it has to be the same—when capacity is optimal.

Now switch your attention to optimal prices. Think about a 1-percent change in the number of riders. With congestion pricing, the revenue from this small group of riders is the cost of the congestion they cause—the amount by which the congestion cost of the other riders would fall if this small group were vaporized. That's the reduced cost of congestion from a 1-percent decrease in the utilization ratio.

Increasing capacity by 1 percent also decreases the utilization ratio by 1 percent. So the reduced cost of congestion from 1 percent more capacity equals the reduced cost of congestion from 1 percent fewer riders, which equals the revenue raised from 1 percent of riders.

But we've already seen that the decreased cost of congestion that results from 1 percent more capacity is the same as the construction cost of that capacity (when capacity is optimal). Pulling all this together, the revenue raised from 1 percent of riders pays for 1 percent of capacity. That means the revenue raised from 100 percent of riders pays for 100 percent of capacity. The system breaks even (see Box 5A).

This implies that off-peak rides—rides at times without congestion—can be free (or can cost a nominal amount for boarding and disembarking). Peak-hour congestion fares pay for capacity; everyone else considers capacity as something already there. And they are right to do so, because the calculation of optimal capacity did not take off-peak use into consideration because it was irrelevant. When engineers design a flood-control project, they don't have to calculate water flow on sunny days.

Of course, the four conditions needed to make congestion fares equal

Proving That Congestion Pricing Breaks Even

≡ The fact that congestion pricing leads to break-even conditions can be shown more quickly mathematically. Let c designate the capacity, x the number of riders, K the construction cost per unit of capacity (so total construction cost is cK), and $g(x/c)$ the individual congestion cost when the utilization ratio is (x/c).

For any number of riders x, the problem is to find capacity c that minimizes total cost:

$$cK + xg(x/c).$$

The first term is construction cost and the second term is aggregate congestion cost.

Differentiating with respect to c yields the first-order condition:

$$K - (x^2/c^2)g'(x/c) = 0$$

or

(1) $$cK = (x^2/c)g'(x/c).$$

Now consider marginal cost pricing once optimal capacity c has been installed. Each additional person increases every individual's congestion cost by

$$(1/c)g'(x/c),$$

and so since there are x people who suffer this loss, each person should pay a fare of

$$(x/c)g'(x/c).$$

Because the total number of people paying this fare is x, total revenue is

$$x(x/c)g'(x/c) = (x^2/c)g'(x/c),$$

which from (1) covers construction cost exactly.

Mohring (1972) is responsible for this remarkable result.

construction cost don't always hold. Planners aren't omniscient, and so capacity is only rarely optimal; many transit systems have increasing returns to scale in the construction of capacity; the utilization ratio sometimes misses important aspects of congestion; and everything else isn't always priced right. Systems that are too large and that have increasing returns to scale lose money when fares are set optimally; systems that are too small make money.

Violations of the fourth condition—no other distortions—have drawn

the most attention. Automobile travel is a substitute for mass transit, and almost everywhere, as we saw in the last two chapters, it is underpriced. At off-peak hours, pollution, accidents, and noise may not be properly accounted for in gasoline taxes; and at peak hours, drivers don't pay for the congestion they cause (unless dynamic congestion pricing is in effect). The peak-hour difference is the larger and the more serious.

This implies that transit fares, especially peak-hour fares, should be lower than they would be if everything were priced right. Even if the other three conditions are met, so long as highways are underpriced, public transit systems shouldn't make money. The reasoning is simple: if you get the full benefit of your actions, you will make the right decisions if and only if you bear the full cost of your actions; but if you get only half (say) of the benefit of your actions, you will make the right decisions if and only if you bear half of the cost of your actions. When car travel is underpriced, people who take public transit instead of cars don't get the full benefit of their decisions; drivers who are less delayed and breathers who have better air would be willing to pay them to take the train but don't. People will make the right choices between cars and mass transit only if transit users pay somewhat less than the total costs they impose on other people. (Of course, this argument for low transit fares depends on the empirical proposition that transit is a good substitute for cars. Empirical evidence doesn't support this proposition strongly.)

On the other hand, if transit fares were reduced by exactly the total benefit drivers and breathers get when one less motorist is on the road, people would make the right choice between cars and transit, but they would make too many trips overall, because trips by both modes would be underpriced. Some trips by both modes would have costs greater than benefits. One kind of problem—wrong mode choice—would be eliminated, but another problem—too many total trips—would be exacerbated.

The optimal policy is complicated to find but it is essentially a moderate one, in that it doesn't completely eliminate either kind of problem and doesn't completely ignore either type of problem. The optimal transit fare is somewhere between the low fare that would sort people between modes correctly and the higher fare that would get the total number of trips right.

You will find similar reasoning used at several other points in this book: when two things are substitutes and you can't get the price of one of them right, you shouldn't try to get the price of the other precisely right either. In economist shorthand, this is called **second-best** reasoning. "First-best" is to get both prices right, but if that's impossible, second-best is to get the price you can control "wrong" by the right amount. Second-best is common sense: when you take a shower and the flows of hot and cold water are both working perfectly, you set both at levels that give you your first-best temperature and volume. But if the flow of hot water is for some reason restricted below the level you want, you reduce the flow of cold water below its first-best level and

achieve a second-best shower with neither first-best temperature nor first-best volume.

Notice that this form of second-best reasoning applies to prices, not to investments. The argument here is that the price of a transit trip should be reduced because the price of a car trip is too low, not that excess investments should be made in transit because excess investments have been made in roads (if they have). In fact, the opposite is true—if for some reason a society has built too many roads or is doomed to do so, then it should invest less in transit than it would have done if road investment were optimal. If you live in Florida, you don't invest in a powerful furnace just to be fair to fossil fuels.

II. Getting It Done

A. The Governance Question

So far in this chapter I have outlined how transit systems should operate in an ideal world—what modes they should use, what schedules they should try to adhere to, what they should do when something goes wrong and they can't adhere to the schedule, what they should charge and when. "Should" here refers to potential Pareto improvements; any deviation from these prescriptions sets up an opportunity to make someone better off without making anyone else worse off. These prescriptions have two immediate uses. For altruistic people running transit systems, they point to which sorts of reforms might be worth striving for, political and administrative constraints permitting. To citizens in democratic countries, they give some guidance for deciding whether the political leaders they select are doing or promising to do things that would really be helpful.

But understanding how transit systems would operate in an ideal world doesn't make them operate that way. Real transit systems are run by people like us—lazy, venal, ignorant, maybe well meaning but still needing to watch out for themselves and their families first. They probably don't know very accurately very many of the important parameters we've identified for calculating optimal service and fares; and even if they knew what the relevant parameters were, they might not have any reason to want optimal service and fares.

The question therefore is whether the governance of transit systems can be set up so that the people running them have both the information and the incentives to make the right decisions. Or even passably good decisions. Nobody I know has a good answer to this question (except a simple "no"). But the question of how transit systems should be governed is clearly an important one—it's not much use figuring out what transit systems should do unless you can figure out how to get them to do it. It has been a subject of intense and furious debate for more than a century.

Transit systems today are governed in many different ways, so it is un-

likely that there is one single best way to govern them always and everywhere. Building owners run the elevators and escalators in their buildings, but the buildings' visitors determine where the elevators stop. The Staten Island Ferry is municipally owned and operated, but the dozen other ferries in New York harbor are private. Most airlines in the United States are owned by investors (employees have a significant stake in a few), but many other airlines in other countries are owned by national governments. In almost all U.S. cities, local governments and quasi-governmental agencies control the largest commuter bus enterprise, but many private bus services, both big and small, operate alongside the public buses. British bus service outside London is almost totally privatized and deregulated. Private jitney services operate in Atlantic City, San Diego, and San Francisco; they are the dominant form of mass transit in many cities in developing countries. American school buses are mostly privately owned and operated, but their activities and payment are fairly tightly controlled by contracts with school districts. Airport shuttles are mainly private, and some are run by the hotels they serve; airports themselves usually operate the monorails and interterminal bus routes. Subways, trolleys, and passenger railroads in the United States are now almost all public, except in amusement parks, although freight railroads are privately owned.

The actual governance question, moreover, is much more complex than a simple public-private dichotomy. Governance is not a single decision but a series of decisions: what kind of buses to buy, how much to train bus drivers and how much to pay them, whether to run a holiday schedule on Washington's birthday, where to leave Mrs. Jones off and when to wake Mr. Smith up, what music to play or permit or ban. Each of these decisions can (and probably will) be made by someone different, responding to different rules and incentives, with different degrees of discretion. Elevators are subject to safety checks and regulation by municipal governments. Airplanes land at publicly owned airports under the direction of publicly paid air-traffic controllers. Airport security was provided by private companies before 9/11 and has been a public enterprise since. Private buses drive on public streets and pull over to pick up and discharge passengers only where public authorities permit them to. Public buses buy gasoline from private vendors and adorn themselves with private advertisements.

The same transit system can see many different allocations of decision making. Bus service in Kingston, Jamaica, for instance, was provided by a heavily regulated private company between the 1950s and 1974. The city took over the bus system, which was then bankrupt, in 1974. Costs became exorbitant, and the city privatized it again in 1983 and deregulated the market (Klein, Moore, and Reja 1997, p. 29).

The history of the New York City subway system also shows this mixing of private and public and the same pattern of changing roles. A private but heavily regulated company, Interboro Rapid Transit (IRT) completed con-

struction of the first subway route in 1904, and it was immensely successful. To expand the subway system, the city in 1913 entered into two very complicated contracts, one with the IRT and one with Brooklyn Rapid Transit (BRT). These contracts—known as the Dual Contracts—gave the companies operational control but divided financing and revenue between the city and the companies. The Dual Contract subways gained as many riders as their backers had hoped, but unanticipated inflation in 1918 made the contracts a financial disaster because state law required a nickel fare.

Foes of privately owned mass transit captured City Hall and the state government in Albany in the 1920s, and they built New York's first wholly public subway system, the Independent Subway System, or IND. Both of the private companies went bankrupt, and in 1940 the city bought them. Since 1940 the New York subways have been publicly owned and operated (and the fare is a lot more than a nickel).

B. Good Matches and Bad Matches

Even though there does not appear to be any universally favored form of transit governance, the historical record and the analysis of good transit operation suggest a few useful precepts.

The first is that information, incentives, and decision-making powers should be tied together as closely as possible. In the 1920s and 1930s, for instance, due to a quirk in New York City's Dual Contracts, the private companies made most of the decisions but profit at the margin went to the city. By all accounts the subways were horribly run at this time. Major General John F. O'Ryan, who headed a panel charged with investigating subway service in the 1920s, concluded that if he had treated prisoners of war the way the private companies treated passengers, he would have been court-martialed (Hood 1995, p. 93).

Elevators, in contrast, are a type of public transit system that seems to work well. The owners of office and apartment buildings make most of the long-run decisions about elevators (how many, what operating algorithm, how often they will be maintained and repaired) and bear most of the long-run burdens and benefits (lousy service will hold their rents down and contribute to vacancies; good elevator service will allow them to raise rents). Short-run scheduling decisions are made by the elevator riders themselves, who have every reason to economize on their own time. Riders don't take into consideration the costs they impose on other riders—my stopping at the eighth floor delays your trip to the fifteenth floor—but if these costs become large, the owners can (and probably will) reduce them by restricting the floors at which different elevators stop.

A second general precept is that strong incentives produce desirable results only when the people who have them enjoy all the benefits and bear all

the costs of their actions. Incentives don't help unless they're the right ones. Elevators don't work well when landlords can't raise rent or when tenants can't move out.

The privatization and deregulation of buses in Britain provide an interesting example of how subtle misalignments of incentives can cause major operating problems. When privatization and deregulation were first proposed in the early 1980s, proponents argued that competitive pressures would force fares and operating costs down, and that lower fares and improved service would increase patronage on many routes. Instead, while operating costs have fallen because of cuts in wages and employment, fares have increased, ridership has fallen (though at about the same rate it was falling before deregulation), and little on-the-road competition has developed.

What happened? Under the rules adopted in 1985, anyone could offer service on any route (provided there were no serious safety or traffic congestion problems), even if another bus service was already operating there. You could even use the same bus stops as other operators. All you had to do was register your schedule with the appropriate authorities forty-two days in advance.

These rules made what economists call "schedule jockeying" the main way in which bus companies competed against each other. This is where the incentive problem develops. Suppose I am operating a lucrative bus route and you want some or all of my profits. If incentives are right, you will have to offer either better service (nicer buses, more helpful drivers, more frequent scheduling) or lower fares to win over my passengers and earn some profit. All of these ways of competing are expensive, but they give customers real benefits.

Alternatively, you can just schedule a bus thirty seconds ahead of each of mine. You will get most of my passengers even if your buses are not nicer or cheaper than mine, and you won't have to incur the additional costs of service enhancements or fare cuts. It should not be surprising that British bus companies wishing to enter new markets turned to schedule jockeying rather than the more expensive (and socially beneficial) alternatives. (Notice that schedule jockeying doesn't occur with airlines or long-distance buses, because customers have reasons to buy tickets in advance for a specific flight.)

How would I respond if you were to start running buses thirty seconds ahead of mine? One way would be to revise my schedule and run my buses thirty seconds ahead of yours. But if I do that, you will respond in kind, and in a little while many customers will be so thoroughly confused that they will stop riding buses altogether. Both of us will lose. So instead of piecemeal schedule revision, the common response of an incumbent bus company threatened by a schedule-jockeying entrant was to practice "route swamping": running so many buses that the entrant has no way of finding enough passengers to make a go of it.

As Klein, Moore, and Reja (1997, pp. 70–71) explain: "Route swamping has a twofold strategic quality: it not only drives out the current challenger, it also discourages future challenges. Larger incumbent firms have been known to maintain 'fighting fleets' that were available immediately to meet any competitive challenge."

Route swamping is why bus-kilometers have increased in Britain since deregulation while ridership has fallen, and why bus companies have consolidated, even across cities. The more routes I have, the easier it is for me to maintain a fighting fleet big enough to swamp any challenger who might surface on one of them.

By publishing schedules and being reliable, firms create congregations of waiting passengers, but they can't protect those congregations from the depredations of other bus companies. "Incumbent firms have an investment in the passengers waiting at the curb; the service they have been providing is what draws the passengers there. Faced with interlopers engaged in schedule jockeying, they turn to route swamping simply as a means of protecting their investment" (Klein, Moore, and Reja, p. 70). It is as if Burger King could send employees with already cooked burgers to solicit business from customers waiting on line at McDonald's, or physicians could drum up business in each other's waiting rooms.

The problem is, the difficulty of protecting their investments in waiting customers gives both challengers and incumbents incentive to act in ways that are not socially optimal: challengers do better by stealing customers than by offering them better and cheaper service; incumbents do better by investing in weapons to scare or drive newcomers away than in ways to make the customers happier. Privatizing and deregulating bus service won't lead to the outcomes its proponents hope for unless there is some way for companies to protect their investments without expensive weapons like fighting fleets.

This problem may have some pretty good solutions. Inventors face a similar difficulty, for example, in keeping other people from stealing their ideas, but patents have evolved to help them do so. Klein, Moore, and Reja propose a system they call "curb rights": along with privatizing and deregulating mass transit, setting aside specific spaces on sidewalks, curbs, and roads that would be exclusively controlled by each bus company. These spaces would be managed as private resources, like physicians' waiting rooms: "Jitneys and other buses would be prevented from picking up passengers within these curb zones (unless they received express permission to do so from the holder of the curb right)" (p. 97). Companies would bid for the rights and could resell them. In essence, curb rights would substitute enforcement by the state for enforcement by the fighting fleets. Privatized transit cannot function without some mechanism—like curb rights, perhaps—that will protect and encourage investment in passenger congregation.

Protecting this sort of investment isn't the only incentive-alignment

problem that arises in the design of privatized transit services. Remember that routes that should operate don't always break even or make money, and sometimes optimal transit service is even free. Most profit-making businesses won't provide these services, and those that do are likely to charge too much.

The elasticity of demand is the key to understanding how routes that should operate don't make money with optimal fees. On these routes, some people get large benefits but other people get much smaller benefits. With optimal fees, the first group of riders pay a lot less than the benefits they receive, and an increase in fares would drive the second group away and so waste capacity. This difference between the benefits the first group receives and the payments it makes causes a misalignment of incentives; the firm deciding whether to optimize can't realize in cash all the benefits of its affirmative decision.

Of course this misalignment, too, can be corrected without resorting to publicly operated transit. The obvious way is for the government to subsidize the private operation of such routes. About 35 percent of U.S. transit agencies contracted out some of their services in the 1980s (Teal 1988, p. 209). But subsidy contracts have their own problems.

The major problem with subsidy contracts is that they can be very expensive for the government when the private company knows a lot more about the service's profitability than the government does. This leads to our third precept about transit governance: The more inefficient the general tax structure is and the greater the advantage private firms have in information about profitability, the less well contracts and subsidies work. Under **inefficient taxation,** taxes raise a lot less money for the government than the pain they cause the taxpayers; if a tax structure is very inefficient, taxpayers would be willing to pay (as a lump sum, of course) a lot more money to the government to stop taxing them than the government is actually collecting. **Deadweight loss** per dollar of revenue is big. As usual, everything here is about potential Pareto improvements.

New York City's Dual Contracts arrangement of 1913 offers a good example of the intuition behind the precept. When the negotiations that led up to the contracts were being conducted, city officials knew that expanding the subway would bring enormous benefits to many New Yorkers, but they didn't have much of an idea about the cost and revenue potential of the lines they wanted built. The companies had a much clearer idea of the profitability of these lines (the IRT had just built a subway and was operating it; the BRT was already running mass transit in most of the neighborhoods where its subways would be built). To make sure the companies would go ahead and expand, the city had to offer a package sufficiently remunerative that the companies would agree even if their beliefs about profitability were the most pessimistic plausible beliefs. If my car breaks down in a strange town and I clearly have no idea what's wrong with it, the mechanics there have every incentive to diagnose the

most expensive malady that is plausible to me, and my ignorance gives me no alternative but to accept their diagnosis. Getting a car repaired under these circumstances can be very expensive; the same is true for getting a subway built. Thus the Dual Contracts arrangement had provisions favorable to the companies, such as a guarantee of existing profits on the IRT lines that were already operating, seniority for the company bonds, giving the companies the right to determine operating costs. If inflation had not intervened the companies would have done very well. Even with inflation, the city probably could have built the Dual Contract subways at less taxpayer cost on its own, since ridership met expectations and the city, acting as its own contractor, would not have had to guarantee existing profits.

So the greater the information advantage that private firms have, the greater the expected cost to taxpayers of subsidizing their activities. The more inefficient the general tax structure, the greater the losses to taxpayers for each dollar of subsidy they have to provide. The two steps work together in the third precept: more subsidy and more loss per dollar of subsidy both make contracting a worse idea.

The extra cost to taxpayers has a secondary effect, too. Some projects that would be worthwhile for them if they didn't have to pay extra (once in the deadweight loss of taxes and once in the expected profit to the contractor) aren't worthwhile when they do have to pay extra. Subsidized transit systems end up smaller than systems directly run by the government. During the late 1920s and early 1930s, the IRT and the city battled for years over whether IRT stations would be expanded to handle eight-car trains instead of five-car trains; the city finally prevailed but had to pay the IRT a lot. The city built the IND at about the same time, with stations big enough for ten-car trains. In light of this experience it is tempting to see the IND itself not as the triumph of anti–private-transit political factions but as an expansion of the system that would have been too expensive to carry out if it had been done through the companies.

This reasoning also explains a good deal of the worldwide trend toward public ownership of mass transit systems during the middle of the twentieth century. Between the 1940s and the 1970s, most bus, subway, and trolley systems in the United States switched from (heavily regulated) private ownership to public ownership; a similar trend was evident in most other developed countries at this time as well. Why did this happen?

After World War II, American mass transit systems lost passengers rapidly. Reasons for this are not hard to find: growing popularity of automobile travel, better roads, suburbanization of housing and the corresponding decrease in density of origins in the morning and destinations in the evening, suburbanization of employment and the corresponding decrease in the density of destinations in the morning and origins in the evening, rising wages and the corresponding increase in the value of waiting time and schedule de-

Did GM Kill the Trolleys?

≡ A popular alternative theory about the demise of many mass transit operations shortly after World War II is that they were victims of a conspiracy by General Motors (GM) to replace trolleys with cars. Goddard (1994) and Kwitney (1981) offer the basic expositions of this story. Most of the evidence comes from an antitrust case, *U.S. v. National City Lines,* 186 F2d 562 (1951), in which GM and several other companies and individuals were found guilty of conspiracy and fined, but only very lightly.

The basic facts are that during the Depression and the early 1940s, GM and several suppliers of buses and the products that buses use (Mack Trucks, Firestone, Phillips Petroleum among them) financed and assisted a bus company called National City Lines (NCL). NCL bought around thirty failing trolley car lines, got rid of the trolleys, and replaced the trolley service with bus service. The antitrust violation was that the financing companies forced NCL to buy all its supplies from them, and to agree never again to use electric-powered transit. NCL's principals became very rich from this enterprise.

I don't see how you can draw from these facts the conclusion that trolleys would still be operating if GM and the other suppliers had acted competitively and not violated antitrust laws. The story shows that trolleys were such a bad business that the present value of future profits from an already-built trolley system was less than the present

lay cost, the baby boom and the advantages of transporting small children by car (see Box 5B). Because density of demand is so crucial to the profitability of mass transit operations, many lines began to lose money, and even more would have if they had been required to charge fares equal to marginal cost.

But there remained lots of reasons why various people wanted at least parts of these mass transit systems to keep operating. Some of the reasons have to do with benefits to riders; I explored these reasons in the first part of this chapter. Other reasons include the effects on land values: buildings next to a subway stop are more valuable because of their easy access to the subway; if the subway goes out of business they will lose value. I will examine these reasons in much greater depth in the next chapter.

To keep transit systems operating, somebody had to subsidize them (or subsidize them more than in the past). The subsidy could come from general government revenues or it could come from fares much higher than the marginal cost; for this discussion the distinction doesn't matter, because the difference between a high fare and marginal cost is, for all purposes, a tax just like a tax on cigarettes—it makes something cost more than it otherwise would and generates revenue that can go to the transit operator.

value of joint profits from building and operating a new bus system, even with the cost of building the buses netted out. Since monopoly generally restricts output, my inclination is to say that GM's conspiracy and monopoly slowed the conversion from trolleys to buses, rather than forcing it artificially.

Moreover, I've seen no evidence that consumer surplus shrank when NCL replaced trolleys with buses. Indeed, Goddard, a conspiracy theorist, cites at least three instances in which NCL acted to increase consumer surplus. In Montgomery, Alabama, and Kalamazoo, Michigan, NCL slashed fares (p. 131). In Baltimore, NCL eliminated the inconvenience of transfers (p. 130). Goddard describes the general modus operandi of Roy Fitzgerald, NCL's chief manager, in the following way (p. 126): "Typically, Fitzgerald cut back on fares immediately, then stood on the streets to see how many people were still walking to work. If need be, he would trim fares some more."

Providing cheap and convenient mass transit with buses, and trying to get as much ridership as possible, seems like a good way to reduce demand for cars, not increase it.

In the absence of any information indicating shrinking consumer surplus, and with good evidence of an increase in profit, it's pretty clear that replacing trolleys with buses was a potential Pareto improvement. After the antitrust case was decided in 1951, there was no move back to trolleys, and so it's very likely that the contractual prohibition on reconversion was not binding. Only a large failure of the market system could have prevented the replacement of trolleys with buses.

By this time, however, incumbent transit operators had acquired information for decades about costs and demand; they knew an awful lot more than the government officials and the general public about how their business operated. Subsidizing them, either by writing checks from the general treasury or by allowing generous fare hikes, was going to be very expensive, since they could claim that they were less profitable than they were and nobody could really tell. So in accordance with the third precept, governments found that taking these operations over was going to be cheaper than allowing them to stay in private hands. Some governments—New Jersey's, for instance—tried subsidizing transit for a decade or two but then moved to direct operation. That's how mass transit became mainly a government responsibility.

Government ownership of mass transit in the United States has run into serious problems of its own, however. Ridership has not picked up to any great extent and on some measures has continued to fall. In 1980 the average bus carried 13 passengers; by 1995 the average had fallen to 8.6 passengers, and by 2001 it had recovered some, but only to 9.3 (U.S. Bureau of Transportation Statistics 2002). Real operating costs may have risen even faster than ridership has fallen.

With this sparse usage, buses do not enjoy a significant advantage over cars in reducing external costs. Since 1991, buses have used more energy (measured in British thermal units) per passenger mile than regular (non-SUV) cars have used, and even when vans, pickup trucks, and SUVs are included in the passenger fleet, buses still come out behind, but only by a little (U.S. Bureau of Transportation Statistics 2002).

Air pollution is probably roughly proportional to energy usage. But because most buses use diesel fuel (in 2001, 71.5 percent of new buses used diesel; see American Public Transit Association 2002), and diesel engines produce far greater particulate pollution than standard car engines do, pollution from buses might be a more serious problem per passenger mile than pollution from cars, since we have seen that particulates are the most dangerous type of vehicle pollution. In addition, data from Deleucchi and Hsu (1998) indicate that buses cause much more damage from noise per passenger mile than cars do. And buses cause considerably more road damage.

What went wrong? One easy answer is that government bureaucracies are naturally self-aggrandizing and so automatically overdo any service they are charged with providing. Power, glory, and an easy life come from a big budget, and so bureaucrats strive to command big budgets. Certainly this explanation points to some important and widespread weaknesses of government functioning, but when applied to mass transit it seems to explain both too much and too little. Is it really true that governments always overprovide—too many police officers, too-big libraries and too many parking spaces next to them, too much water, too many teachers and too much chalk for them to use, too-clean cafeterias in schools that serve too-tasty food, dog licenses that are too big and too shiny for their intended use? If not, what is special about mass transit that explains overprovision in this sphere but not, for instance, in snow plowing or street sweeping? And why does overprovision take the form of grandiose subway systems and big buses with almost no passengers rather than, say, swarming fleets of minibuses?

To try to answer these questions, it's best to start with the fourth and final precept: the more likely it is that government decision making will be influenced by considerations other than public well-being, the better private ownership performs. Private ownership and the informational advantages that it entails make it more difficult for government officials to achieve their ends, and so they will achieve those ends less well. The public should be unhappy about this if the goals of the government officials are consonant with their own goals, and happy about it if they are not.

What has been striking about the recent governmental forays into mass transit operation has been the emphasis on the effects of mass transit on property values, not on the benefits to riders. Federal involvement in mass transit did not originate in the congressional committees concerned with transportation or public works; instead, it began with the 1961 Housing Act in the hope of spurring urban renewal (Meyer and Gómez-Ibáñez 1981, pp. 41–

43). New subways and trolleys have often been sold on their "economic devel-opment" aspects and usefulness in reviving downtowns (they have also been sold on environmental grounds). And some of the new subways have indeed spurred considerable office and commercial development along their routes. Politically, mass transit has often been seen as more of a real estate issue than a transportation issue.

This emphasis explains why rail projects are more popular than bus proj-ects. A subway station or a trolley stop is not something that is going to go away soon; it represents a long-term government commitment. I can make long-term investments, like building an office tower, secure in the knowledge that my transit access is not subject to political vagaries and whims. A bus stop, in contrast, offers no such security. Bus service can be rerouted, cur-tailed, or eliminated at the snap of a politician's finger or the drop of an elec-tion return. As a real estate proposition, rail has advantages that do not appear in the transportation calculations we have been making. (Ironically, the flexi-bility advantage of buses, their ability to adjust to changing times better than rail, turns out to be a political liability.)

But maybe the more popular and political analysis of mass transit is right—that the real estate and "economic development" benefits are real and important and should not be ignored. Maybe the new subway systems and the record of government-operated mass transit systems are not failures. Why am I so sure that economists' neglect of these arguments for mass transit, and their consequent dour outlook, is correct? To answer this question, we have to understand how urban land markets work. That is the subject of the next chapter, and so I will not answer that question here. But I will answer it even-tually.

III. Conclusion

We have come full circle. In the first part of the chapter, we analyzed an ideal transit system and saw that with U.S. residential and employment densities close to what they are now, no economist would support any expansion of rail transit and few would support very large expansions of bus transit. But cur-rent U.S. residential and employment densities were created by a series of public policies, including transportation policies; they didn't drop directly from the hand of God. So to really understand transportation policies—both cars and mass transit—we have to understand what goes into creating resi-dential and employment densities, and whether we have the right ones now.

The second part of the chapter looked at what sort of governance struc-tures would implement good mass transit policies. The key unanswered ques-tion here turned out to be how you should think about the benefits that real estate owners accrue from transit systems.

So both parts of this chapter—and the previous chapters, too—point in

the same direction: we need to look at land, and real estate, and buildings; why things are located where they are and cost what they cost. To that task we turn in Chapter 6.

Questions

1. The study by the Berkeley researchers (Keeler et al. 1975) that determined the relative costs of car, bus, and rail transportation was completed many years ago. Much has changed since then, including our understanding of medicine and the environment. Following is a list of some of these changes, many of which have been discussed in this chapter and in Chapters 3 and 4. For each change, discuss how it would modify the Berkeley results, if at all.
    a. Car technology and operation
    b. Rising real wages
    c. The discovery of the problem of global warming
    d. The discovery of the dangers of particulates
    e. The rising price of gasoline
    f. More working wives
    g. The Web
    h. Better, cheaper road paving
    i. Terrorism
    j. Anything else you think is important
2. Do you think that congestion pricing on highways would raise enough money to pay for road construction? Explain why or why not.
3. Two people board an elevator on the fourth floor. The door stays open for a while after they board. One of them, A, pushes the button for floor ten, and the other, B, pushes the button for fourteen. The door opens at six and one more person, C, gets on. She pushes the button for eight and gets off there. A gets off at ten, and B gets off at fourteen. The elevator was never crowded and no one was intending to pick their nose during the trip. There was a person waiting on fourteen to go down. No one else was waiting for the elevator. The elevator passengers all weigh less than 300 pounds, and all of them have time that is about average in value among educated New York City residents. In terms of social cost, whose trip was the most costly? Whose was the least? Why?
4. In Queens, along Jamaica Avenue, people's desired departure times are such that five people want to leave for Manhattan every minute during rush hour. People lose time if they have to wait or come early; they all value time lost the same way—at $.10 a minute—no matter whether it's waiting time or coming-early time. It costs $25 for a big bus to make a trip to Manhattan,

and the bus can take a lot of people (it's hard to tell the exact capacity be-
cause passengers can crowd in very tightly if they have to; they don't mind).
A trip to Manhattan always takes 26 minutes.

a. If the Metropolitan Transit Authority (MTA) wants to minimize the total
cost to society—the cost of operating the bus and the cost of waiting or
coming early and the cost of being transported—how often should it run
the buses? Nobody knows the timetable—as on the subway, passengers
believe that buses arrive randomly. How many people will be on each
bus? How much must the MTA charge to break even?

b. Suppose a small bus is available that can carry a maximum of 35 people.
A trip to Manhattan in this bus also takes 26 minutes, and it costs only $9
to operate the bus. If only small buses are available, how often should the
MTA run them to minimize the total cost to society? If the small buses
have to break even, how much must the MTA charge? If the MTA has a
choice between only small buses and only big buses, which should it use?

5. Given the information in question 4:

a. How would your answers change if the MTA published a timetable that
people were aware of, and that the MTA adhered to?

b. In what directions would your answers change if we were talking about
Manhattan's East Side, where people make a lot more than $10 an hour,
rather than Jamaica Avenue in Queens? In general, should buses ridden
by rich people be bigger or smaller than buses ridden by poor people?

6. Everyone lives on an island where accidents are the only external cost of
driving—pollution and congestion don't occur. The drivers on this island
are not very proficient; sometimes they hit another motor vehicle and
sometimes they don't. For an individual driver on a particular trip, the
probability of an accident depends on the ratio of the number of drivers x
to the number of lane-miles c on the island. The more likely you are to find
another driver on a randomly chosen piece of road, the more likely you are
to have an accident. Specifically, the probability that a particular driver will
have an accident is $0.5(x/c)$.

If an accident occurs, each driver suffers physical damage of 500; total
damage in an accident is 1,000. Every accident involves precisely two cars,
both of them being driven.

The only decision drivers make that affects whether they have an accident
is whether they make a trip at all. Cars have only one speed, and drivers are
all nearly blind. All trips cover the same distance.

Drivers differ in the benefit they derive from making a trip. On any given
day, the number of drivers who derive a benefit of at least B from making a
trip is

$$1,000 - 10B,$$

which is essentially an inverse demand curve for trips. Driving has no pri-
vate costs except those connected with accidents.

a. Suppose laissez-faire without courts prevails. Each driver is responsible only for the private harm she suffers. There are 100 lane-miles of road. Find the equilibrium number of trips. (*Hint:* Let x denote the number of trips. Find the expected accident cost of a trip as a function of x. Plug in the inverse demand curve to find how many trips will be taken if that is the cost of a trip. Equilibrium requires that the x be the number of trips people want to take.)

b. What's the individual expected accident cost of a trip if x trips are being taken? What's the aggregate accident cost as a function of x? (*Hint:* Multiply this answer by the number of trips.) What's the marginal social cost of a trip? (That is, the derivative of total cost with respect to x.) Find the optimal pay-at-the-pump insurance cost per trip. (*Hint:* Subtract private cost from marginal social cost.)

c. Show that double-strict liability is equivalent to the optimal pay-at-the-pump insurance cost per trip.

d. Suppose that no lane-miles are currently built, and the government knows the number of trips x that are going to be made every day. Once a lane-mile is built, capital cost will amount to 2.5 per day for it. So total cost consists of construction (capital) costs and accident costs. As a function of x, what number of lane-miles minimizes total cost?

e. For any capacity c, suppose the government institutes double-strict liability, with the excess revenue going to the government. That is, when there's an accident, nobody gets reimbursed anything and each party pays 500 to the government. Find the equilibrium amount of driving x.

f. Find the optimal capacity. (*Hint:* Values for x and c have to satisfy both the condition in d and the condition in e.) At this capacity, how much revenue does the government collect? What are the government's costs for constructing the lane-miles?

g. Does the government make money, lose money, or break even? How does this result relate to optimal capacity?

References

American Public Transit Association. 2002. "The Bus Market." Accessed July 24, 2002, at *www.apta.com/stats/vehicles/busmkt3.htm.*

Baum-Snow, Nathan, and Matthew Kahn. 1998. "The Effects of New Urban Rail Transit: Evidence from Five Cities." Discussion paper, Columbia University.

Deleucchi, Mark, and Shi-Ling Hsu. 1998. "The External Damage Cost of Noise from Motor Vehicles." *Journal of Transportation and Statistics* 1(3):1–24.

Goddard, Stephen B. 1994. *Getting There: The Epic Struggle between Road and Rail in the American Century.* New York: Basic Books.

Gómez-Ibáñez, José. 1985. "A Dark Side to Light Rail?" *Journal of the American Planning Association* 51:337–351.

Hood, Clifton. 1995. *722 Miles: The Building of the Subways and How They Transformed New York.* Baltimore: Johns Hopkins University Press.

Kain, John F. 1997. "Cost-Effective Alternatives to Atlanta's Rail Rapid Transit System." *Journal of Transport Economics and Policy* 31:25–49.

Keeler, Theodore, et al. 1975. *The Full Costs of Urban Transport.* Monograph 21, part 3, Institute of Urban and Regional Development, University of California, Berkeley.

Klein, Daniel B., Adrian T. Moore, and Binyam Reja. 1997. *Curb Rights: A Foundation for Free Enterprise in Mass Transit.* Washington, DC: Brookings Institution.

Kwitney, Jonathan. 1981. "The Great Transportation Conspiracy." *Harper's,* February, p. 14.

Meyer, John, and José Gómez-Ibáñez. 1981. *Auto, Transit, and Cities.* Cambridge, MA: Harvard University Press.

Mohring, Herbert. 1972. "Optimization and Scale Economies in Urban Bus Transportation." *American Economic Review* 52(4):591–604.

Romilly, P. 1999. "Substitution of Bus for Car Travel in Urban Britain: An Economic Evaluation of Bus and Car Emissions and Other Cost." *Transportation Research D,* 4:109–125.

Teal, Roger F. 1988. "Public Transit Service Contracting: A Status Report." *Transportation Quarterly* 42(April):207–222.

U.S. Bureau of Transportation Statistics. 1999. "G-7 Highlights." Accessed at *www.bts.gov/itt/G7HighlightsNov99/G-7book.pdf.*

———. 2002. "National Transportation Statistics." Accessed at *www.bts.gov/btsprod/nts.*

Vickrey, William S. 1987. "Observations on Modern Transit." Discussion Paper 356, August, Columbia University.

Chapter 6

Land

───

In 1879, America's most popular book on economics offered the following insight about getting rich:

> Take now . . . some hard-headed business man, who has no theories, but knows how to make money. Say to him, "Here is a little village, in ten years it will be a great city—[in] ten years the railroad will have taken the place of the stage coach, the electric light of the candle; it will abound with all the machinery and improvements that so enormously multiply the effective power of labor. Will, in ten years, interest be any higher?"
>
> He will tell you, "No!"
>
> "Will the wages of common labor be any higher . . .?"
>
> He will tell you, "No . . ."
>
> "What, then, will be higher?"
>
> "Rent, the value of land. Go, get yourself a piece of ground, and hold possession!"
>
> And if, under such circumstances, you take his advice, you need do nothing more. You may sit down and smoke your pipe; . . . ; you may go up in a balloon or down a hole in the ground; and without doing one stroke of work, without adding one iota of wealth to the community, in ten years you will be rich! (George 1900, pp. 291–292)

Almost all the land on this planet is worth almost nothing. In 1998 you could buy an acre in Vermont for about $200; undeveloped Siberian land was even cheaper than that. But a little bit of the land is very expensive. Some land sales in midtown Manhattan in 1998 were for prices around $1,000 a square foot (or $40 million per acre).

Why these great disparities? Surely you can't use differential fertility to explain them (any more than you could attribute the popularity of the Lin-

───

To get the most from this chapter, you should be familiar with these concepts: elasticity, excess burden, factor substitution, potential Pareto improvement, and present value. You can find these terms in the Glossary. Congestion pricing will also be referred to.

coln Tunnel at 8:30 on a weekday morning to the great views). Instead, Manhattan land is so much more valuable than Vermont land because you can accomplish so much more there, for all the reasons I outlined in Chapter 2. We saw in that chapter that there were great advantages to locating a whole lot of some activities on a little bit of land. It's much easier to find a cheap engagement ring, a good French restaurant, a skilled urologist, somebody to buy your company or your latest painting, a spouse, or a good bridge tournament if you're in Manhattan than if you're in Vermont.

But you can't do more than one thing in the same place at the same time. The same site in Manhattan can't be a jewelry store, a French restaurant, a urologist's office, a brokerage, an art gallery, a bridge parlor, and your home simultaneously. Even using the same site as a jewelry store on even days and a French restaurant on odd days would be ridiculously expensive (where would you store the jewelry cases odd days?). You could stack the uses one on top of the other in a multistory building, but even this is expensive (the taller the building, the greater the percentage of otherwise available floor space that has to go into structural support and elevators), and it still doesn't put everything in the same place—the distance from the top to the bottom of a 100-story building is about four New York City blocks.

In cities, there are more great things to do than great places to do them in. Somehow something has to allocate different activities to different places. That is the subject of this chapter; the technical name for this approach is **location theory,** the study of how activities are allocated to different places and **land rents** are determined. The primary mechanism that market economies use to allocate different activities to different places is the price of land (for instance, people don't farm in midtown Manhattan because the land costs too much). So this chapter will look at what determines the prices of various pieces of land, what activities end up at what locations as a result of those prices, and whether the resulting allocation is a good one.

You can't just simply build a sewer, a house, a gas station, an office building, or a church—you have to build each of these things *somewhere.* In the next few decades trillions of dollars will be spent to build things like these in cities around the world, especially in developing countries. Some people think that the United States has ended up putting these things in the wrong places—that they are too spread out and so prevent full realization of the economies of scale that make city life attractive and productive. You have to understand land markets and location theory to assess the arguments in this debate.

I. Purchase Price and Rent

A. Capitalization

Before beginning to examine these questions, we need to clarify some terminology. You see two different kinds of transactions in real estate markets: pur-

chase and rental. When you *purchase* something, you acquire control over it until you (or your heirs) sell it to someone else. When you *rent* something, you acquire control for a fixed period of time, after which control reverts to the person you rented from. What you pay to purchase something is called its price or value; what you pay to rent it is called its rent.

In this chapter I will refer to both numbers, and it is important to know how they are linked. The relationship I will use is a very simple one.[1] Suppose you have enough money to purchase a piece of land. Call this V, for value. You can either purchase it now, or rent it every year forever. These two courses of action are equivalent, since they both give you control forever, and so they should cost you the same amount. Why? Because if one method of gaining control were more expensive than the other, no one would use it. Because we see both methods being used, we know they are equally expensive. (Alternatively, we can just define the rental-equivalent of a purchase price as the rent that does this trick. Or vice versa.)

If you rent, you will pay R per year. If you purchase you must give up your lump sum V. Let i denote the rate of interest. If you held your money in a bank instead of purchasing the land, the amount of interest you would get every year would be iV. This is what you give up every year when you purchase. To make the two methods equally expensive, we need

$$R = iV$$

or equivalently

$$V = R/i.$$

Economists say that purchase price is the capitalized value of rent. The **capitalized value** of a stream of income is the amount for which you would sell the right to receive this stream of income. (**Present value** is another name for capitalized value.) Since I am not concerned in this chapter with changes in the rate of interest, rent and purchase price will always be proportional to each other and move in the same direction. Notice, though, that when interest rates are low, the purchase prices of real estate will be high, and when interest rates are high, purchase prices will be low.

B. Bubbles

Occasionally, exceptions occur to the general rule that prices are the capitalized value of rents. These exceptions are called **bubbles.** In a bubble, people buy property in the hope of reselling at a higher price. They hope to sell to other people who are buying in the hope of reselling to other people at a

1. Inflation, risk, and taxes complicate the relationship, but in this chapter I will ignore these factors.

higher price, who are buying in the hope of reselling at a higher price to still other people, and so on. People are willing to be suckers in the hope of finding bigger suckers, and they can succeed so long as prices keep rising fast enough.

A chain letter is a classic example of a bubble: you pay money for something essentially worthless—membership high enough up in the chain—in the hope that people below you in the chain will do the same. Chain letters and other bubbles can keep growing and keep making money for some of their participants so long as new blood is continually being added.

Bubbles have occurred at some time or another in the markets for practically every imaginable asset, from tulip bulbs in seventeenth-century Holland to Internet stocks in late 1990s America. Land markets are no exception. In the 1980s, for instance, the famous bubble in Japanese real estate sent prices to dizzying heights—at one point it was estimated that the land under the imperial palace in Tokyo was worth more than all the land in California.

Bubbles, though, are evanescent, as the name implies. Eventually they burst. As a bubble in economics terms matures, it can continue only if it finds more and more investors willing to pay prices higher and higher than fundamentals justify. Sooner or later the supply of ever more gullible suckers runs out (or, more precisely, the belief of rational investors that there are more people to sell to runs out).

Because bubbles burst, and are probably rare, I will not pay attention in this book to their role in determining land prices. But in analyzing any particular incident—for instance, the American real estate boom at the beginning of the twenty-first century—you should be aware of the possibility of a bubble.

II. Location Theory Basics

A. Location Theory When Everyone Is the Same

As usual, it is best to start thinking about a complex, difficult subject by breaking it down into simple problems that are so easy and unrealistic that our little minds can grasp fully what is going on.

Suppose there is some place that everybody would like to be close to—Disney World, midtown Manhattan, or the Golden Gate Bridge, for instance. I don't care why they want to be close to it—maybe because their jobs are there and they have to commute every day, maybe because they like to look at it or listen to the foghorns, maybe because they are store owners who want their business to be near a major train terminal.

Assume everyone is identical and that the amount of space that each uses up is the same and fixed. Everybody, say, lives in a trailer of the same size. Then everybody can't be right next to the place they all want to be near; some have to be far away. You can figure out how many people can locate within 300 meters, how many within two kilometers, how many within five kilometers,

and so on, and that's all that can fit. If you know the size of the population you're dealing with, then you can figure out how big the area is that they have to occupy, and the land mass that they should occupy is the land mass of that size closest to the site.

Assume everyone is identical, too, in how much they like being near the site in question. In other words, everyone is willing to pay the same amount to be another kilometer closer to the magnet site. To be concrete, say people are willing to pay $40 a month extra to be a kilometer closer.

If every possible site is owned by landlords who charge what the market can bear, it is easy to see what will happen. Rents are going to have to decline by $40 a month for every kilometer in distance from the magnet site. Why? Suppose some location three kilometers from the magnet site is more than $40 a month more expensive than locations four kilometers from the magnet site. Then everyone will prefer the four-kilometer site to the three-kilometer site. More people than can fit will want to be at four kilometers; this will bid up the price there. Simultaneously the landowner at three kilometers will find no takers for her location; she will cut the price she is asking. The opposite happens if the rent at three kilometers is less than $40 more than the rent at four kilometers. The only way to prevent bidding wars, changing prices, and unwanted vacancies is to have rents decline by $40 for each additional kilometer from the magnet site. Under these conditions, everyone is indifferent about where he locates. Closer locations offer better access but require higher rents; distant locations the reverse. The rate at which the rents decline per kilometer is called the **rent gradient.**

The people who end up furthest from the magnet site have to pay essentially nothing. Suppose the landowners at the farthest occupied locations charged a positive amount. Then the landowners a little farther out could charge a smaller positive amount, less than what the other landowners were charging minus $40 a kilometer, and suck away all the demand from the inner landowners. Then the farthest occupied locations wouldn't be either farthest or occupied any more. Since there must be farthest occupied locations, rent must go to zero there.

If the population we are concerned with is fixed, these two conclusions let us figure out what rent will be at every location. The population size tells us where the farthest locations are; rent is zero there; rents rise by $40 a month for each kilometer you get closer to the magnet site. If the magnet site is somewhere in the middle of the occupied area, a graph of land rents looks something like Figure 6.1.

Displacement cost in the figure refers to the loss people suffer from not being right at the magnet site—the cost of commuting, in the most common interpretation.

Figure 6.1 should look quite familiar; it's very similar to Figure 4.4 in the discussion of congestion pricing. Distance is on the horizontal axis instead of

Figure 6.1 Land
rents around a mag-
net site.
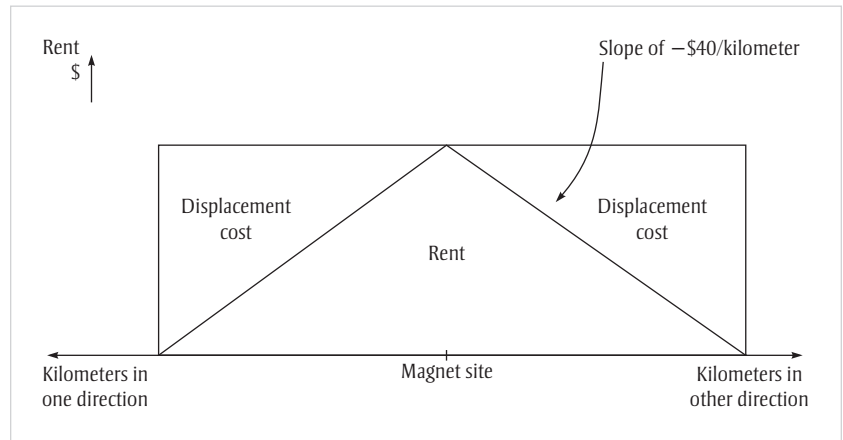
time; "rent" appears instead of "toll"; "displacement cost" instead of "schedule delay cost." But these differences are mainly cosmetic—location theory is just congestion pricing with some of the labels switched around. Location theory is about rationing space; congestion pricing is about rationing time; all of the basic principles are the same. (Indeed, one way to set up a system of dynamic congestion pricing would be to give each minute at the Lincoln Tunnel to a different "time owner"—say, a different nonprofit organization—and let them set prices to clear the market.)

Rents thus have the same kind of interpretation that congestion tolls have: they are what you pay because of the costs that you impose on other people. If the person located two kilometers from the magnet site were to be vaporized, everybody located farther out could move in a little bit and reduce their displacement cost; the rent paid by the person at two kilometers is exactly equal to what the people farther out would be willing to pay in the aggregate to have him vaporized.

Like congestion pricing, location theory makes a distinction between situations in which the population is fixed—demand for living in the city is inelastic—and situations in which the population can vary—demand is elastic. In location theory, perfectly inelastic demand is called the **closed-city case;** it assumes that people can't migrate either into the city or out of it, and so when conditions in the city improve, the general population benefits. Perfectly elastic demand is called the **open-city case.** In the open-city case people can migrate, and so if the package of access to the magnet site, rent, and displacement cost is better than other packages offered elsewhere in the world, people will migrate in and population will grow. If it's worse, population will shrink.

For most questions, both the closed- and open-city assumptions are extreme and unrealistic. People migrate into and out of cities all the time. Kids grow up and move away; older people move to retirement communities;

and people of working age seek better opportunities. So the closed-city assumption is unrealistic. On the other hand, many people are not completely rootless: where they are living now has many sentimental associations for them; they know how to get around and who the best plumbing contractors are; their families and friends are located nearby, and so are the cemeteries where their parents are buried. A 1 percent difference in land prices will not get people like this onto the plane for Australia tomorrow morning. So the open-city assumption is unrealistic too.

Why do we use these extreme and unrealistic assumptions, when in any given time period the demand elasticity is certainly some intermediate value? Because they simplify the analysis and because for many problems they're pretty good approximations. We saw in Chapter 4 how much easier it was to analyze the extreme elasticity cases instead of the intermediate case.

When are these assumptions good approximations? When you want to think only about what happens in the next few years and about small changes, the closed-city case is a good approximation. Raising the subway fare by 2 percent is not going to have a big impact on a city's population, and even major changes are not going to have a major impact on migration right away. The open-city case is a good approximation for large changes and long periods of time. Over decades, no matter how much you love a city, mortality guarantees that you will leave it, and the cumulative effect of small annual rates of migration becomes immense. The open-city assumption is also a good approximation for changes that affect only a small part of a large city.

How are land rents determined in the open-city case? The basic two conclusions don't change—land rents decrease by $40 a month for each additional kilometer from the magnet site, and rent at the farthest occupied location is zero. But we don't know the population, so we can't start off knowing where the farthest occupied location is. In the open-city case, instead of working from the outside in, you work from the inside out.

Think about somebody living right next to the magnet site; they have no displacement cost whatsoever. Rent at this location has to be just enough, considering how attractive the magnet site is, to make somebody indifferent between being there and being someplace else in the world. If the rent is higher than that, people will migrate away and no one will visit the magnet site at all; if rent is lower than that, people will flood in from around the world and bid rent up. So in the open-city case, the relative attractiveness of the magnet site determines rent at the closest location. Rent decreases at $40 a kilometer as you move away from the closest location and finally falls to zero at some point. That's the farthest occupied location, and it determines what the population is.

In either the open-city or the closed-city case, the greater the population that wants access to the magnet site, the higher the rents at the locations right next to it (the more people who would gain from your being vaporized, the more you have to pay to maintain your location). Traditionally, the main busi-

ness district of a big city is thought of as the magnet site. Then location theory explains why rents are higher in cities like New York than in cities like Omaha.

Of course, the higher rents in New York raise an interesting question: Why doesn't everybody leave New York and move to Omaha? There can be only one answer: because as a magnet site Manhattan is more attractive—wages are higher and there's more to do at night. Why are wages higher? Not because rents are high—if that were the reason, firms would just move to Omaha and their workforce would follow them. Wages are high in Manhattan because work in Manhattan is very productive—for all the reasons discussed in Chapter 2. Thus you shouldn't be appalled to learn that a one-bedroom apartment rents for $3,400 a month in Manhattan; you should be delighted that the wonderful chemistry of city life is so potent that people are willing to pay such apparently absurd prices.

B. Different Displacement Costs per Kilometer

Putting more realism on the bare bones of this outline of location theory is not too difficult. Suppose people or businesses differ in how much they are willing to pay to be close to the magnet site; some are willing to pay $40 a month to be a kilometer closer and others are willing to pay only $20 a month. Some households, for instance, might have two people making trips rather than one; or some people might value their time more highly or have worse modes of transportation or just hate driving.

Congestion pricing theory tells us what will happen: high-displacement-cost people will locate close to the magnet site; low-displacement-cost people will locate farther away. Figure 6.2 illustrates this profile of rents: it gets flatter as you move farther away. This is just like Figure 4.9. In the region where high-displacement-cost people are located, rents go down by $40 a kilometer; in the region where low-displacement-cost people are located, they go down by $20 a kilometer. Notice that this arrangement is efficient; no Pareto improvements are possible. But if a high-displacement-cost person were in the low-displacement-cost region (or vice versa), a Pareto improvement could be made: a low-displacement-cost person who was located closer to the magnet site would require only a bit more than $20 a kilometer to be persuaded to change places, but the high-displacement-cost person would be willing to pay a bit under $40 a kilometer. In this situation land prices guide activities into the right places for them.

If there are many different groups, instead of just two, then the groups line up in order of displacement cost: those with the highest displacement cost per kilometer on the inside, then those with the next highest displacement cost, and so on. Rent gradients look like Figure 6.3: the gradient is steepest close in and gets flatter and flatter as you move away from the magnet side. Once again the arrangement of groups is efficient.

Actual land values in Manhattan in 1903 followed a pattern very much

Figure 6.2 Land rents with two classes of users.
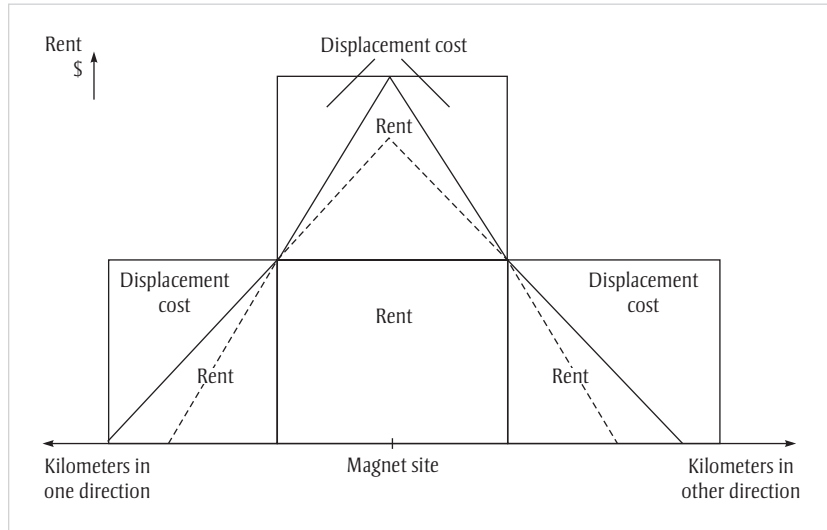
Figure 6.3 Land rents with many classes of users.
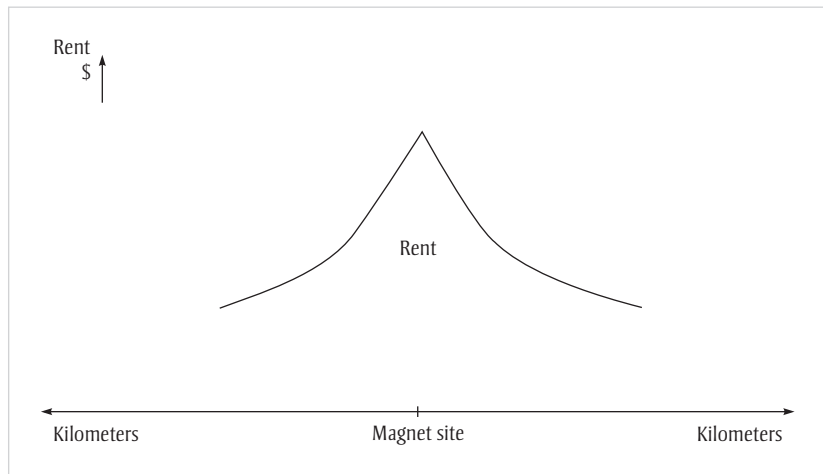
like that predicted in Figure 6.3. The map in Figure 6.4 illustrates (from Hurd 1903; cited in Willis 1995). Prices were highest ($400 per square foot) at the New York Stock Exchange and fell by more than over $100 a block in that neighborhood. Farther away, say, by the Brooklyn Bridge, a distance of several blocks changes the land price by only $5.

This theory about how different activities arrange themselves also implies segregation: all of the activities with the same per-kilometer displacement cost will locate in the same region, and none will be outside this region. Stockbrokers will be in one area; warehouses in another.

Notice that groups with the highest displacement cost per kilometer pay

[To view this image, refer to
the print version of this title.]

Figure 6.4 Land values in Lower Manhattan (1903), in dollars per square foot. From Carol Willis, *Form Follows Finance: Skyscrapers and Skyliners in New York and Chicago* (New York: Princeton Architectural Press, 1995), p. 172. Used with permission.

very high rents. What happens if they can't afford these rents? Is the allocation still efficient? The answer sounds harsh but is not illogical: those activities end up not happening, and this is efficient. The analogy is the situation with elastic demand in which, when tolls are optimal, some trips through the Lincoln Tunnel that people think about making and might have made under other circumstances don't get made.

To see this, consider a seashore town. Rents will decline as you move away from the oceanfront, and at the oceanfront will be the activities with the highest displacement costs—things like marinas (carrying boats through the streets is costly), hot dog stands, and fancy hotels; convenience stores will be a little farther away (carrying suntan lotion is easier than carrying either a boat or a hot dog), and then residences. Another activity that has very high displacement costs is freezing ocean water and letting people ice skate on it. If you were going to set up an operation like that, there would be great advantages to having it as close to the oceanfront as possible; every meter you were away from the oceanfront would cost you a lot. But your locating near the oceanfront imposes costs on lots of other people. So oceanfront ice skating is something that doesn't happen. Since the meager net benefits you and your customers would get from such an endeavor are not enough to justify the costs it would impose on other activities displaced from the oceanfront, it's not something that should happen, either.

This sounds fine for businesses and business activities, but what about people and places to live? To say that people should disappear is certainly not the same as saying that oceanfront ice skating should disappear. But the argument is not about disappearance or appearance—it's about access to a particular magnet site. There are many places to live—remember that most land is essentially worthless—and many magnet sites. Not having access to a particular magnet site and not living are different propositions.

There is a traditional view of how cities are arranged that grows out of this idea that activities should be arranged in order of decreasing displacement cost. Think of the center of a city as the magnet site. This is sometimes called the **central business district,** or CBD. This area will be filled with those whose activities have the highest cost of displacement from the center of activity: high-powered executives, lawyers, and wheeler-dealers who rely on face-to-face meetings but whose time during working hours is so valuable that they cannot go traipsing far afield to have those meetings. Here too would be restaurants and hotels to facilitate those meetings, and people who have to help the high-powered folk: accountants, secretaries, junior lawyers, waiters, cooks, chambermaids.

In the late 1930s, one study explained why the financial industry was concentrated in Lower Manhattan:

> Dependence on network and telegraph, telephone, and cable lines coming
> to a central focus has tended to tie the financial district to one spot. It is im-

portant that financial institutions remain close to the shipping lines, Sub-treasury, Federal Reserve Bank, Customs House, Clearing House, and then principal depositors such as railroads, utilities, as well as other banks so that business may be transacted speedily and directors conveniently reached for meetings . . . Coffee, sugar, cotton, and cocoa markets are required by exchange rules to have bank "margin depositories" within a given radius for accepting down payment against "future" trades and hence these exchanges must stick close together in a limited area. (Armstrong and Hoyt 1941, p. 125; cited in Willis 1995, p. 172)

Since the CBD has the best transportation connections and the heaviest concentration of foot traffic, it's also the place for major department stores and civic centers (because of the high volume of traffic); moving them away is costly. Slightly farther out from the hub of activity would be manufacturing and shipping operations. Some of these establishments need to be near the main freight terminals to get their inputs from the rest of the world or to send their outputs there; others need to be near the wheelers and dealers.

Farther out come residences, and the jobs that support residences. Being closer to the CBD is better for the people who live here, since it cuts down their commuting time. You would expect two-earner families to live closer than one-earner families, since their commuting costs per kilometer are twice as great, and you would expect people like police officers and firefighters, who do a lot of their commuting at weird and therefore uncongested times, to live farthest away. Some manufacturing might also be in the residential zone—the sort of manufacturing in which it's cheaper to ship output to the CBD than to move workers there.

This traditional view of how cities work is called the **monocentric city model.** It was probably a very good model around 1900, when railroads and streetcars were the dominant mode of transportation. In a world with trucks, cars, and airplanes, cities are unlikely to have a single center. But CBDs are still important centers of employment and commerce, even if airports and interstate highway intersections also act like magnet sites. Land generally gets less valuable as you move away from Manhattan, but not without interruptions caused by airports and suburban office centers.

C. Different Land Intensity

When we studied congestion pricing, we saw that if every commuter had the same costs per minute for being late or being early to work, it made sense to schedule buses and high-occupancy vehicles to go through the Lincoln Tunnel closer to the most desirable time rather than regular cars, and sports utility vehicles to go through further from the most desirable time than regular cars (see Chapter 4, question 4). A minute's delay cost the passengers on a forty-passenger bus forty times as much as it cost the passenger in a one-passenger

car, and an SUV takes up more road space than a regular car does and so delays the cars behind it by more.

Just as people use different kinds of vehicles to get through the Lincoln Tunnel, so too do they use differing amounts of land to carry out the activities they are interested in. For housing, some people live in rambling ranch houses with half-acre lawns; others live in townhouses flush with the sidewalk, and share walls with their neighbors; still others live in fifty-story apartment buildings and share the same plot of land with hundreds of other households. Some universities have hundreds of hectares of parking lots and one-story classroom buildings; other campuses are united around subway stations and waiting for elevators. A Home Depot uses a lot more land than a French restaurant; a steel plant uses more than an office building.

Everything else being equal, then, we would expect to see land-intensive uses like Home Depots and golf courses on the outskirts of metropolitan areas, and land-saving uses like apartment buildings and multistory office buildings close to the CBD. This is efficient, and the land rent gradient makes it happen. It's another reason, too, why land rents decline more steeply at the center than at the periphery: moving a fifty-story apartment building out by a kilometer adds a kilometer to the commute of hundreds of workers; moving a ranch house out by a similar amount affects only a worker or two.

These relationships hold no matter why it is that some people live in apartment buildings and some live in ranch houses. Maybe some people are just born recluses and others are born extroverts; the recluses will live at the periphery in ranch houses and the extroverts at the center in apartments. At the other extreme, maybe everyone is identical, and all find some kind of sufficiently fancy apartment just as good as every imaginable kind of ranch house. Then apartments and ranch houses will be built so that no one prefers one to the other. Once again apartments will be in the center and ranch houses on the periphery.

In both cases, if you plot persons per hectare against distance from the city center, you get a line that slopes down. Such a plot is called a **density gradient.** When you can use varying amounts of land for a housing unit, you also have to make a distinction between the land rent gradient—what a hectare of land costs—and the residential rent gradient—what a standardized house costs. The residential rent gradient slopes down because of commuting cost, and the land price gradient slopes down because of both commuting cost and differing land-use intensity.

Understanding the mechanism that makes people use more valuable land more intensely is important. The key is the ability to substitute land for capital (and vice versa) and the declining marginal productivity of each. (This is an example of factor substitution, from microeconomic theory.) Suppose you are thinking about constructing an office building. If you spread out and use a lot of land, you won't have to use a lot of capital: you won't need to install a lot of

fast elevators; walls and steelwork won't need to be strong enough to support a lot of weight; you may not even need to dig foundations.

If, however, you build a tall building on a small plot of land, you will have to use more capital. A 1930 study of office buildings in New York, for instance, found that it took $9.10 in building costs to construct one square foot of net rentable space in a fifteen-story building, $12.60 for one square foot in a seventy-five-story building. The big differences were in the cost of structural steel, brickwork, elevators, and heating and ventilating (Clark and Kingston 1930, table 10; cited in Willis 1995, p. 84).

The cheaper land is, the more of it you will use for whatever you want to do, and the less capital you will use. There's no sense investing in something like a bank of elevators when you can spread out over more land instead. Expensive land, therefore—land close to the center of a city—will be covered with tall buildings; cheaper land will have lawns and low buildings.

The ability to choose different land-intensities of development is a separate reason why land rent gradients get steeper in the monocentric city model as you approach the CBD, and flatter farther from it. Suppose all consumers of land are identical. If everyone has to live in the same style of single-family house and use the same amount of land, the land rent gradient will have the same slope everywhere. Now give every landowner the option to build an apartment house or a two-family house instead of the single-family house of the required style. This option increases the value of land only where it is likely to be exercised—which is close to the CBD. The increasing value of the option to raise land intensity thus increases the value of land close to the CBD more than it increases the value of land far from the CBD. Land would not be anywhere near as valuable as it is in most CBDs unless you could build skyscrapers there—and nobody would want to build skyscrapers there unless land prices were that high.

Thinking about these different gradients is often helpful in thinking about the implications of various policies. The answers are not always the first ones you think of (see Box 6A).

Differing land intensity in housing is one reason why rich people and poor people often live in different parts of a metropolitan area. Rich people consume more land than poor people do—just as they consume more of practically everything—and so we would expect to find them on the periphery. That's how things generally work out in the United States. But land consumption isn't the only determinant of location—displacement costs matter, too. Since rich people generally have higher wages than poor people, their commuting time is generally more expensive. This should make them, rather than poor people, live closer to the CBD (just as, in Chapter 4, rich people used the Lincoln Tunnel at the best times when only schedule delay costs differed). So displacement costs and land consumption work in opposite directions; location theory alone can't predict that rich people always and ev-

See for Yourself

≡ Suppose that a city restricts the height of residential structures (Washington, D.C., for instance, prohibits any buildings taller than the Washington Monument). The maximum height is set at five stories, the height that would normally occur at five kilometers from the city center. Draw two residential rent gradients, one for the city in the absence of height restrictions, and one for the city with height restrictions. Use the open-city assumption. Who gains from the height restriction? Who loses?

Under the open-city assumption, residential rent adjusts to ensure that no location in the city is any more attractive or less attractive than the other occupied locations in the world. Consider a location more than five kilometers from the city center. The height restriction does not affect the attractiveness of this location at all—it still requires the same commute to the same job opportunities. Nothing changes in the rest of the world, either. So the rent that tenants here pay cannot change.

Now consider a location less than five kilometers from the city center. From the same argument, the rent tenants pay here cannot change either, since neither commuting costs nor options in the rest of the world change. Thus the residential rent gradient is the same without the height restriction as it is with it.

The people who lose from the height restriction are those who own land less than five kilometers from the city center. The rent per unit that they can collect stays the same, but the restriction keeps them from adding floors that would be profitable at those rents. Owners of land more than five kilometers from the city center gain or lose nothing, since the rents they collect stay the same and the restriction doesn't prevent them from doing anything they want to do. The land rent gradient gets flatter within five kilometers of the city center, even though the residential rent gradient stays the same.

Nobody gains. The height restriction makes the city's population smaller by preventing people who could contribute something from living there.

erywhere will live far from the CBD. It's not surprising that in most of Europe and in New York City, very rich people live close to the center of town.

III. Transportation and Other Improvements

A. How Transportation Improvements Affect the Land Market

Now we can see what happens to density and land rents when transportation improves—when cars are invented, for instance, or when New York City and

the transit companies build a subway system. This is useful for understanding the history of how cities have changed, and for assessing the benefits and costs of prospective transportation projects.

Better transportation means lower displacement costs per kilometer: if the cost of traveling a kilometer falls by 50 percent, the displacement cost you used to incur at ten kilometers you now incur at twenty (assuming transportation is the only displacement cost). So the rent gradient will flatten; the rate at which rents decline will decrease. A one-kilometer difference in distance from the magnet site will no longer imply as big a difference in rent.

This flattening of the rent gradient happens in both the open-city model and the closed-city model. But the implications are very different. In the closed-city model, the city's boundary is fixed by the available population, and so flattening the gradient means lowering rents everywhere (except at the boundary). In the open-city model, rent at the magnet site is fixed by the relative attractiveness of the magnet site, and so flattening the gradient means raising rents everywhere (except at the magnet site). I need to elaborate on these conclusions and qualify them a little.

In the closed-city model, when the amount of land that each business and person uses is fixed, the size of the city is fixed; changes in transportation cost won't change how many people commute or how far they commute. Each person's residence becomes less of an obstacle to those who live farther out, and so what she should pay goes down. Similarly, what people on the outside are willing to bid to move to the inside goes down. At the extreme, if transportation cost went to zero, nobody would pay anything for land; any site would be just as good as any other site, and no one would be getting in anyone else's way. Figure 6.5 illustrates what a transportation improvement does to land rents in the closed-city case.

This conclusion has to be modified slightly when people can choose the intensity with which they use land—when they can decide whether to have a lawn or a stereo, privacy or a doorman, a big supermarket with a single worker at each checkout counter or a small store with two workers each at half as many checkout counters. Then the lower price of land will cause people and firms to use land more intensively and spread out more. Land prices will rise relative to what they would have been with the transportation improvement but without the option to increase land intensity, but not enough in most locations to get back to pre-improvement levels. The city boundary will move out, and the inner areas will become less dense.

This is probably the model that public health reformers had in mind when they supported the New York City subway in the late nineteenth and early twentieth centuries. Many poor people still walked to work then and so had to live very close to their jobs. As a result, Lower Manhattan was jam-packed with workers, many of them immigrants. The Lower East Side in 1910 had 700 people per acre, the greatest residential density ever recorded. Such

Figure 6.5 Land rents with better transportation in a closed city.
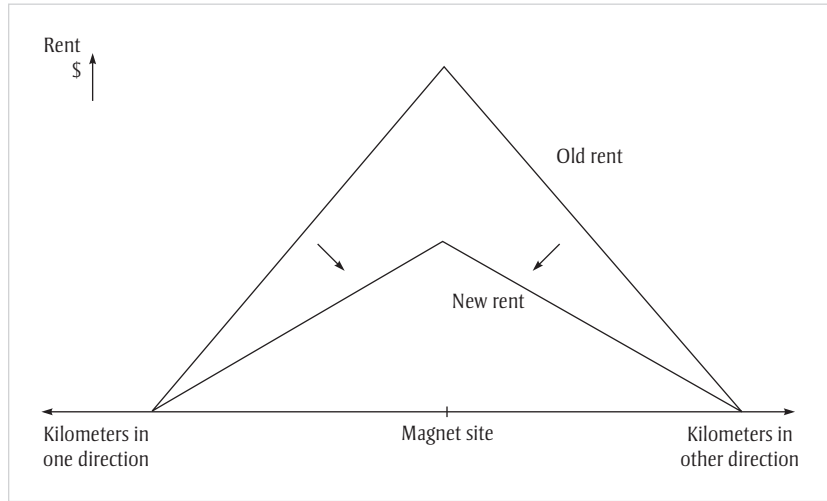
densities appalled the public health movement, and reformers looked at the
subway as a way to move people out of the Lower East Side and away from the
pathologies (moral as well as physical) that they saw lurking there. Subways
would substitute sunshine for slums.

The open-city model, on the other hand, predicted that subways would
make conditions on the Lower East Side worse. First, consider what happens
with the fixed land intensity. At any distance from the magnet site, this city is a
better place to live than it used to be—everything else is just as good, and
transportation is even better. Since it used to be just as good a place as every
place else in the world, this makes it better than other places now. So people
flood in from other places and bid up land rents. Rents rise until no location is
any more attractive than it was before transportation improved. The city
boundaries expand, too, because locations previously too far away from the
magnet site to command positive rents are now close enough that people are
willing to pay something to live there. Figure 6.6 illustrates.

People's ability to change the intensity with which they use land modifies
these conclusions slightly. When land prices go up, people will find ways to
economize on land; this will mean that the original rent hike will hurt even
less than it would if people couldn't adjust. So land rents must rise even more,
and the population must rise even more. Thus in the open-city model the
subway would make conditions on the Lower East Side worse—more immi-
grants would flood in, rents and density would increase. Queens and the
Bronx would be developed, but so many more people would migrate to New
York that it wouldn't make a difference.

The open-city model is probably what the outer-borough developers and
Manhattan business interests had in mind. The outer-borough developers ex-

pected a much larger population boom than the closed-city model would imply, and the Manhattan business interests surely did not expect the subway to reduce Manhattan property values. The supporters of the subway were expecting it to bring different, and indeed incompatible, results.

Who was right? The time since the subway was constructed has been long enough and mobility in the United States (indeed, in the world) is great enough that the open-city model has to be the appropriate interpretation of what happened. The outer boroughs filled up, and Manhattan property values went up. New York City's population rose from 4.8 million in 1900 to 7.6 million in 1940. It looks like the developers and Manhattan businesses were right and the public health reformers were wrong. But at the same time, densities in Manhattan fell, from 161 people per acre in 1900 to 130 in 1940 and 105 in 2000. Part of the story is that commercial activity displaced some residences, but even in neighborhoods that remained always residential, density fell. Does this mean that the closed-city model is really appropriate?

Probably not. Here is my interpretation: In the thought experiment we did to develop the open-city model, we let transportation improve in one city and kept everything else everywhere the same. In particular we assumed that every place else stayed just as attractive as it was before transportation improved. But if you think about the subway and a period like 1900 to 1940, this assumption is clearly untenable. Life outside New York City clearly got better during this period. Real gross national product per capita rose 70 percent in the United States, and the income of a typical urban wageworker or clerical worker's family more than doubled (U.S. Bureau of the Census 1972, tables F1-5 and G582-601).

Life in New York City clearly got better, too; otherwise it would have been deserted by 1940. Life got better through higher incomes, better sanitation,

Figure 6.6 Land rents with better transportation in an open city.
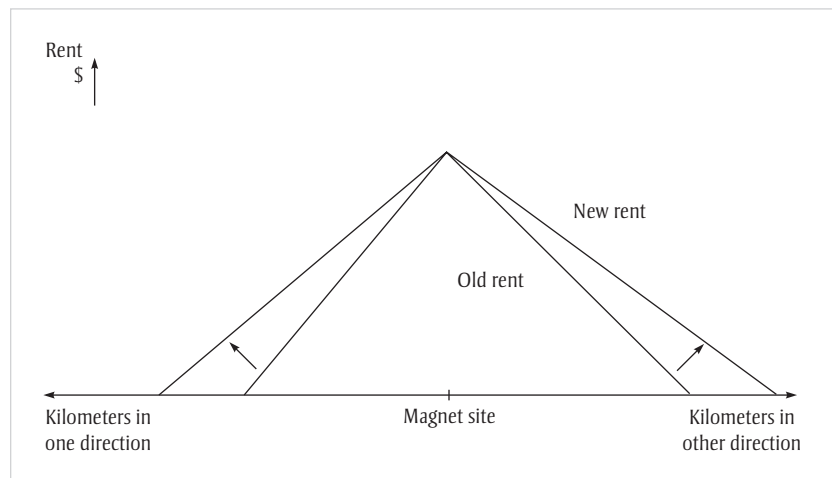
and nicer housing, and part of the package of nicer housing and better sanitation was lower density. In my view, the public health reformers were really wrong about the subway: it didn't do anything to reduce population density in lower Manhattan. But they got lucky. Income rose in the rest of the United States and the rest of the world and did for them the job they hoped the subway would do.

Rising incomes outside New York City also indicate that the subway probably affected wages in New York City, too. In the open-city model, the subway would have made New York's population and labor force bigger even if wages didn't change. A bigger labor force, however, implies more opportunities for increasing returns to scale, and if these opportunities are important, as I argued in Chapter 2, then marginal product and wages should rise. Higher wages cause a further boost to land rents, but this boost does not necessarily increase density, since part of the income may be spent on more land and more privacy.

In the same way we can also analyze the difference in the effect of cars versus mass transit. Some people think that mass transit makes a city denser than highways do. Is there a basis for this belief? In the open-city model, the answer depends on how well mass transit works. If mass transit is better than car commuting—faster, cheaper, more convenient, more comfortable—then it will raise land prices and induce greater density. If it makes commuting better only for people who live close enough to the stations, then it will increase density only near the stations and not elsewhere. If mass transit isn't better than cars, it won't increase density—it will just make a city smaller.

In the closed-city model, mass transit will increase density only if it's worse than cars. But in a closed city, anything that makes transportation worse increases density, and building a subway is one of the more expensive ways to make transportation worse.

Thus, under certain circumstances mass transit increases density. But it isn't magic. Empirical evidence on this question is sparse, and mixed. Landis and Cervero (1999) concluded that after twenty years, the BART system in California's Bay Area had possibly spurred office development in downtown San Francisco but had not increased business or residential densities elsewhere. Bollinger and Ihlanfeldt (2003) found that employment in Atlanta neighborhoods that were closer to stops on the newly opened subway didn't grow faster.

B. Measuring the Benefits of Better Transportation and Other Improvements

This analysis gives us the tools to examine the question of whether changes in land prices should be counted among the benefits of transportation improvements. People often argue in favor of transportation improvements and other policies by saying that they will raise property values, but it's not always clear that raising property values is a good thing. If all we pay attention to is land

rent, for instance, then something so clearly desirable as a magic invention that reduces all transportation costs to zero looks like a disaster, since it drives all land rents to zero too. Similarly, when life in the rest of the world becomes more attractive but nothing changes in New York, land prices in New York go down, even though nothing has become worse.

We need to be very careful when thinking about the ethical significance of the land-market repercussions of changes in transportation. The best way is to proceed step by cautious step. Remember that the ultimate goal is to find out which transportation enhancements are potential Pareto improvements—that is, which ones generate more gains to their winners than losses to their losers. Transportation enhancements do two things: they change the cost of various kinds of trips, and they change the rents of various parcels of land. We want to examine the combined effect on different people of both changes.

Let's start with the open-city model because it's simpler. Ordinary consumers who don't own land are faced with two changes: the cost of some kinds of trips goes down, while rents they have to pay for land go up. (In an open city, transportation enhancements may also make a city more productive by raising its population, but we'll ignore that effect in this section.) The rent increase, we know from the construction of the open-city model, is exactly the right size to leave every consumer just as well off after the transportation enhancement as she was before it. What better transportation gives, higher rents take away. So ordinary consumers without land are neither winners nor losers; they are willing to pay nothing for the transportation enhancement, and they do not need to be compensated for it.

Thus in the open-city model all of the benefits of a transportation enhancement accrue to landowners (just as in congestion pricing with infinitely elastic demand, all of the benefits accrue to the toll collector). Landowners receive higher rents, and these rents are capitalized in the form of higher land values. If you can figure out how much a transportation enhancement raises land values, you know what its benefits are. A transportation enhancement that raises land values by more than it costs is a potential Pareto improvement; one that does not, is not.

Notice that I am *not* saying that the benefits of transportation enhancement are the *sum* of the benefits of lower-cost trips and higher land values; the benefits are higher land values *only*. One way to understand why this is so is to realize that higher land values are good for owners of land but bad for consumers of land. Here is the basic equation for the open-city case:

benefits of better transportation
    = gains to consumers of better transportation, holding land prices
       constant
    = loss to consumers from higher land prices, given better
       transportation
    = gain to landowners from higher land prices.

The gain that consumers get from better transportation is exactly offset by the loss they suffer from higher land prices; what's left over is the gain to land-owners.

Is there any sense, then, in estimating anything about how consumers value better transportation? The answer is often yes, because this number can provide a good approximation of what landowners will end up gaining (since it is the source of their gain), and it may be easier to estimate prospectively than the gain to landowners. Given that the transportation enhancement is in place, the loss to consumers from the rise in land prices exactly equals the gain to landowners. So net benefit can be thought of as what's left over: how much consumers would be willing to pay for better transportation if they had assurances that land prices wouldn't rise. Often it is easier to guess what this number will be for a transportation project that hasn't been carried out yet than it is to estimate what will happen in the land market. But once you've estimated what consumers would be willing to pay for better transportation in the absence of increases in the price of land, it's not legitimate to add the gains that accrue to landowners from higher land rents, since those gains are made directly at the expense of consumers.

The closed-city case is also worth considering, since a short-run analysis is sometimes appropriate and since sometimes migration restrictions are in place. In the closed-city case, consumers gain twice: once from better transportation and once from lower land prices. But consumers' gain from lower land prices is exactly offset by landowners' loss. Thus in the closed-city case the net benefit of a transportation enhancement is what consumers would be willing to pay for better transportation, holding the price of land at its old (high) level. For the purpose of finding net benefits, you can ignore the land market.

Similarly in the open-city case, a transportation enhancement in one city reduces population and land prices (ever so slightly) in all other cities (relative to what it would have been without the transportation enhancement). But the loss to the landowners in those other cities is offset by the gain to consumers, and so you can ignore it in calculating net benefits.

On the other hand, if you're looking for potential Pareto improvement you can look at either the transportation market or the land market but not both, but if you're looking for political support, both markets are relevant. Here the difference between transportation enhancements that are reasonably permanent, like roads and subway stations, and those that might only be temporary, like bus routes, is very relevant. A permanent enhancement increases (in the open-city model) the annual rent of land when the land market adjusts, but it raises the value of land—what you can get for selling your land— almost immediately (even before construction is complete). This is because the value of land is the expected present value of future rents. So land values go up before rents rise.

Damm et al. (1980), for instance, found that Washington, D.C., property

values went up in areas where Metrorail service was anticipated, even before the system was constructed. Other studies elsewhere also found that rail access was capitalized into real estate values: see Gatzlaff and Smith (1993) for Miami home prices, Voith (1993) for Philadelphia home prices, Cervero (1994) for Atlanta and Washington office buildings, and Baum-Snow and Kahn (1998) for homes in several cities; Cervero and Landis (1999) found mixed evidence for the impact of BART.

Consumers, too, benefit from better transportation right after the transportation enhancement goes into effect, although when the land market adjusts these benefits will vanish. So for a transportation enhancement that looks to be permanent, political support right at the start is much greater than what would be justified by long-run benefits. Since short-run political support may matter more for government officials than what people will think in thirty years, permanent transportation improvements may be unduly favored in the political selection of projects.

Finally, transportation enhancements aren't the only policies that change land values, and so they aren't the only policies to which these lessons apply. Sewers, baseball stadiums, schools, police stations, and parks all change land values. In the open-city model especially, practically anything that makes any piece of a city either more or less attractive gets capitalized in land values. This includes policies that are not about bricks and mortar. Abolishing slavery, for instance, made land in the Mississippi Delta less valuable, and allowing gambling made land in Atlantic City more valuable. We should think about these changes in land values the same way we think about changes caused by transportation improvements.

IV. Land Taxes

Another question left unresolved in the discussions of transportation thus far is how to pay for the roads, buses, and trains that would not pay for themselves with optimal tolls or fares. Similarly, we need to know who should pay for schools, parks, and police stations. We know now who benefits from these things, at least in the long run—landowners. This raises the question of whether landowners should make up the associated deficits. There is no rule of justice that says that people should be fully taxed for every benefit they receive; after all, if there were such a rule, we would all have to be totally miserable. Still, the greater the congruence between who a project's ultimate beneficiaries are and who pays for it, the more likely any potential Pareto improvement is to be an actual Pareto improvement, and the converse. Then projects with greater benefits than costs will be more likely to have unanimous support and so be undertaken, and projects with more costs than benefits will be more likely to be unanimously rejected.

Economists find other advantages to taxing land to make up transporta-

tion deficits. As I noted in discussing privatizing mass transit, most taxes cause greater losses to taxpayers than they raise in revenue for the government—taxpayers would be willing to pay more to the government to stop taxing them than the government is collecting. The more that people change their behavior because of the tax, the greater is this deadweight loss. By "change their behavior" I mean consume less of the taxed commodity.

But think about a tax on land. Land of all kinds is in essentially fixed supply. The amount of land within a three-kilometer radius of Times Square won't go down because it's being taxed. So land taxes have no deadweight loss. What taxpayers lose is exactly what the government gains. This is a strong argument for using land taxes rather than, say, taxes on income, improvements, luxuries, or imports.

When you consider these advantages, you should not be surprised to learn that land taxation has been a popular cause since the nineteenth century, both with economists and with a slice of the general public. The key name associated with this is Henry George (1819–1897). George was a sailor, prospector, printer, reporter, San Francisco newspaper editor and publisher, orator, and political activist. He knew poverty personally and saw widespread deprivation in the midst of a booming economy. His major book, *Progress and Poverty,* published in 1879, sold more than two million copies and became the most influential American economic treatise of the nineteenth century. He ran for mayor of New York in 1886 and finished second in a tight, three-way race (future president Theodore Roosevelt was third). More than a century after his death, Henry George's work is still being propagated by several organizations and foundations.

In *Progress and Poverty,* George asks why material progress has done so little to alleviate poverty. "When the conditions to which material progress everywhere tends are most fully realized—that is to say, where population is densest, wealth greatest, and machinery of production and exchange most highly developed—we find the deepest poverty, the sharpest struggle of existence, and the most of enforced idleness" (George 1900, p. 6). The problem is private ownership of land. This is unjust, he says ("Private property in land . . . cannot be defended on the score of justice" [p. 338]); it causes poverty because wealth is siphoned off to landowners; it discourages industry because governments are forced to use other taxes, and because speculators sequester usable land from the market; and it propagates panics and depressions by inflating speculative land bubbles that inevitably burst.

The solution George proposes is a "single tax" to take land rent away from private owners and give it to the community, which really earned it. George didn't want to dispossess landowners, only to take the full annual rental value of land as a tax. (Since the value of land is the capitalized value of land rent, taxing the rent of land and taking its value are equivalent.) This tax would replace all others. With a single tax, speculation, unemployment, and

depressions would end; capital and labor would be unfettered. "Wealth would not only be enormously increased; it would be equally distributed . . . The non-producer would no longer roll in luxury while the producer got but the barest necessities of animal existence" (pp. 452–453). The single tax would "raise wages, increase the earnings of capital, extirpate pauperism, abolish poverty, give remunerative employment to whoever wishes it, afford free scope to human powers, purify government, and carry civilization to yet nobler heights."

Notice that the single tax is really quite similar to the idea that the proceeds of congestion tolls on highways should be used for public purposes, not private gain. Most economists believe that there are at least serious efficiency arguments to be made for land taxes, even if few are convinced about its carrying civilization to nobler heights. Taiwan (Sun Yat-sen was an avowed Georgeist), Australia, New Zealand, and several cities in Pennsylvania all tax land more heavily than improvements. Hong Kong, where the government owns a great deal of the land and leases it out to private developers, might also be considered a something like a Georgeist jurisdiction.

Could a land tax really replace all other taxes? The obvious reaction is that a land tax could never raise enough money to fund the level of government activity we are used to. The federal Bureau of Economic Analysis estimates the developed land in the United States was worth about $5 trillion in 1987; undeveloped land was worth very little. Even if this value of land represented capitalization at a very high interest rate of 10 percent, total land rent would have been no more than $500 million a year—only half of government expenditures in that year.

This reaction, however, is wrong. The value of land is low because other taxes are high. Suppose that one state, say Illinois, abolished all taxes and all Illinois residents became exempt from federal taxation. This would make Illinois a very attractive place to live and work, and more people and businesses would move to Illinois. In the open-city model, they would keep moving to Illinois until land prices in Illinois rose so much that people in Illinois were no better off than people elsewhere. Thus people in Illinois after tax abolition would be no better off than people in Illinois before tax abolition. The rise in the value of Illinois land would be equal to the total burden of the taxes that were originally collected—the sum of government revenue and deadweight loss.

So a confiscatory land tax could eventually collect a great deal of the revenue that state and local governments are now raising from taxes that produce deadweight losses, if enough people believed that other taxes would never be reimposed. Land taxes could also replace federal taxes if immigration were unrestricted. This conclusion becomes more certain if we interpret "land" to mean any scarce, geographically specific resource, not just dirt—petroleum, rights to use parts of the electromagnetic spectrum, the right to immigrate to

the United States or to the European Union, or entry into the Lincoln Tunnel at rush hour, for instance. A single tax on land, broadly conceived, is not impracticable.

Would it be desirable? Here again George's arguments have considerable strength. Reducing deadweight losses is generally a good thing, and a tax on land would probably lighten the tax burden on poor people. But land taxation suffers from two closely related informational problems.

First, finding the value or rent of a piece of land at a particular moment is very, very difficult. You can find out today's price for a two-liter bottle of Diet Cherry Coke by going to the supermarket and seeing what people pay; so long as a few sales take place today you can get an answer. Similarly, if you want to know what two-bedroom apartments near Route 1 in Edison, New Jersey, rent for, all you have to do is examine this month's rent checks. But with land, unlike two-liter bottles of Diet Cherry Coke and two-bedroom apartments in Edison, for an average month there are very few transactions to observe. In a built-up urban area, most land changes hands only as part of a package that includes buildings as well as land—only rarely, for instance, does someone buy a house separately from the land that it sits on, or vice versa. The same is true for rentals. Trying to separate out the value of land from the value of the rest of the package is like trying to figure out the value of steering wheels from the prices of used cars. Under these circumstances, the rare sales of truly vacant parcels become very important in estimating the value of every parcel of land. Since land prices appear to vary greatly across both time and space, the sale of a vacant lot six months ago at 55th Street and Sixth Avenue may be a very poor indicator of the value of land today at 57th Street and Third Avenue; but it may be the only available indicator. Vacant lots, too, may be unrepresentative of the surrounding neighborhood—if the rest of the neighborhood is built up, there's probably some reason why they're vacant.

At the same time, because different pieces of land are often pretty good substitutes for one another for many purposes, the task of estimating land values is far from hopeless. The price of land on West 119th Street tells you a lot more about the price of land on West 120th Street than the price of light bulbs tells you about the price of yogurt. Often the price of land per square meter varies smoothly from place to place, and so knowing about a few prices lets you infer something about many more.

But prices of land per square meter don't *always* vary smoothly. Sometimes people have sentimental or religious attachments to particular places (think of Jerusalem or the footprint of the World Trade Center) and they are willing to pay large amounts to control those sites, and to keep others from using them inappropriately. Often the shape of a parcel matters—a 1-meter-wide, 300-meters-long curlicue is not going to be as valuable as a rectangle 10 meters wide and 30 meters long, even though they have the same area. Physical characteristics may matter, too: whether land is rocky or swampy or used

to be occupied by a gas station that let its underground tanks rot; whether snowmobiles are allowed to cross it or large apartment buildings may be constructed on it. If all you observe on a particular parcel is a restaurant that's phenomenally successful, how can you tell for sure whether that success should be attributed to a great location or a brilliant owner?

The second informational problem arises when the government uses information about what is actually going on to assess the value of land (it's hard to imagine that it could completely ignore it). Private-sector actors also have a hard time figuring out what the best use for a particular parcel of land is at a particular moment (and so they don't know "true" land values either). Putting the right kind of activity in the right place at the right time requires effort, persistence, skill, and some luck. The main reason that people make the effort that goes into this process is that they hope to get rich from taking advantage of some misallocation the rest of the world didn't see. People won't make this effort if the government is going to take all or most of their profits.

Thus if land taxes end up being a tax on developer effort, they have the same sort of distorting consequences most other taxes do. But that doesn't mean they shouldn't be used. It means only that land taxation is not a magic elixir.

We use medicines all the time that are not magic elixirs. And just as with any medicine we need to understand the possible side effects, the important questions about land taxes are how big the likely side effects are—how much less effort developers will exert, how inaccurately assessors will gauge the value of land—and how they can be alleviated. Only by answering these questions can we decide how large a role land taxes should play in an optimal fiscal system. It wouldn't be surprising if that role were considerably larger than their current role in most of the United States.

V. Conclusion

Urban land prices are determined by access: the easier it is to get from a location to some place that a lot of people want to go to, the more expensive is the land in that location. Land prices ration physical access the same way congestion pricing rations points in time. As a result, activities with higher transportation costs per unit distance tend to be close to the center of a city, everything else being equal, and those that use a lot of land per unit of activity tend to be far from it. If you can carry out the same activity (raise a family, for instance) with varying amounts of land, the closer you do it to the center of the city, the less land you will use (and the more of other kinds of inputs). This arrangement of activities cannot be improved upon in such a way that the losers from any change will lose more than the winners will gain.

Better transportation makes land-price and density gradients flatter.

With improved transportation in the open-city model, land prices go up and the entire benefit accrues to landowners; in the closed-city model, land prices go down and the entire benefit accrues to consumers. In the open-city model, which is probably the relevant one for this question, a mass transit enhancement will cause greater densities than a highway system, especially in the vicinity of stations, if in fact it moves people more quickly, cheaply, and conveniently to where they want to go than a highway system; otherwise it will not raise density. Higher income reduces density, since space is a normal good.

You can measure the benefits of better transportation in the transportation market or in the land market, but not in both.

Land taxes were a very popular idea in the late nineteenth century; Henry George's book sold more than two million copies, and he was almost elected mayor of New York. Land taxes have some significant advantages: they have no excess burden, and they tax the ultimate beneficiaries of many public activities. But they are tough to administer and so cannot reasonably be expected to replace all other taxes.

Questions

1. After 9/11, Lower Manhattan became a less desirable place to work and to live. Many people had traumatic memories, services and subways were severely disrupted, and serious questions lingered about environmental pollution from the World Trade Center's collapse. Some law firms and financial services businesses moved to Midtown, Manhattan's other and larger central business district, or to Jersey City, directly across the river.

   In response to this situation, a large portion of the $20 billion that the federal government pledged to help New York recover from 9/11 has gone into subsidies designed to induce firms and residents to stay in Lower Manhattan. Business tenants who stay in Lower Manhattan get money from the federal government if they stay long enough. Office rents are not regulated in New York City, and competition among building owners is fierce. Lower Manhattan is a small part of a large metropolitan office market.

   a. Relative to the situation in which no subsidies were available, what happens to equilibrium net rents that tenants pay in Lower Manhattan? (Net rent = rent actually paid minus subsidy received.) What happens to rents actually paid?

   b. Who gains from the subsidy?

   c. What's the effect on the value of office buildings in Lower Manhattan?

   d. Is this program an appropriate public policy response to the events of 9/11? Explain why or why not.

2. In your job as economic development officer for the city of Henriola, you spend a lot of time talking to owners of businesses, trying to get them to move to Henriola. Their number-one complaint is that the land prices are too high, especially for sites in good locations. Most of the business owners who complain about high land prices end up locating elsewhere. What should you do about this? Why?

3. Up until 1996, commuters to New York City along New Jersey Transit's Morris and Essex train line had to transfer in Hoboken to the PATH subway service into Manhattan. This transfer was time-consuming and disruptive. In the evening, making the connections was also nerve-wracking. In 1996 New Jersey Transit inaugurated its Midtown Direct service. This eliminated the Hoboken transfer and saved about ten minutes on each trip for everyone going to Midtown or north of there. Fares were adjusted so that the monetary cost of a trip to Midtown stayed about the same. Assume everyone works in Midtown or north, and that substitution between land and capital in the production of housing is impossible.

   a. Draw the rent gradient before and after Midtown Direct service if the Morris and Essex corridor is a closed city. Who gains? Who loses?

   b. Do the same under the open-city assumption.

   c. When do the gains and losses occur?

   d. Over a five-year period, which do you think is a more useful model, closed city or open city? Why?

   e. Do you expect to see changes in wages earned in New York City as a result of Midtown Direct? If so, what kind? If not, why not?

4. Consider a circular, monocentric city with 500,000 low-income consumers and 200,000 high-income consumers, each of whom has a completely fixed demand for land of 2787.84 square feet (one ten-thousandth of a square mile). High-income consumers have greater transportation costs per mile: $500 per mile per year, as opposed to $250 per mile per year for low-income consumers. Assume rural land rents are zero. Which income group lives closer to the city center? Why? Draw the rent gradient. What happens to the rent gradient if transportation costs for the rich decrease to $400 per mile per year but transportation costs for the poor stay the same?

5. Consider a closed city in which all employment is at the city center. Suppose that the work week is shortened from five days a week to four days a week, but hours per day are adjusted so that output and income stay the same. Draw two residential rent gradients, one for the city with the five-day work week and one for the city with the four-day work week. Draw them on the same graph, and label them carefully so that one can be distinguished from the other.

6. Mudville has lousy weather, frequent tornadoes, poor drinking water, an incompetent local government, and (as everyone knows) a baseball team that never wins. Pleasantville has wonderful weather, occasional showers of dollar bills, great drinking water, a superb local government, and its team wins

all the time. People who live in Mudville and Pleasantville do all their work by telecommuting. Everyone has the same innate preferences about consumption and leisure, but people have different incomes. What is a necessary condition for the residents of Pleasantville with an income of $50,000 a year to be better off than residents of Mudville with the same income (there are some of each)?

7. In the model discussed in section II of this chapter, suppose there are two magnet sites, ten kilometers apart, on a skinny peninsula. Each magnet site is a closed city with enough people to spread six kilometers in each direction. Draw the rent gradient.

References

Armstrong, Robert, and Homer Hoyt. 1941. *Decentralization in New York.* Chicago: Urban Land Institute.

Baum-Snow, Nathaniel, and Matthew Kahn. 1998. "The Effects of the New Urban Rail Transit: Evidence from Five Cities." Mimeograph. Federal Reserve Bank of New York and Columbia University.

Bollinger, Christopher R., and Keith Ihlanfeldt. 2003. "The Intra-Urban Spatial Distribution of Employment: Which Government Interventions Make a Difference?" *Journal of Urban Economics* 53:396–412.

Cervero, Robert. 1994. "Rail Transit and Joint Development: Land Market Impacts in Washington, D.C., and Atlanta." *Journal of the American Planning Association* 60(1):83–94.

Clark, W. C., and J. I. Kingston. 1930. *The Skyscraper: A Study in the Economic Height of Modern Office Buildings.* New York: The American Institute of Steel Construction.

Damm, D. S., et al. 1980. "Response of Urban Real Estate Values in Anticipation of the Washington Metro." *Journal of Transport Economics and Policy* 14(3):316–336.

Gatzlaff, D., and M. T. Smith. 1993. "The Impact of the Miami Metrorail on the Value of Residences near Station Locations." *Land Economics* 69(1):54–66.

George, Henry. 1900. *Complete Works of Henry George,* vol. 1. New York: National Single Tax League.

Hurd, Richard M. 1903. *Principles of City Land Values.* New York: The Real Estate Record and Builders Guide.

Landis, John, and Robert Cervero. 1999. "Middle-Age Sprawl: BART and Urban Development." *Access* 14(Spring):2–15.

U.S. Bureau of the Census. 1972. *Historical Statistics of the United States.* Washington, DC: U.S. Government Printing Office.

Voith, R. 1993. "Changing Capitalization of CBD-Oriented Transportation Systems: Evidence from Philadelphia, 1970–1988." *Journal of Urban Economics* 33(3):361–376.

Willis, Carol. 1995. *Form Follows Finance: Skyscrapers and Skylines in New York and Chicago.* New York: Princeton Architectural Press.

Too Many Cars? Too Much Lawn? Too Much Blight?

Some people claim that cities in the United States have developed too sparsely, particularly in the years since World War II. Residential densities are too low to permit much use of mass transit, and automobile commuting encourages low densities. Americans would be better off, many commentators argue, if they lived closer together, worked closer together, and used more mass transit, as they used to do. The shorthand statement of this problem is that there is too much **sprawl.**

In this chapter we will begin to analyze the complaints about sprawl. First, in a comparison of Tokyo and New York, we will see that U.S.-style cities are not inevitable; cities with a lot less sprawl work perfectly well. We will ask whether better transportation policies would make (or would have made) U.S. cities look more like Tokyo, and whether cities in developing countries should imitate Tokyo or Phoenix. Next, we will look at whether land taxes reduce sprawl, examining the effects of transportation and land taxes by applying the lessons of the preceding chapters.

In section IV we will explore a different story about sprawl: the idea that U.S. cities are too sparsely populated not because transportation policies have been deficient but because externality problems, such as like crime, environmental degradation, and rowdyism, have not been handled well. If Newark were safer and if the fragile environments of the West Jersey highlands were better protected, more people would live in Newark and fewer in West Jersey "McMansions."

In section V, I digress somewhat from the topic of sprawl in order to explain the property rights approach to externalities. We will look at many externality issues in the rest of the book—some related to sprawl and some not—and the property rights approach will prove to be a handy tool.

To get the most from this chapter, you should be familiar with these concepts: externalities, Pareto improvement, potential Pareto improvement. You can find these terms in the Glossary. You will also need to understand open-city and closed-city cases.

I. Tokyo and New York

The easiest way to argue that U.S. cities could be denser than they are today is to produce an example of a wealthy city that is considerably denser than any American city. Tokyo is such an example. Hatta and Okhawara (1994) makes the comparison.

Even though New York is by far the largest, densest, and most transit-dependent metropolitan area in the United States, Tokyo is bigger, denser, and more transit-dependent. Both metropolitan areas cover about the same land area. About twenty-nine million people live in (metropolitan) Tokyo and thirteen million work there; sixteen million live in (metropolitan) New York and seven million work there. (Exactly what a particular metropolitan area consists of is nowhere uniformly defined; Hatta and Okhawara used definitions that tried to be as consistent as possible between the two cities.) Although Chiyoda ward in Tokyo is nowhere near as dense as Manhattan (possibly because the danger of earthquakes until recently inhibited high-rise construction), Tokyo stays at Brooklyn-like densities as far out as the distance from New York to the Jersey shore. The result is that, for the metropolitan area as a whole, Tokyo has about 5,900 residents per square mile, while New York has about 2,900. The Tokyo subway carries roughly twice as many passengers a year as the New York subway; Tokyo train service has five times as many regular riders as New York metropolitan-area train service. (New York mass transit has improved in quality and ridership has grown since the time that Hatta and Okhawara wrote, but it had been so far behind Tokyo that the ratios are little changed.) In 1980 only 3.2 percent of commuters to Chiyoda ward used cars; only 4.2 percent of commuters to the four central wards of Tokyo traveled by car. The volume of transit ridership in Tokyo is great enough that commuter lines are profitable without significant government subsidies, and trains run with astonishing frequency. During rush hour, Chuo-line trains arrive at Tokyo Station every three minutes (almost like an elevator); the New Haven–line trains of Metro North arrive at Grand Central Terminal every twenty minutes.

It is difficult to see how you could fit thirteen million jobs into some (probably expanded) version of the New York metropolitan area without either extraordinarily long commutes, huge traffic jams, or a mass transit and density pattern like Tokyo's. If the economies of scale we talked about in Chapter 2 are important and extend into this range, then cities with only seven million jobs are not going to be as productive, per job, as cities with thirteen million. So New York's low density and poor transit system may be making it miss out on economies of scale that Tokyo is realizing. In that sense, New York's low density and poor mass transit are responsible for a major loss of productivity—and the even lower density and poorer mass transit of other U.S. cities are responsible for much larger losses.

II. Is Tokyo Better?

One possible conclusion to be drawn from this exercise is that Tokyo's plan-
ners have done a better job than New York's. But there is an alternative, more
benign interpretation of the U.S. experience. Tokyo may be more dense simply
because the alternatives to Tokyo that the Japanese have are not as attractive as
the alternatives to New York that Americans have. If Japanese people had al-
ternatives as attractive as, say, Nebraska and California, you would need lawns
and cars to entice them to work in Tokyo, just as you need lawns and cars to
entice most Americans to work in New York. In 2000, even in the New York
metropolitan area (the larger Consolidated Metropolitan Area), 71 percent of
households had at least one car and almost 40 percent lived in single-family,
detached homes (U.S. Bureau of the Census 2000).

In the benign interpretation, the historical trend of declining density and
declining mass transit use in the United States is a normal and natural pro-
cess. Mills (1972) and Mieszkowski and Mills (1993) show that density and
land-price gradients have been getting flatter in U.S. cities since about 1820;
suburbanization is by no means a post–World War II phenomenon. Con-
tinuing improvements in transportation alone would be sufficient to explain
why these gradients have been flattening, but there have been many other con-
tributing factors as well: frozen food, television, and air-conditioning, as I
mentioned in Chapter 1; electricity, so that factories didn't need to be near
flowing water or some other power source; telephones, possibly, and rising in-
come—greenery and privacy are things that people are willing to spend more
money on as they grow richer. Flattening density gradients and the eclipse of
mass transit in the United States, in this view, are simply progress. Sprawl is a
fact of modern life.

The benign view does not argue that government actions played no role
in the trend toward sprawl, that it sprang unaided from the pure workings of
an untrammeled market. Clearly, governments in the United States could have
stopped or reversed the trend at any time by taking one of any number of dif-
ferent actions: by banning automobiles, or stopping road construction, or
paying for all high-rise construction, or not insuring mortgages, or not polic-
ing traffic at all, or enforcing a speed limit of fifteen miles per hour. Rather,
the benign view of the U.S. trend is that government actions (and inactions)
were more or less justified by cost-benefit analysis: that most road construc-
tion produced more in benefits than it absorbed in costs, that gas taxes rea-
sonably meet the cost of air pollution, that larger transit subsidies would cost
more than they would be worth, and so on. This view does not require that all
relevant government policies have been wise, only that they not have been sys-
tematically unwise to such an extent that they created a vastly different density
gradient.

I have already identified some relevant government policies that are probably unwise and probably affect density gradients—the lack of congestion pricing, the lack of responsibility for accidents that auto insurance creates, the nontaxing of employer-provided parking, the silly structure of mass transit fares, the general undertaxing of land and consequent overtaxing of other commodities. The next few chapters will bring up more issues that affect density—mortgage deductibility, mortgage guarantees, average rather than marginal cost pricing of sewer and water service, zoning, crime, race. Many people will find current and historical policies in these areas deficient, too. But we will also look at other unwise policies that promote centralization—subsidies to building rail transit, Western-style toilets, and many economic development activities, for instance. The United States (and Japan, too) subsidizes agriculture heavily, and these agricultural subsidies shift land out of urban use and into farming. They make cities too dense and too small.

Probably the general drift of bad policy is to encourage low densities, but no one knows how much actual, measurable difference bad policies have made. To figure it out, you would first have to find optimal policies on all these areas and then figure out how U.S. cities would have grown if these policies had been in place. It would be a mammoth project, and no one has come close to doing it.

Nor should anybody. If policies are bad, they are bad regardless of what they do to density gradients. The cities that are built today and will be built in the future, in the United States and in the rest of the world, will be built with and for new technologies. Lagos shouldn't end up looking like either New York or Tokyo or Phoenix. Currently, Chinese cars pollute more than American cars do; optimal density for Beijing is going to be different from optimal density for Los Angeles. Whether or not the policies of Robert Moses in the 1930s in New York were the right ones or the wrong ones depends in large measure on the technologies that were available then—how people drove, shopped, worked, communicated with each other, and played. People in Lagos in decades to come won't be doing any of these things like 1930s New Yorkers did, and so figuring out how New York would look if Moses had followed better policies is not going to help policy makers in Lagos today.

The important question is whether the policies—toward car safety, mass transit, housing, agriculture, sewers, and so on—are the best they can be, not what a city looks like. If late twenty-first-century Lagos had excellent policies and looked like twentieth-century Phoenix, that would be wonderful. If it had excellent policies and looked like fourth-century Rome, that would be wonderful, too. Cities are a way of helping people live their lives better; they're not an end in themselves.

III. Land Taxes

One of the bad policies that most American cities pursue is that they undertax land and overtax improvements; this was one of the conclusions from Chapter 6. This also probably contributes to sprawl. This idea was expressed by Gaffney (1964, p. 132):

> Think how many miles of wires, ducts, pipes, and conduits there are in a skyscraper, not to mention stairs and elevators; all social overhead capital supplied at private expense, and taxed besides. If the same floor space were sprawled over a section of suburban prairie, the connecting wires et al. would not only be much longer, but supplied at public expense and/or heavily subsidized. We tax vertical transportation but subsidize horizontal. With such a large and systematic fiscal bias at work, is it surprising we get urban sprawl?

Reducing taxes on structures and increasing them on land would lower the relative cost of structures (since the price of land to users would stay the same or go up a little) and so induce more building on the same amount of land. In an open-city model, cities would become both more dense and physically larger (since some construction in outlying areas would become profitable).

Would this be desirable? It would be a potential Pareto improvement. The underuse of land taxes thus is reason to believe that cities are too sparse, not to believe that they cover too much ground.

IV. Flight from Blight

There's an alternative story about U.S. cities, another explanation for why they might be too sparse and too big. It's called the **flight-from-blight hypothesis.** The idea is that policies have done a worse job than they should have in promoting health, safety, and morals in the older, more central parts of metropolitan areas. As a result, people who could afford to have moved to sparsely populated suburbs in greater numbers than they should have. Because crime is high in inner cities, for instance, people flee from them and build homes far away, in areas that should instead be left for fields and forests.

Two kinds of everyday experiences give the flight-from-blight story some plausibility. The inner parts of many U.S. cities have lots of problems, and so do many of the people who live there. This is especially true for older cities in the Northeast and Midwest. More activity used to occur in these areas, and markets don't seem to be working right in them now. On the other hand, knocking down trees and tearing up farms to build housing developments and shopping malls on the outskirts of metropolitan areas has obvious environ-

mental costs. If inner cities could be restored to their former glory, the thinking goes, the farms and forests could be spared and everyone (except maybe some greedy developers) would be better off.

A generalized yearning for the old days, like this one, however, is not the same as a convincing demonstration of potential Pareto improvements. Several steps are missing.

The first step is to show not just that externalities occur but that they are excessive. Cars are going to crash, influenza is going to spread, distraught lovers are going to murder each other, and thugs are going to rob gas stations—these are facts of life that by themselves are of no interest to economists. The relevant question is not whether these events occur at all but whether some instances of these events could be avoided cheaply enough that the winners from avoiding them could compensate the losers. Thus to address flight from blight, we need to know not whether blight exists but whether too much blight exists.

Second, we need to know where the excess blight is, assuming it is somewhere. The optimal level of traffic accidents, for instance, is almost certainly higher in more densely populated areas, where streets are narrower and lined with parked cars and children play on the sidewalks rather than in playgrounds or backyards. Or racial tensions may be greater when you can hear your neighbor's toilet flushing and he has to listen to your child's birthday party. If this is the sort of blight that people are taking flight from, then it is surely no cause for worry—and it may even be worth cutting down a few trees to facilitate their move.

Third, we need to specify clearly what people's alternatives are. Flight from blight makes little sense in the open-city model, for instance. In that model, if downtown Los Angeles is dirtier or more crime-ridden than it should be, there is no impact on the population of suburban Orange County or San Bernardino County. The people who would have lived or worked in downtown Los Angeles stay in Mexico or move to Denver instead. The only way to induce fewer people to move to Orange or San Bernardino County is to make those places less attractive, and making downtown Los Angeles nicer doesn't do that.

Finally, even if excess blight is concentrated in the more densely populated parts of metropolitan areas, flight from blight is a serious problem only if the reduction in activity that excess blight causes is great. If excess blight causes little movement, then the densities and the dimensions of U.S. metropolitan areas are fundamentally all right, even if a few jobs or residences are in the wrong places.

On this last question, economists have done a substantial amount of empirical research, and their conclusions have generally not supported the flight-from-blight hypothesis. Mieszkowski and Mills (1993, p. 140) find that the timing of suburbanization argues against a large role for flight from blight:

"The most rapid pace of suburbanization . . . occurred between 1920 and 1950, a period not known for racial tensions, school desegregation, rising local taxes and high crime rates. In fact, between 1940 and 1950, national crime rates fell." After reviewing a number of cross-section studies that tried to find correlations between blight and flight, they conclude that cities with more crime, worse schools, and higher taxes don't lose either jobs or residents faster.

The cross-section studies do show that metropolitan areas with a higher percentage of African Americans in the central city decentralize faster. This finding would strongly support flight from blight if African Americans were blight, but that is ridiculous. African Americans are some of the most educated, talented, productive, and accomplished people the world has ever known. This finding in the cross-section studies tells us that race relations affect spatial location, but it doesn't tell us what to do about race relations. That is a topic I will discuss in detail in Chapters 11 and 12.

The cross-section studies are hardly definitive evidence against the flight-from-blight story, however, because they do not measure the effect of *excess* blight; they just measure blight. Finding out how much blight is excessive and where it's located is a much more difficult task, one best accomplished by looking at each possible source of blight in some depth. That is the strategy I will follow in this book.

This piecemeal strategy has a great advantage over trying to put together a more overarching approach to sprawl: it makes for better policy recommendations. What's wrong with sprawl (if anything) is that it precludes certain potential Pareto improvements. If we can identify those potential Pareto improvements directly and figure out what to do about them, then we really don't have to spend any time thinking about sprawl, defining it, or measuring it. For instance, if metropolitan areas are too sparsely populated because inner-city schools are not as good as they should be, then the appropriate policy is to improve inner-city schools, not to build subways or preserve farmland. Lousy schools are a problem no matter how much sprawl they cause or don't cause.

V. Negotiation and Property Rights

Policies toward externalities are going to be a major theme in the rest of this book, just as they were in the preceding chapters on transportation. In those chapters, all you needed to know about externalities was that if you make people bear the full cost of their actions, all will be well (or efficient, at least). For much of the following text, that basic knowledge will not be enough. A more sophisticated understanding of externalities will be required.

The possibility of negotiation is one area in which greater sophistication would help. In the discussions of traffic policy, negotiation was not feasible:

you can't roll down your window in the Lincoln Tunnel queue and shout out an offer to all the drivers behind you that you will let them pass you if they collectively pay you enough. Sometimes, though, negotiation is feasible, and when it is, policy prescriptions are different.

Consider, for example, two people who will be in a small room together for a while. One is a smoker who would like to smoke; the other is a nonsmoker who would be annoyed if the smoker smoked. Suppose they can talk with each other. Then the nonsmoker will try to bribe the smoker not to smoke. If smoking causes the nonsmoker more harm than it's worth to the smoker, the nonsmoker will succeed and the smoker won't smoke. If not, the smoker will smoke. In either case, after they have talked, no potential Pareto improvement is possible: if the smoker isn't smoking, it's because smoking causes more harm than benefit; if the smoker is smoking, it's because the opposite is true.

The same result will occur in a society where smokers need permission to smoke, so long as negotiation is possible. (Perhaps the nonsmokers can call the smoking police to stop the smokers, but can also refrain from doing so.) In this case it is the smoker who will take the initiative and try to bribe the nonsmoker. If smoking causes more harm to the nonsmoker than it's worth to the smoker, the nonsmoker will refuse the smoker's entreaties and there will be no smoking. If not, the bribe will be accepted and the smoker will smoke. Thus, in terms of efficiency, so long as they can talk it doesn't matter who needs permission from whom: the smoker will smoke if smoking causes more benefit than harm, and will not smoke if the opposite is the case. Negotiation will lead to efficiency. This result is known as the **Coase theorem,** after Ronald Coase who published the first account of it in 1960.

How does the Coase theorem modify the way we should think about public policy toward externalities? Suppose the smoker and nonsmoker negotiated, the smoker paid the nonsmoker for permission to smoke, and then lit up. If a public-spirited analyst then came on the scene and were naive about the Coase theorem, she would observe the nonsmoker wheezing and coughing, the smoker puffing, and conclude that a tax on smoking could produce a potential Pareto improvement, but she would be wrong because she missed out on the negotiations. If the parties have already negotiated, nothing more can produce a potential Pareto improvement. If a tax is imposed and it stops the smoker from smoking, the gain to the nonsmoker is less than the loss to the smoker. Government intervention isn't needed and doesn't help when the parties have negotiated; just looking at a physical relationship that appears to be an externality doesn't tell you everything you need to know.

That doesn't mean that government intervention can never produce a potential Pareto improvement. Sometimes—as in the case of traffic and air pollution, for instance—negotiation is impossible. The Coase theorem means that, in arguing for a government intervention, you have to show not only that

what someone does affects someone else, but also that the parties cannot negotiate themselves out of the problem.

The Coase theorem also suggests another type of policy, aside from taxes, subsidies, or tolls, to alleviate the problems caused by externalities—helping the parties negotiate. The smoker and the nonsmoker could negotiate only because we imagined there were clear rules about who could agree to what and what would happen in the absence of an agreement and that there was some way to enforce an agreement. Without clear guidelines, negotiation can't take place. One way to alleviate externalities, therefore, is to provide such guidelines. This is generally a job for governments.

This idea of encouraging negotiation by straightening out ambiguities is called the property-rights approach to externalities. Property is the branch of law concerned with who gets to do what with what, and so "establishing property rights" is just another way of saying "writing down clear rules for negotiating."

Notice that ownership is not a simple, absolute concept; writing down clear rules for negotiating can take a long time, and there are many different ways of doing it. My owning the Coca-Cola trademark may or may not imply that I can keep someone else from using the URL *www.coca-cola.com.* My owning a cigarette may or may not permit me to smoke it in an elevator. My owning a mountain bike may or may not permit me to ride it in Central Park. My owning a tire iron may or may not (it probably does not) permit me to beat you over the head with it.

Lawyers sometimes refer to property as a "bundle of sticks." By that they mean that ownership is really a collection of rights. What we call "owning a tire iron" consists of the right to keep other people from using it (except the government, and even the government can use it only under certain conditions), the right to carry it in your trunk, the right to change your own tires with it, the right to prop open a door you own with it, the right to display it in your living room, the right to melt it down and recast the metal into little statues of yourself, the right to throw it at rabbits and snakes (real or imaginary) in your backyard, and so on. These are the sticks that are in the bundle. "Owning a tire iron," in most of the United States today, does not include the right to use it to remove other people's tires unless they say you can, the right to attach it perpendicularly to your front bumper, the right to throw it at imaginary rabbits on a crowded sidewalk, the right to melt it down and recast the metal into little statues of Mickey Mouse and sell them on the street, the right to have it displayed in the Metropolitan Museum of Art, and so on. These are some of the sticks not in the bundle labeled "owning a tire iron."

The Coase theorem says that, so far as efficiency is concerned, if people can negotiate, it doesn't matter which sticks are in which bundle. But every stick has to be assigned unambiguously to one and only one bundle, and it's

the government's job to assign them. Does owning a tire iron give you the right to melt it down and sell statues that *almost* look like Mickey Mouse? Can you put your tire iron on your dashboard and drive with a small child in the car? Can you throw it at a snake in a national forest? I don't know. Even such a prosaic and well-worn concept as owning a tire iron has ambiguities, and changes frequently. In more technologically dynamic arenas like the Internet, organ transplants, and the deep seabed, sticks that are in no bundle are constantly turning up. Efficient negotiation can take place only after those sticks have found a home in some bundle.

Finally, while the Coase theorem indicates that no matter what sticks are in which bundle, negotiation will produce an efficient outcome, it doesn't say that nobody should care about what sticks are in which bundle. If you own the Coca-Cola trademark, it sure makes a lot of difference to you whether or not that means you also own the URL *www.coca-cola.com*. Considerations like perceived justice and fairness are important in designing property rights. (In fact, one interpretation of the Coase theorem is that if negotiating opportunities are good enough, justice and fairness are the only things you need to worry about in assigning property rights, because efficiency will take care of itself.)

VI. Conclusion

Up until a century ago, moving to a large city meant that you were more likely to die soon; if you didn't die right away, your children were more likely to die at birth, and if they survived, they were more likely to be stunted and sickly. Population density was at the root of this public health disaster, and until these conditions were ameliorated, no city could survive with many more than a million people in it.

Improvements in medicine, nutrition, and transportation, along with rising incomes, have combined over the past century to make cities healthier places to live, healthier even than rural areas. As human diseases mutate and evolve, however, there is no guarantee that this happy state will continue indefinitely.

Lowering the population density in U.S. cities was an important process for improving urban health a hundred years ago—it created positive externalities then. More recently, though, some have argued that the process has gone too far. Essentially, two arguments have been advanced for why sprawl is excessive in the United States. One emphasizes transportation: policies have been too friendly to cars and too unfriendly to mass transit. The other argument emphasizes externalities in the older central parts of metropolitan areas: because these factors—sources of urban blight—have not been controlled as well as they should be, too many people have fled to the suburbs.

Certainly there are many reasons to think that reducing sprawl might be a potential Pareto improvement. For instance, Americans drive too much and cause too much congestion, pollution, and damage from traffic accidents. Land taxes are underused. Parking is subsidized. Correcting these problems would result in significant potential Pareto improvements, even if it didn't reduce sprawl much. So would reducing excessive central-city blight. Other distortions, such as farm subsidies, work in the opposite direction—they make cities too small and too dense.

The major reason U.S. cities are larger and less dense today than they were a hundred years ago is that people are richer and technology is better. The change was not produced totally through some conspiracy or by a mistake. The prescription to relieve excessive sprawl is identical to the prescription for the problems that make sprawl excessive. Making potential Pareto improvements has some value; reducing sprawl per se does not.

We need to look at sources of urban blight one by one. This will require a more sophisticated approach to externalities than was used in the transportation chapters. To think seriously about externalities, you have to consider possibilities for negotiation and you have to understand the various ways property rights can be put together. Like externalities, property is a much more complex idea than it first appears.

Questions

1. Tetracycline is an antibiotic that has quite a few different uses. It can relieve the symptoms of some comparatively minor ailments (acne and conjunctivitis, for instance), as well as cure some serious diseases (typhus and pneumonia, for example).

   The bacteria that tetracycline acts against live short lives and multiply rapidly. Occasionally a mutant that is resistant to tetracycline pops up. The more often tetracycline is used in the world, the higher the proportion of resistant bacteria becomes, and so the less effective tetracycline is. Eventually tetracycline will lose enough effectiveness that people will have to turn to other antibiotics, which will be more expensive and less effective.

   a. Explain the externality involved in tetracycline use. Outline a potential Pareto improvement.

   b. Describe a Pigouvian policy for realizing this potential Pareto improvement.

   c. Describe a Coasian policy for realizing this potential Pareto improvement.

2. "Cybersquatting" is a practice in which ordinary people acquire e-mail addresses and Web site addresses (URLs) with names that are popularly associated with famous companies. The companies then often have to buy the names back from the squatters. In the short run, what does the Coase theorem imply about public policy toward cybersquatting? Should the Internet be allowed to develop along Coasian lines?

3. The total benefit (measured in dollars) that a musician gets from making music is

$$B_m = 25 \ln D,$$

where D denotes the decibels involved. A sleeper lives next door. The total benefit that she gets is

$$B_s = 20 \ln (20 - D),$$

where 20 is the loudest sound technically possible.
   a. Find the level of sound that maximizes total benefits for musician and sleeper.
   b. Suppose no ordinance about noise exists. What will happen if the two can bargain freely? Is this efficient?
   c. The city adopts an ordinance forbidding noise levels above 5 decibels. Being drawn and quartered is the penalty for violating the ordinance, but the city has hired no one to enforce the ordinance. It will respond only to complaints. What will happen? Is this efficient?
   d. The city revises the ordinance. The penalty for violating the ordinance is now $1 for each decibel over 5. Enforcement is by inspectors who are perfectly effective, omnipresent, and absolutely incorruptible. Complaints have no standing—the ordinance will be enforced whether or not the neighbors complain. What will happen? Will this be efficient? How does this compare with the situation in part c? If you were on the city council, would you support the revised ordinance?

4. An **urban growth boundary** (UGB) is a law that prevents expansion of a city beyond certain boundaries. Suppose a city has a binding UGB—that is, if there were no law, the area devoted to urban uses would be larger. Assume all business activity takes place in the central business district, and all land outside the CBD is used for housing workers who commute to the CBD. All households are identical.
   a. Draw the land rent gradient and the density gradient that would prevail if there were no UGB.
   b. Since landowners have the option of putting more apartments on a plot of land and using more or less land to produce a unit of housing, you have to distinguish between the rents consumers pay for housing and what the landowners collect per square meter. Draw the gradient of

housing rents that consumers actually pay, still assuming no UGB in effect.

c. Consider the open-city case. Suppose the UGB is in effect. Draw the gradient of housing rents that consumers actually pay. How does this compare with the gradient without the UGB? Are consumers better off or worse off? How does the city's population change?

d. Draw the density gradient when the UGB is in effect.

e. Draw the gradient of rents that landowners collect per square meter. How does it compare with the gradient without the UGB? Which landowners, if any, are better off? Which landowners, if any, are worse off?

f. In the open-city case, is the UGB a potential Pareto improvement?

g. Now consider the closed-city case. Draw the density gradient with the UGB and compare it with the density gradient without the UGB. How does the city's population change?

h. For the closed-city case, draw the gradient of land rents per square meter that landowners receive and compare it with the gradient without the UGB. Why does it change the way it does? Which landowners, if any, are better off? Which landowners, if any, are worse off?

i. For the closed city, draw the gradient of rents that consumers pay, and compare it with the gradient without the UGB. Are consumers better off or worse off?

j. For what purposes would the closed-city analysis be more useful? For what purposes would the open-city analysis be more useful?

k. What's missing from this analysis of UGBs?

5. Some cities, such as Chicago and Milwaukee, have CBDs located next to lakes or oceans and so cannot expand in several directions. Other cities, such as Calgary and Atlanta, have no such natural boundaries and can expand in all directions. If you were to compare density and land price gradients in cities that have natural boundaries close to the CBD with those in cities that do not have such natural boundaries, what would you expect to find? Why? Apply your answer to the UGB question.

6. Some cities, such as San Francisco, are located on hilly land and some—Philadelphia, Shanghai—are located on flat land. If you were to compare density and land price gradients in hilly cities with those in flat cities, what would you expect to find? Why?

References

Gaffney, Mason. 1964. "Containment Policies for Urban Sprawl." In Richard L. Stauber, ed., *Approaches to the Study of Urbanization: Proceedings of the Inter-University Seminar on Urbanization in the Missouri River Basin,* pp. 115–133. Lawrence: University of Kansas Governmental Research Center.

Hatta, Tatsuo, and Toru Okhawara. 1994. "Commuting and Land Price in the Tokyo Metropolitan Area." In Yukio Noguchi and James M. Poterba, eds., *Housing Markets in the United States and Japan.* Chicago: University of Chicago Press.

Mieszkowski, Peter, and Edwin S. Mills. 1993. "The Causes of Metropolitan Suburbanization." *Journal of Economic Perspectives* 7(3):135–147.

Mills, Edwin S. 1972. *Studies in the Structure of the Urban Economy.* Baltimore: Johns Hopkins University Press.

U.S. Bureau of the Census. 2000. "American FactFinder." Accessed at *http://factfinder.census.gov/servlet/DTGeoSearchByListServlet?ds_name=DEC_2000_SF3_U&_lang=en&_ts=106933343360.*

Chapter 8

Rules

⎯⎯⎯⎯

*A*nyone who has experienced childhood knows what you do when you see people doing something wrong: you tell them to stop. And you threaten them with some punishment if they don't. This approach is simple, direct, and often effective. If policy makers know what they want to accomplish, prohibiting people from doing otherwise is the easiest, most surefire way of getting it done. Moses transmitted ten *commandments,* not nine relative price-ratios.

It's not surprising, then, that from the earliest of times life in cities has been governed by rules and regulations on top of rules and regulations. When Socrates broke the rules of the city of Athens, it led to his downfall. Today, you can't drive west on West 118th Street in Manhattan or east on West 119th Street, park near a hydrant (even if the hydrant has been out of service for twenty years), buy a drink for a female bar employee in Newark, spit on the street, sell a single-family house without a smoke detector or an indoor bathroom, operate a pornography store in a residential neighborhood, toot your horn next to a hospital, drink alcohol in a public park, park a truck on your front lawn, or keep your children locked up in a closet, even if they have misbehaved.

Governments aren't the only entities that make rules. Firms do, too: what you can wear to work, how you should answer the phone, what Web sites you can access from your work computer, what time you have to show up for your job, whether you can bring your dog or your paramour to work with you, whether you can take money from the competition or give it to your co-workers. In Chapter 2 I compared cities with firms; maybe they both use so many rules for similar reasons.

Neoclassical economists generally like prices a lot more than rules (by rules here I mean rules about quantities—*how much* you can spit on the

⎯⎯⎯⎯

To get the most out of this chapter, you should be familiar with these concepts: consumer surplus, externalities, monopoly, Pigouvian taxes, potential Pareto improvement, and sunk costs. You can find these terms in the Glossary. The concepts of Coasian bargaining and sprawl from previous chapters will also be used.

street—not rules about prices). The reasoning should be familiar: if the government or the firm has enough information to set a price so it reflects the true (marginal) cost of an activity, then the only instances of an activity that will occur are those in which benefits exceed costs (to everyone), and these are precisely the instances that should occur if no further potential Pareto improvements are to be possible. Rules, in contrast, appear to neoclassical economists to be blunt instruments: they don't differentiate well between those instances of an activity that should be done and those that should not. Replacing the blunt ax of a rule with the sharp scalpel of a price correctly set almost always causes a potential Pareto improvement.

Neoclassical economists tell horror stories about regulations, too. To cut down on air pollution and congestion, for instance, Mexico City banned driving for a certain number of cars every day, based on the last digit of their license plate numbers. Cars with the last digit zero, for example, could not be driven on Mondays, those with the last digit three could not be driven on Tuesdays, and so on. Many households reacted by buying a second car. Having two cars with sufficiently different license plate numbers guaranteed that every day you would have at least one car you could use. Once households had more cars at their disposal, they drove more. The regulation designed to reduce congestion and pollution ended up increasing both (Eskeland and Feyzioglu 1995).

The neoclassical case against rules, however, is almost surely overstated. It compares ideal prices—the prices you could set only with excellent information—with actual rules—rules set in a context of poor information. In fact, if a government (or a firm) has the superb information it needs to set precisely the right prices, then it can set rules that work just as well as prices: it can permit those instances in which benefits exceed costs, and prohibit the others. When information is very good, rules work just as well as prices.

The real question, then, is whether rules work better than prices when the information the government has is poor. The answer here is the usual, "It depends." In the first section of this chapter I will offer some general principles for when rules work better than prices and vice versa.

In the next part of the chapter we will look at the different kinds of rules and regulations that governments promulgate to govern urban life and keep people from getting in each other's way too much. We will examine the legal bases for rule making, and some of the reasons why rules may end up being unjust or inefficient. This chapter will have a legal flavor, since lawyers (rather than economists) are the ones who specialize in thinking about what wise rules are like.

After that we will look at zoning, the most controversial form of local-government rule making, and at health codes, fire codes, and building codes. The last part of the chapter is about the regulation of street behavior.

By making it easier for people to live and work in close proximity harmo-

niously, a wise system of rules should increase density and help realize more of the economies of scale available in cities. Thus wise rules should be an antidote to sprawl. Critics of current regulations, however, maintain that they contribute to excessive sprawl. The goal of this chapter will be to sort out these arguments.

I. When Rules Work Better than Prices

There are several reasons, some obvious and some not, for using rules rather than prices or **Pigouvian taxes** to mitigate the harm that externalities cause. Sometimes collecting money is simply infeasible under current technology, as at a traffic intersection or when a building is on fire. Other times the external harm of an act is so great that the optimal number of times for the act to be committed is clearly zero. Arson and murder (except in self-defense) are examples of this case.

But rules are used in many other situations in which collecting money is feasible and the actions in question are not heinous. The arguments for and against rules in these situations depend on the size of the social losses that poor information is likely to cause. One argument for using rules deals with the information that policy makers have, and a second argument concerns the information that people implementing the policies have.

A. Information for Making Policy

Rules work better than prices when the marginal external cost of what the public is doing is very sensitive to the quantity of what the public is doing, and when the marginal private benefit is not very sensitive to the quantity. Prices work better than rules in the opposite cases. This is the argument about policy makers' information.

For instance, a traffic intersection should be governed by rules because the marginal external cost of one car going through—about zero—is a lot less than the marginal external cost of two cars going through at the same time—a collision—while the marginal private benefit of each passage does not vary as much. Firms use rules rather than prices to schedule deliveries of intermediate inputs because when you have two cars otherwise finished, for example, the eighth tire to be delivered is a lot more valuable than the ninth. Supermarkets use prices to sell cereal because the cost to them of each box that disappears from their shelves is simply the wholesale replacement cost, which each individual supermarket thinks of as approximately constant.

What's the intuition behind these precepts? Consider a small cruise ship with enough lifeboats for 100 people. The ship has 20 crew members, all of them deferential or heroic enough that they would give up a place in a lifeboat to any passenger. Thus each passenger beyond the eightieth endangers an ad-

ditional crew member, while the first 80 passengers pose no such danger. The marginal social cost of carrying passengers beyond a total of 80 is much greater than the marginal social cost of carrying 80 passengers.

If you're running the cruise ship and care about the crew (either because you like them or because you'd have to pay them more if they were in danger), you'll set a rule: no more than 80 passengers. You know the quantity of people beyond which marginal passengers harm the crew, but you don't know for sure how many passengers you'll have at a given price. If you set a price without a quantity rule and more than 80 passengers book passage, the consequences could be tragic. So when you don't know demand but you do know a quantity at which marginal social cost changes (80), it makes sense to use a rule rather than a price.

In contrast, consider a copy machine in a public library. The library must replace the machine's ink, toner, and paper when they run out. Since all of these items are for sale in regular stores, the cost of a copy to the library—the marginal social cost—is pretty much the same no matter how many copies have been made in a given day. There's no magic number of copies after which the marginal cost suddenly jumps up. So setting a rule such as "Only 80 copies a day" would be a bad idea for the library. If patrons happened to want to make 200 copies some day and were willing to pay for the ink, toner, and paper for each of them, the rule would preclude a potential Pareto improvement. Setting the price equal to marginal social cost makes a lot more sense. Then no matter how many copies patrons wanted to make at that price, the outcome would be Pareto optimal.

Thus when you know what the marginal social cost is going to be, as at the library, set a price. When you don't know what marginal social cost is going to be but do know where it changes, as on the cruise ship, use a rule. For situations in between, use a rule for those more like the cruise ship case and a price or tax for those more like the library case.

To be a bit more analytical, let's look at marginal social cost curves for the cruise ship and for the library. Figure 8.1 shows the marginal social cost curve for the cruise ship: it goes up at 80. Figure 8.2 shows the marginal social cost curve for the library: it's flat.

Now think about how a price or a tax looks in a diagram. The price is the same no matter how many copies are made or passengers board, and so it appears in the diagram as a flat line, as in Figure 8.3. By contrast, a rule looks like Figure 8.4. You can have as many passengers or make as many copies as you want so long as it is less than 80, but to go over 80 invites a large punishment. (This assumes the rule is enforced by some finite punishment, like prison. Sometimes rules are physically impossible to break: your car will get towed if you park next to a hydrant, and a college student who doesn't pass the required swimming test will never graduate, no matter how much tuition he pays. In these cases we can think of the cost of breaking the rule as infinite: no matter how great the benefits of doing so, nobody will do it.)

Figure 8.1 Marginal social cost for the cruise ship.
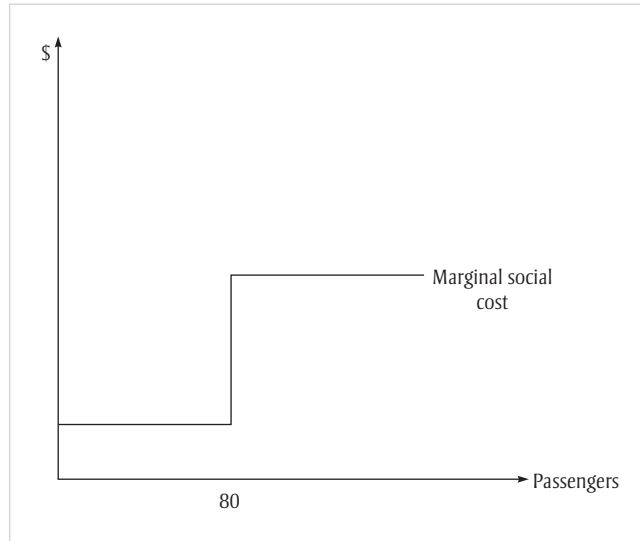
Figure 8.2 Marginal social cost for the library.
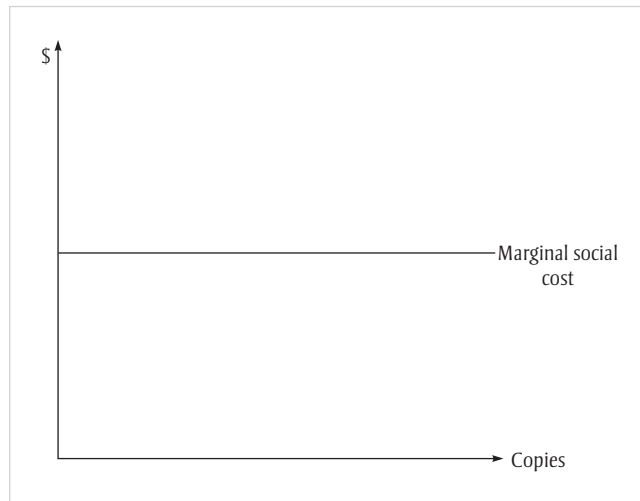
Compare Figures 8.1 and 8.2 with Figures 8.3 and 8.4. The correspondence is obvious: when marginal social cost looks like a tax (8.2 looks like 8.3), use a tax; when marginal social cost looks like a rule (8.1 looks like 8.4), use a rule. And the government should make the form of its response conform as closely as possible to the form of the social harm. As nearly as possible, the government should choose between rules and taxes to make the "punishment" fit the "crime."

To demonstrate a bit more rigorously how this correspondence comes about, we need to introduce demand curves (private marginal benefits). If the government, the cruise ship owner, or the library actually knew where these

Figure 8.3 A tax.
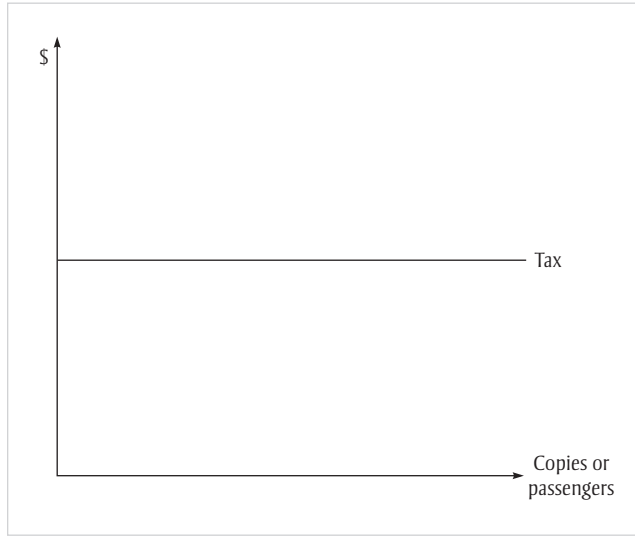
Figure 8.4 A rule.
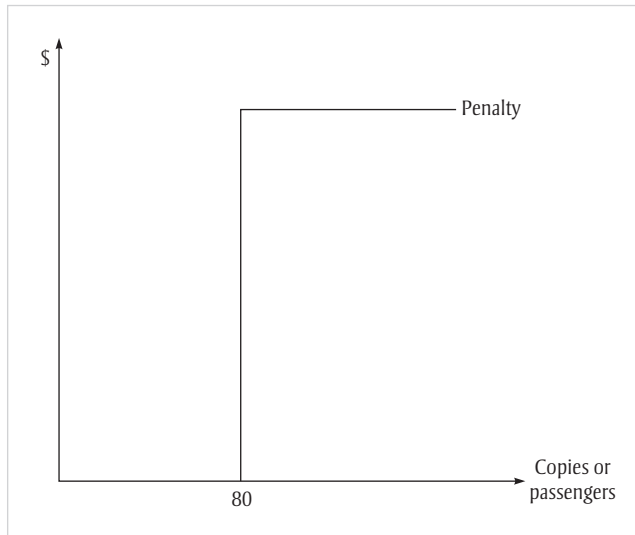
curves lay, of course, it wouldn't make any difference whether they used rules or prices: they could calculate both the optimal quantity and the optimal price, and implement both by implementing either. The choice between rules and prices, then, boils down to the question of which works better when the government doesn't know where the demand curve is, or is mistaken about it. Working with mistakes is easier than working with ignorance.

So in the cruise ship case, suppose the owner thinks the demand curve is *D* in Figure 8.5, but it's really *D**. If the owner is using a rule—no more than

Figure 8.5 Dead-weight loss from the wrong price.
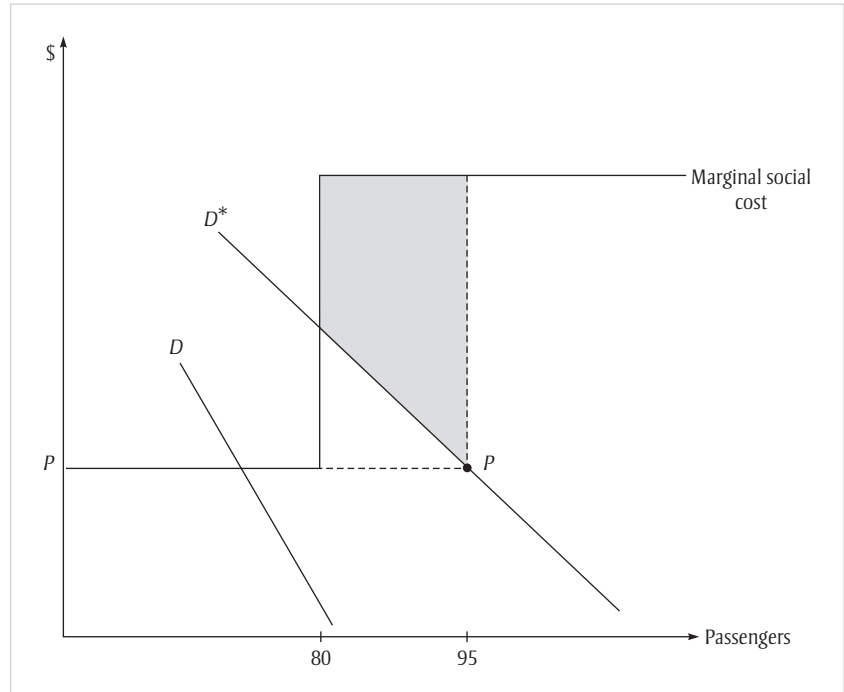
80 passengers—it will be surprised by how many passengers show up, but there will be no deadweight loss. Only 80 passengers will be allowed on board, and that's the right number.

On the other hand, if the owner is using a price to control the number of passengers and thinks the demand is D, it will set the price or tax at P. But since the demand curve is really D^*, this price will draw 95 passengers, which is too many. Deadweight loss will be the large shaded area. In the cruise ship case, as I indicated before, being wrong about demand has much more dire consequences when you're using prices than when you're using rules.

The opposite is true for the library. Once again, suppose the library believes the demand curve is D but it's really D^*, as in Figure 8.6. If the library is using prices, this mistaken belief makes no difference: it sets the price equal to P, which equals marginal social cost, no matter what it thinks demand is. Actual use of the copier is Q^*, which surprises the library but is optimal. If the library uses a rule, it sets quantity at Q, since that's what it thinks the optimal quantity is, and this creates the large shaded area of deadweight loss.

B. Information for Enforcement

The second argument for using rules rather than Pigouvian taxes concentrates on enforcement costs—how hard it is to find out what people have done in

Figure 8.6 Deadweight loss from the wrong rule.
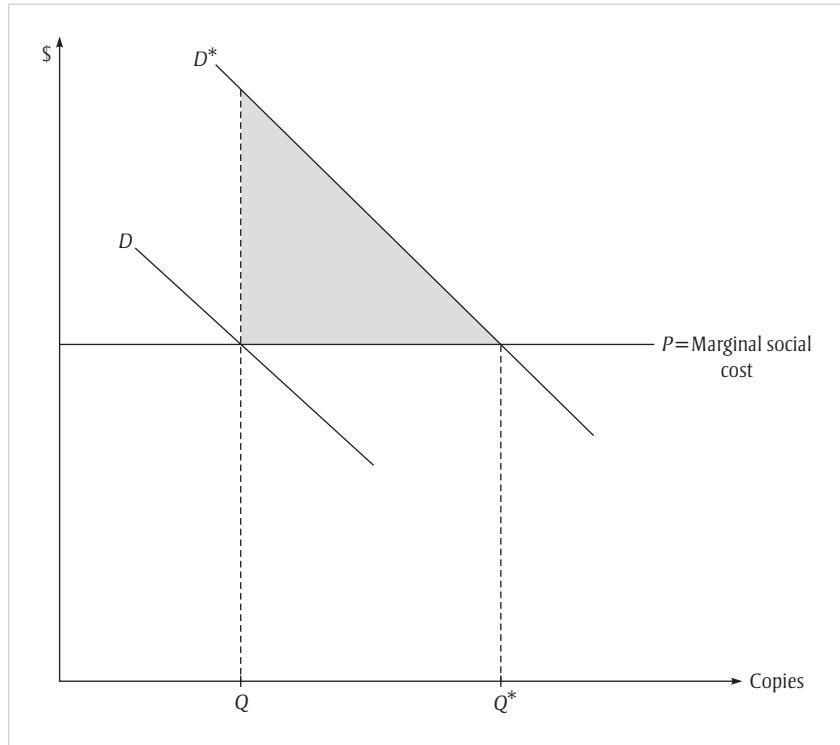
order to tax or punish them appropriately. Community reporting is cheap and
often important, and it's much easier for neighbors and the community in
general to know whether someone is breaking a rule than whether she has
paid appropriate taxes. Glaeser and Shleifer (2001) are responsible for this ar-
gument.

For instance, consider Sunday liquor sales. The externalities these cause
could be mitigated by either putting an extra tax on them or prohibiting them
(or prohibiting them during certain hours). If sales are prohibited, then it's
easy for neighbors, clergy, and passersby to tell whether a liquor store is in
compliance, and easy for a judge or jury to decide guilt or innocence. In con-
trast, a passerby on her way to church on a Sunday morning has no way to tell
whether anyone is actually paying the appropriate Pigouvian surcharge on li-
quor. Even an auditor poring over the liquor store's books a month later will
have a hard time being sure which sales occurred on Sundays, since cash regis-
ter records can be postdated or backdated, and trusted customers may not
have had to pay when they picked up their liquor (or the owner could have
formally sold himself a bunch of liquor on a Saturday night for distribution
on Sunday). Neighbors and churchgoers are much easier to find than super-
sleuth auditors, and they don't cost as much. Thus, so far as enforcement is

concerned, rules are likely to be a cheaper and more effective way of dealing with Sunday liquor sales than are Pigouvian taxes.

In general, then, when community monitoring is cheap and easy, and when tax collectors are likely to be lazy or corrupt, rules will often work better than taxes to enforce the desired behavior. Some other examples are hunting and fishing limits, rules on where buildings can be built, height limitations, restaurant and bar closing hours, noise ordinances, and laws on endangering the welfare of children.

II. Police Powers

City governments have the legal authority to adopt regulations of all sorts because they possess what is known as **police powers.** To understand the legal setting for the rules that city governments make, you have to understand these powers. The term does not refer (directly, at least) to the guys who wear blue shirts, carry guns and radios, direct traffic, and eat donuts; instead, police powers is a technical, legal term. A government with police powers has broad discretion to "promote health, safety, morals, and general welfare" in some community. It can forbid spitting on the streets and punish miscreants, stack up sandbags when a flood is imminent, and quarantine people with cholera or SARS. I don't have police powers, nor does General Motors—we can't order that a vacant building be boarded up or a vicious dog be destroyed—but the corporate body called the City of New York does. That's one important way that it's different from GM and me.

Specifically, Ernest Freund, one of the earliest legal commentators on the police power, cites two main attributes: "It aims directly to secure and promote the public welfare, and it does so by restraint and compulsion" (1904, p. 3). Police legislation differs from criminal legislation, he says, because "the peculiar province of the criminal law is the punishment of acts intrinsically vicious, evil, and condemned by social sentiment; the province of police power is the enforcement of merely conventional restraints, so that in the absence of positive legislative action, there would be no possible offense" (pp. 21–22).

Because regulations are often a good idea, it's not bad to have entities that can promulgate them. Because circumstances vary and change rapidly, and regulations need to adapt, the promulgating entities should have some discretion. Having someone around who has police powers can make all of us who are subject to those powers better off (even if occasionally each of us individually wishes for a special exemption).

Discretion and the power to coerce, however, are a dangerous combination. An entity with discretion and power can coerce all the rent for an entire city into its own hands simply by extortion—threatening terrible punish-

ments for anyone who does not pay enough. Discretion and power can even be self-defeating: if there are some investments that people will make only if they can be rewarded for them later, then those investments will never be made. The coercive entity could never convince people that anything would ever restrain its capacity enough to let them harvest those rewards. Instead of making us all better off, police powers, if untrammeled, could make us all worse off—possibly including even the entity with police powers—and prevent the realization of the economies of scale on which cities are based.

Police powers, therefore, are circumscribed in many different ways. One of the chief ways in the United States is constitutional restriction. The U.S. Constitution and almost all state constitutions have four chief clauses that limit governmental exercise of police powers: the due process clause, the equal protection clause, the First Amendment, and the takings clause.

The due process clause has two parts—procedural and substantive. Procedural due process requires that governments follow some sort of already prescribed tests before they exercise police powers against someone. Rules changes must be publicized, comment opportunities afforded, hearings held. Punishments can be meted out only after trials have been conducted. Some agent of the government may not just wake up one morning and say, "I don't like what you're doing" or "Go to jail." Procedural due process can't keep governments from doing nasty and vicious things, but it can keep them from doing them quickly and with an element of surprise. Procedural due process also requires that a fair number of different individuals be involved in government actions that can seriously harm people. Bigger conspiracies to violate someone's right are harder to put together and keep together than smaller ones.

Substantive due process requires that the rules be reasonably related to the purpose of police powers—to promote the health, safety, morals, and general welfare of the community. Rules can't be adopted for reasons of whimsy, revenge, the personal enrichment of politicians, or the glory of God. Public purposes, however, are often broadly construed—courts have ruled, for instance, that expanding casinos is a public purpose in Atlantic City—and "reasonable relation" is a low hurdle. Some plausible argument has to be advanced for the gross benefits (not net benefits) of the rule in question; nobody has to convince anyone it's the best possible rule. Like procedural due process, substantive due process makes abusive coercion somewhat more difficult, but not impossible.

Equal protection says that rules must be written and applied impersonally. You can make different rules for different groups, but the definitions of the groups must have some "reasonable relation" to the public purpose being promoted (if the group is a protected class, like African Americans or women, the standard is higher). Substantive due process and equal protection work together to help protect people against personal vendettas and extortions.

The First Amendment protections of religion, speech, assembly, and po-
litical activity also circumscribe some police powers. But the protections are
far from absolute—fraud can be prohibited, and so can hanging political signs
on telephone poles. Rules that infringe on First Amendment rights have to
pass a higher standard than having a "reasonable relation" to some public pol-
icy objective; the government has to show a "compelling interest." Arguments
that support regulation of food stores won't necessarily support regulation of
book stores.

The final constitutional check on police powers is known as the takings
clause. It's in the Fifth Amendment: "Nor shall private property be taken for
public use, without just compensation." This refers to an ancient practice
called **eminent domain.** When a government is building a road or a firehouse,
it can force landowners to sell it the property it needs—but it must pay them a
fair price. "Fair" is what a judge says is fair.

The connection with police powers is that a government agency can't
force you to turn over your store or your driveway unless the government
compensates you. Nor can a government adopt a series of regulations for the
sole purpose of making your property less valuable so it can acquire that
property more cheaply (even if it's for some good project). That much is clear.
Much else about the takings clause is unclear.

The problem is that property is a "bundle of sticks," as I explained in
Chapter 7, and there is no hard-and-fast rule about which sticks or how many
sticks have to be removed from the bundle before the takings clause applies. If
you lose the right to hold cockfights in your backyard or to pay workers $2.00
an hour to work at your sewing machines or to rent out a third-floor apart-
ment without a fire escape, you've surely had sticks removed from the bundles
that constitute your backyard, your machines, and your third-floor apartment.
But almost no courts and almost no legal scholars would hold that a compen-
sable taking had occurred. On the other hand, the Supreme Court has held
that forcing the owner of an apartment building to install cable television, for
example, and a restaurant owner to permit a bicycle path through her parking
lot, are instances of property taking and thus are compensable.

What is the general principle that sets these cases apart, or should set
them apart? Many eminent legal scholars and economists are trying to figure
this out, and so is the Supreme Court. The important point is that the takings
clause provides some sort of protection for citizens against aggressive govern-
ments with police powers, but as with the other constitutional clauses, the
protection is not complete. And no matter how the Supreme Court rules in
the next few years, the takings clause isn't going to be much help for people
without property. Police powers are inherently discretionary; anything discre-
tionary is subject to abuse.

That's why constitutional limitations aren't the only constraints on police

powers. Instead of trying to limit governmental discretion, other constraints try to change the incentives of the entities exercising police powers so that they have little or nothing to gain from abusing their discretion.

Democracy, for example, makes it unlikely that governments will long use their police powers in an obvious way to abuse a majority of voters. This is not insignificant. Democratic governments may have the discretion to use their police powers in this way, but they usually don't want to. At the same time, however, democracy offers only porous protection for minorities, outsiders, and people who can't vote.

Another way to change the incentives of a government with police powers is to make it difficult for the government to gain much from using them. The idea is the same one that led ancient kings to employ eunuchs to guard their harems: if people can't gain much from abusing their power, they won't abuse it.

For governments with police powers, the same kind of safeguard comes from the prohibition on taking money for making rules or making exceptions to them. A government you could buy rule exceptions from would have a lot of incentive to become a protection racket: the more rules it made, the more exceptions it could sell, the more money it could make to pursue whatever goals it wanted to pursue. If a government can't take money, the ways it can gain from its power to make rules are more limited; the same is true for the ways it can gain from abusing its power to make rules. Limiting the ways a governing body can gain from abusing its power makes it less likely to do so.

III. Zoning

A. What It Is

The type of urban regulation that economists have studied the most is **zoning.** Zoning is "the division of a community into districts or zones in which certain activities are prohibited and others are permitted" (Fischel 1985, p. 21). Zoning generally deals with large, hard-to-move, long-lasting objects, like buildings, trees, and parking lots; once in place, they don't require frequent monitoring.

In each zoned district, zoning codes spell out what kinds of buildings are permitted or prohibited—single-family houses or apartment buildings, gas stations or laundromats, office towers or bodegas, convents or fast-food restaurants, drug rehab clinics or junkyards. Zoning codes set limits on the size, bulk, and location of buildings: how high they can be built, how far from the curb and from other property lines they must be set back, how much total floor area they can encompass. Zoning can set requirements for off-street parking spaces, for lawns and trees—even, in some cases, for the color of

buildings. The size of a lot, the number of buildings on it and the relationships between those buildings are also the subject of zoning codes. Zoning thus is a series of rules.

Zoning codes are not absolute; several different kinds of exceptions are possible. A variance permits a specific exception to a specific provision: a building thirty-five feet high in a district where only thirty-foot buildings are normally permitted, a Jacuzzi that will come within four feet of a property line when a six-foot setback is required, a lawyer's office in a residential district. Special permits (or conditional uses) deal with activities that are generally allowed in no zone. Gas stations, video game arcades, fast-food restaurants, and drug rehab clinics are examples of uses that often fit this description; whenever they are permitted, it is an exception. "Nonconforming use" refers to buildings or activities that started up in a zone as permitted but became prohibited when the zoning was changed (or was first implemented). Many towns allow nonconforming uses to continue so long as they don't expand or change drastically. Finally, towns can change the boundaries of zoned districts and the rules that apply.

People who want to obtain variances or special permits apply to the town's zoning board of adjustment (or sometimes the planning board). The board of adjustment is a group appointed by the town's governing body. It has discretion—it can approve, disapprove, or approve subject to conditions, and it doesn't have to give detailed reasons—but it is bound by procedural due process. Notice must be given, hearings must be held, neighbors must be given a chance to object. Most of the time, however, it's fairly easy to predict how a board of adjustment will act—if nobody objects, it will approve the variance or special permit; if a lot of neighbors object, it won't.

Zoning came into use in the United States in the beginning of the twentieth century. Frankfurt am Main in Germany was the first major city to establish different zones and enact different regulations for each; that was in 1891 (Ellickson and Tarlock 1981, p. 40). New York City's zoning code of 1916 is usually thought of as the first in the United States because it was comprehensive, received great publicity, and was widely imitated. To be sure, Chicago in 1893, Washington in 1899 (to make sure the Capitol was not overshadowed), and Boston in 1906 enacted building height limitations before New York (in fact, the New York code included no height limitation per se); Los Angeles in 1909 designated some zones of the city as industrial and some as residential (ibid., p. 39). However, these earlier American ordinances were in many ways closer to building and fire codes, which I will discuss later, than to the zoning codes that followed New York's code of 1916, and they did not have the impact that New York's did.

The 1916 New York zoning code was heavily supported by progressives and reformers, who saw it as a way of harnessing untrammeled capitalism and

making the city more beautiful and livable, but three specific economic problems made some key businesses support zoning as well, and it was this business support that was chiefly responsible for the code's adoption.

The first economic problem was sunlight. Office buildings needed it. Willis (1995, pp. 24–26), in her study of skyscrapers, explains:

> Until the introduction of fluorescent bulbs in the 1940s, sunlight was the principal source of illumination for interiors. Although electric wiring was universal by the 1890s, incandescent bulbs remained weak and inefficient. The main form of artificial light, desk-top lamps, produced about three to four foot-candles (one foot-candle is the amount of light on a surface held one foot away from a burning candle). Outdoors, summer sunlight radiates about 10,000 foot-candles, and even cloudy days can provide 200 to 500 foot-candles . . . Indirect lighting from large ceiling lamps was considered the most healthful type of illumination since it produced the fewest and faintest shadows, but using multiple light bulbs added considerable heat to a room—one 500-watt bulb gave off the equivalent of one pound of steam per hour—which posed problems in hot weather.
>
> The quality and rentability of office space thus depended on large windows and high ceilings that allowed daylight to penetrate as deeply as possible into the interior. Ceiling heights were at least ten to twelve feet, and single windows were as big as possible without being too heavy to open . . . Many buildings had mechanical ventilating and cooling systems (there was no air conditioning until the 1930s), but open windows provided the fresh air vital to comfort, especially in summer months.

Because of the need for natural light and ventilation, offices were hardly ever designed to be more than twenty to twenty-eight feet deep.

If a big, tall building were built next to yours, it could therefore pose a serious problem for you. Nor were those affected only the immediate neighbors (a situation that could be resolved through Coasian bargaining): the Equitable Building at Broadway and Cedar Street, for instance, rose up forty stories straight from the street when it was completed in 1915, and it cast a shadow over four blocks—four of the most valuable blocks in the United States.

The second economic problem for New York businesses was overbuilding. In 1916 the New York office market was going through one of its periodic episodes of high vacancy rates (exacerbated by the 1.2 million square feet of new space for rent in the Equitable Building, the world's largest office building at the time). Owners of existing buildings wanted to prevent massive new construction that would drive down their rents and drive up their vacancy rates. (Chicago's 1893 height limits were also adopted during a time of slack demand in the office market.)

The final economic problem concerned garment factories and, to some extent, insurance companies. Garment factories employed many poor immi-

grant women; insurance companies employed many women of somewhat higher social standing, but they were still working-class women. Owners of the high-toned stores that lined Fifth Avenue were afraid that if garment factories and insurance companies were allowed to operate too close to their businesses, the working women would overwhelm their wealthy clientele, and their location would lose its cachet. Who would want to shop at Tiffany's if it were surrounded by crowds of poorly dressed Jewish and Italian women getting off work at the garment factory?

These interests coalesced to produce the 1916 New York code, which divided the city into three kinds of districts—residential, commercial, and unrestricted—and ensured that no factories could operate near Fifth Avenue. Office buildings were restricted to a maximum floor-area ratio—how much space they could build on a lot of given size—and subject to rules about how high up a facade could extend at the lot line, and how quickly it had to step back after that. There were no height restrictions on towers, but they could cover only a quarter of a lot. The 1916 zoning code thus created the ziggurat-like structures with multiple setbacks and tall spires that form the classic New York skyline.

Zoning spread rapidly throughout the United States after New York adopted its code. The political forces were somewhat different outside New York. For example, owners of single-family homes, who hoped to protect their investment against all sorts of encroachments, were particularly important. But even industrialists liked zoning. Common law allows people to sue to enjoin "nuisances" from operating, and industrialists hoped that by operating in zones explicitly designated for manufacturing they would be protected from nuisance suits (Ellickson and Tarlock 1981, p. 41).

The federal government also contributed to the spread of zoning in the 1920s. Herbert Hoover, then secretary of commerce, convened a group in 1921 to help states establish zoning laws; this group produced the Standard Zoning Enabling Act (SZEA). With the SZEA, states wishing to let their municipalities write zoning codes didn't have to reinvent the wheel; they could simply appropriate the language and procedures that the Commerce Department had developed. By 1930, thirty-five states had adopted the SZEA; almost every state had laws similar to the SZEA by 1960.

The other important federal action regarding zoning in the 1920s was judicial rather than executive: the Supreme Court's *Euclid v. Ambler* decision in 1926. The Court held in this decision that zoning was a constitutional exercise of police powers. Euclid, Ohio, is an independent suburb of Cleveland, about ten miles from the city's central business district. Ambler Realty Company had bought a large tract of vacant land there. Situated between a railroad and a major thoroughfare, the parcel was ideal for industrial development, and that was Ambler's goal. In the early 1920s, however, Euclid adopted a zoning ordinance that placed Ambler's parcel in a residential district, which substantially

reduced its value. Ambler sued to overturn the ordinance as a violation of due process.

Ambler had a reasonable chance of prevailing. State courts in Maryland and Texas had recently ruled that the objectives zoning sought to accomplish were too dissimilar from the traditional objectives of police powers for zoning to be constitutional. The federal district court in Ohio agreed and ruled in favor of Ambler in 1924. "The result to be accomplished [by Euclid's law] is to classify the population and segregate them according to their income or situation in life," the judge wrote in that decision.

Euclid appealed, but the U.S. Supreme Court's preliminary decision upheld Ambler by a vote of five to four. Justice Sutherland, however, a member of that majority, had doubts about the decision and asked for more time to think about the case. Eventually Justice Sutherland changed his mind and wrote the majority opinion upholding Euclid's zoning ordinance. Police powers were not static, he wrote, and zoning was a legitimate exercise of police power so long as it bore a "reasonable relation" to "health, safety, morals, and general welfare."

(Although this litigation may have had long-lasting consequences for zoning in the United States, it does not seem to have made much difference in the development of Euclid. Ellickson and Tarlock [1981, p. 51] report that several years after the Supreme Court decision, the Ambler parcel had been rezoned for heavy industry and a manufacturing plant had been built on it.)

The half century following *Euclid v. Ambler* was remarkably stable for zoning in the United States. Codes everywhere were similar, and justifications bowed strictly to the SZEA and Euclid language. Beginning in the 1970s, new concerns—the environment, aesthetics, historical preservation—began to animate the people who wrote zoning codes, and the courts began to accept these new interpretations of police power. The structure of zoning codes changed, too. Residences, instead of being allowed everywhere, began to be prohibited in commercial and industrial zones; more special permits were added; greater opportunities for discretion and negotiation were incorporated in codes under headings like "planned unit developments" and "university zones." But the basic framework created by New York, the SZEA, and *Euclid v. Ambler* have survived.

B. Rules Rather than Taxes?

Why not use Pigouvian taxes to restrain the activities zoning is concerned about, force people to bear the full social costs of their decisions, and then let them decide on their own what to do? In other words, why not use prices instead of rules?

Recall the principle from section I: rules work better when the marginal benefit to the builder (or the person regulated by zoning) is relatively con-

stant, and when the marginal cost to the public is increasing (relatively)—that is, not constant. To what extent do these conditions apply to the activities that zoning typically regulates?

For the most part with zoning, the condition of a relatively constant marginal benefit for the builder seems to be satisfied. An additional floor in an office building or another townhouse in a suburban development or another square meter of buildable footprint on a lot is going to provide the builder with about the same amount of increased profit (or increased consumer surplus) as the last one did.

The changing-marginal-external-cost condition is also met in many traditional zoning contexts, but not in all. In the 1916 New York skyscraper problem, if a new building is being constructed next to an existing twenty-story tower, the marginal cost of most of the floors below twenty may not be too great: the existing tower loses some sunlight, but many of the other neighbors have already had their sunlight blocked by the existing tower; whether the new building next to it has fifteen stories or sixteen stories makes little if any difference to them. If the new building is bigger than the old one, however, the consequences are different. Then going from twenty-five stories to twenty-six stories will blot out the sunlight of additional neighbors who were not affected by the existing building. The marginal external cost is not constant in this case: it changes once the building becomes the tallest one around.

Noisy, smelly factories are similar: the first one on a residential block is going to cause a lot of damage; with the second one, nobody may notice the difference. Residential neighborhoods may also work the same way: adding a one-family house on a block of one-families hurts the neighbors by increasing congestion and taking away some privacy, but adding a two-family house hurts more than twice as much because it brings in neighbors of lower income with whom existing neighbors may not be so eager to socialize.

The genius of zoning—as opposed to similar prohibitions on activities—lies in recognizing that, in many cases, deviations from homogeneity cause big changes in the marginal external cost schedule. Taxes and pricing are not a good alternative to zoning in these cases.

It's not clear, though, that all instances of zoning satisfy this criterion. The most questionable are the controversies over large, relatively isolated suburban developments. The marginal external cost of the one-hundredth house on a 100-acre tract does not seem to be much different from the marginal external cost of the four-hundredth house or the eight-hundredth house. Taxes and fees might work better than zoning regulations in these situations.

Sometimes, in fact, prices are used instead of rules in controversies over developing large parcels of land. Exactions, which I discuss below, are one example. The best example is parks. Open space has many benefits for a town, and so developing open land causes harm external to the developer. The marginal external harm per acre of developed land, however, doesn't change much

as the size of the development increases, and so preserving open space is usually done by paying the owner—buying the land—rather than making rules. Towns that restrict development this way bear most of the cost themselves, and so are likely to do so efficiently.

C. Analyzing Zoning

Think of the description of property rights as a bundle of sticks. Zoning is a way of assigning sticks to bundles. A zoning code assigns sticks such as "the right to build a thirty-story apartment building" or "the right to operate a convenience store" to certain pieces of land and withholds them from others. Some pieces of land get sticks like "the right to object if the neighbors want to install a swimming pool," or "right to object, and probably prevail, if anyone on the block wants to open up a carwash"; others do not. Seen in this light, as economists would like it, zoning appears to be not so much an exercise of police powers as the simple administrative tidying up of title claims.

The Coase theorem says that if certain conditions hold, it doesn't matter how this tidying up is done as far as the availability of potential Pareto improvements is concerned. So long as everyone knows which stick is in which bundle, people will negotiate the efficient use of every stick. But for almost all zoning matters the Coase conditions do not hold, and so zoning is more than a simple tidying-up of property claims. The conditions on negotiating and sale are usually violated: often, many parties—too many to negotiate effectively—are affected by the activities zoning addresses, and almost always, buying and selling are prohibited because zoning is a police power.

To see the consequences of these violations of the Coase theorem, consider four different zoning concerns: the three that led to the 1916 New York ordinance, and the frequent contemporary suburban concern about "overdevelopment" and maintaining a community's character.

The Coase considerations are strongest in the case of sunshine. If all I am concerned about is whether my neighbor will build a tall building and block my sunlight, then in the absence of zoning I can negotiate with him and bribe him not to do so, if this outcome is efficient. If I cannot bribe him not to build, then it is efficient that he build. Zoning in this case is just a reassignment of sticks, a transfer from my neighbor to me: he loses the right to build a very large building, and I gain the right to (a certain amount of) sunlight. It should not be surprising, then, that big-city height limitations and zoning codes have tended to be adopted during the busts after office building booms: at those times few large owners consider the right to build a important part of their portfolio, but many consider the right to sunshine for existing buildings to be extremely desirable, since the volume of existing buildings has just grown so much during the boom. When the market turns, this reasoning of-

ten reverses and height limitations do, too. For example, Willis (1995, p. 168) shows Chicago height limitations over time on the same graph as annual additions to office space; there is considerable coincidence between the two series.

The New York sunlight example makes the interpretation of zoning as a transfer even clearer because it is one of the few cases in the United States in which people can buy and sell a right created by zoning. In downtown and midtown Manhattan, if you own a building that is shorter than the maximum height allowed by zoning at that particular location, you can sell to any adjacent property owner the right to build the floors you didn't. This is called a sale of air rights. For instance, if adjacent properties are permitted buildings of one million square feet but one property holds a building only 600,000 square feet in size, then the owner of that building can sell the "leftover" 400,000 square feet to the other property owner, who would then be permitted to build up to 1,400,000 square feet. If I own the smaller building, and constructing the large building is a potential Pareto improvement, the contrast between zoning and no zoning is stark: under zoning the neighbor pays, and without zoning he doesn't. If the big building does not offer a potential Pareto improvement, the contrast is similar: under zoning nothing happens, but without zoning I pay him (not to build).

It's only when we consider large numbers of neighboring properties— the Equitable Building's four-block-long shadow—that zoning appears to be something other than a transfer. These neighbors would find it difficult if not impossible to coordinate a bargaining strategy. Even if the loss the Equitable Building imposed on its immediately adjacent neighbors was less than the profit and consumer surplus it generated, building the Equitable would still create a net loss if the losses of the nonadjacent neighbors were large enough. So zoning could create a potential Pareto improvement if it kept buildings that harmed large numbers of nonadjacent neighbors from being erected.

But then again, it might not. The harm that the distant neighbors suffer, while real enough, may be trivial compared with the benefits a large new building provides, and there is no obvious way of finding out which is the case.

(Notice that restricting air-rights sales to immediately adjacent neighbors fits in with this rationale for zoning. Suppose A and B are located two blocks from C and D. Whether A sells air rights to B or not should make little difference to C and D—so long as the total bulk that A and B construct is the same, small differences in the exact location of that bulk are not going to matter much to owners two blocks away. If A were to sell air rights to C, however, D would be much worse off. Probably this is a major reason why such sales are prohibited.)

The second problem that the 1916 code addressed—fear of overbuilding—is a good example of how zoning can promote inefficiency. Overbuilding

and the consequent fall in rents harm large numbers of property owners, and so losers from overbuilding make up a group almost certainly too large to organize itself to offer appropriate bribes to keep new buildings from being built. Here, zoning can fill a gap that Coasian bargaining cannot. But economic theory tells us that this is a gap that should not be filled. Tenants gain the exact amount that owners of existing office buildings lose when rents fall. Tenants are probably more numerous and hence even more difficult to organize than owners, and they may not realize that a big new building like the Equitable helps them in the long run, even if they never contemplate moving there.

To see this traditional result, look at Figure 8.7, which illustrates how zoning prevents overbuilding in the office market of a particular city. Without zoning, the quantity of office space would be *Q*, and the price would have to fall to *P* to clear the market. If zoning prevented the stock of office space from expanding beyond *q*, prices could be maintained at the higher level *p*, and owners would gain the additional profit represented by the shaded rectangle in Figure 8.7. For owners, the construction of additional office space between *q* and *Q* would represent overbuilding because it would cause them to lose the shaded rectangle.

Figure 8.7 Preventing "overbuilding" causes deadweight loss.
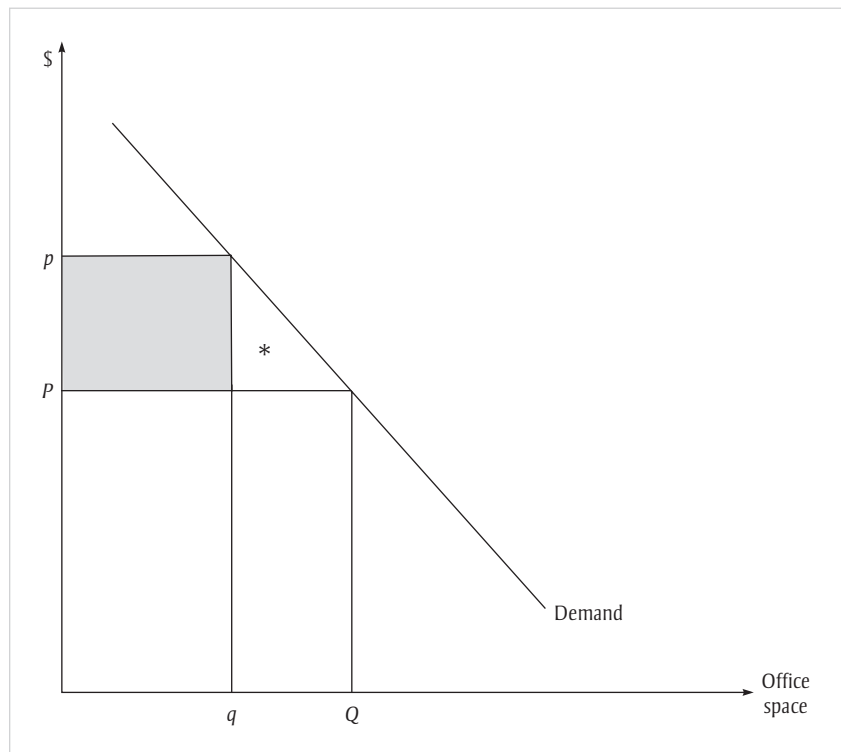
The shaded rectangle represents a direct loss for the tenants that rent office space, however. Every dollar that building owners gain from the price rise from P to p is a dollar that tenants lose—and tenants get nothing in return for the higher rents. Zoning is inefficient in this example because the higher prices and lower quantity of office space prevent additional tenants from renting space—the consumer surplus represented by the triangle with the asterisk in it. This triangle represents the deadweight loss from zoning designed to combat overbuilding.

Preventing overbuilding, then, is an inefficient way for owners of existing office buildings to transfer income to themselves from tenants and from the owners of land on which potentially profitable buildings could be erected. Everybody would be better off if the new big buildings were built, rents were forced down, and tenants and the owners of the new buildings were compelled to compensate the existing owners for their losses. But zoning doesn't act that way—it's a police power.

The problem of the Fifth Avenue shops is similar. Without zoning, poor women have a right to stroll on Fifth Avenue, and their employers have a right to locate factories close enough to Fifth Avenue that their workers would be tempted to stroll there. To get rid of the poor women, the Fifth Avenue merchants would have to get together and either pay the poor women to leave, or, more likely, pay the garment factories and insurance companies to operate elsewhere. With zoning, they don't have to; Fifth Avenue property owners have added a new stick to their bundle—the right to keep poor women away.

Is this efficient? Could the winners compensate the losers? Nobody knows. Maybe it is efficient; maybe it isn't. The important point is that if it were not efficient, there is no corrective mechanism to undo it. Removing the possibility of sales and compensation removes the corrective mechanism that tends to get rid of inefficiencies.

The fourth zoning problem, zoning in the suburbs, is the type of zoning that has attracted the greatest intellectual attention in the past several decades, but the basic concerns today are no different from those of 1916 Manhattan. Adding more people to an existing suburb hurts existing residents in several different ways.

First, residents may enjoy the undeveloped land that would be lost if new houses were built. Farms are pretty to drive by, and the vacant lot down the street may be a good place for kids to play. Abutting a golf course means you have privacy almost all the time. I would rather look at and wander through trees and forests than houses and streets (especially new houses and streets, since good trees in residential neighborhoods take decades to develop); a lot of other people would too. Rabbits never use noisy leaf blowers on Sunday mornings.

Second, nearby construction is itself a serious disamenity. Big trucks roaring down the street, loud noises early in the morning, dust, strange and

dirty people eating lunch almost wherever they please, machines and half-constructed buildings left overnight as alluring but dangerous play areas for children—all of this makes many current residents leery of any construction project whatsoever.

Third, development increases congestion. Roads, parking lots, supermarkets, and schools all get more crowded, at least in the short run.

Fourth, existing residents may find they must pay for various services, if those services are not priced at marginal cost. Schools, police, and fire suppression will all become more expensive, and the formulas that allocate the costs of these activities may place some of the burden of the increase on existing residents. For instance, if the property tax pays for these services, the total cost is proportional to population, and if new residents own on average less taxable property than existing residents, then existing residents will pay more for the services, although the quality will not improve.

Finally, existing residents may not want to associate with the type of people they think the new residents will be, or be associated with them. For many people, passing on their values to their children is an important goal, and since children acquire many of their values from friends in the neighborhood and in school, living in a community where practically everyone shares the same values may be the only way of accomplishing this goal. New residents may not share values, religion, language, or skin color with existing residents, and so it may be hard to integrate them into the civic and cultural life of the town. Crime, fighting, and discord may rise; your children may be picked on at school. Educational achievement may fall if teachers need to spend more time on discipline problems and if the new parents aren't as well educated and so can't help their kids as much. The name of a town may connote something—wealth, prestige, artiness, ecological consciousness, respect for history—and the existing residents may value being able to advertise that they share this characteristic when they write their address; an influx of new residents would jeopardize the value of this advertising. Or existing residents may just be bigots and snobs.

Of course, there are offsetting benefits from the increasing returns to scale created by growth. A larger population promises more specialized stores and services, more competition, better mass transit, more potential friends and partners who live nearby—all of the benefits of city life. But for reasonably wealthy people in the suburbs of a large metropolitan area—people equipped with cars, telephones, faxes, and computers—most of these benefits may be realized already outside of the town they live in. People who live in Maplewood in northern New Jersey go to Livingston for supermarkets, Millburn for Chinese restaurants and running equipment, South Orange for Indian restaurants, West Orange for Mexican restaurants, Union for hardware, Edison for clothing, and Manhattan for work and concerts. The marginal benefits to them of additional population in Maplewood, where they live,

are likely to be small—too small in many cases to offset the losses they would suffer.

Suburban zoning is thus designed in many instances to make it difficult for new residents, especially those not as rich as current residents, to move to a town. This is sometimes called exclusionary zoning. Suburban towns frequently will zone little or no land for apartments and other multifamily dwellings, and those apartments that are permitted usually must meet stringent standards for off-street parking. Building an apartment over a store is impossible in most cases. Single-family homes must have large lots, off-street parking, and in some cases a minimum bulk. Some communities ban mobile homes. When they zone a plot of land for commercial and industrial purposes that are not viable, it's just about the same thing as zoning it for vacant land.

Zoning puts the right to a quiet, low-density, high-income environment in the bundle of sticks attached to already-developed land in the suburbs. Current residents vote, and the people they elect decide on zoning. Possible future residents of a suburb don't have a direct say in that suburb's zoning. The benefits they would get from moving there are (partially) reflected in the price of developable land, but the owners of developable land are likely to be a distinct minority in a suburb that is already half developed (since undeveloped land has, almost by definition, fewer votes per acre than developed land). Like the owners of existing office towers in 1916 Manhattan, owners of developed suburban land want to pull in the welcome mat once they have entered. Zoning lets them do this.

As in the Manhattan cases, there is no pressure for suburban zoning to be efficient. If developers and future residents could pay for zoning standards to be relaxed, then current residents could not costlessly impose standards—they would have to forgo payment every time they raised or enforced a standard. Current residents would choose to squelch only those proposed developments that hurt them more than they benefited outsiders and landowners; the others they would be paid for and would approve.

But zoning is a police power and cannot be sold. Zoning, for current residents of half-developed suburban towns, is like inheriting 300 gallons of ice cream and being forbidden to sell any of it. Ice cream is very good, and you would fight hard against anyone who tried to deprive you of this inheritance. Because of the terms of the inheritance, ice cream will be free for you (even though your eating is socially costly) and you will overeat it by an inefficient amount.

In sum, there is good reason to believe that many partially developed suburbs zone too restrictively. Notice, however, that this conclusion doesn't apply to largely undeveloped areas—farming communities, for instance. In those communities, large numbers of residents may see themselves as owners of potentially developable property, and so would bear some of the burden of overly restrictive zoning.

D. Fungibility

The conclusions about the inefficiency of zoning depend crucially on the premise that zoning cannot be bought or sold. The technical way of expressing this premise is to say that zoning is not **fungible.** The SZEA and *Euclid v. Ambler* imply that zoning should not be fungible by defining it as a police power. *Should,* however, does not always imply *is.* We have already seen that New York air rights are fungible, for instance, at least among adjacent properties. Is the language in the SZEA and *Euclid* merely pious rhetoric that doesn't describe how zoning really works?

Probably not. Although money sometimes changes hands and a lot of wheeling and dealing occurs in actual zoning practice, almost all of the time, especially in suburbs, the transactions differ significantly from those that the Coase theorem identifies as necessary for efficiency. Four different kinds of agreements can be reached, but none of them resembles a Coasian bargain.

1. Bribery

Bribery of zoning officials, for instance, is not uncommon (though it may be less prevalent now than it used to be), but even the most widespread corruption of this sort would be unlikely to produce the right kind of fungibility. Bribes are not made in the right amounts. The bribe a developer has to pay a corrupt politician is bounded below by the expected punishment the politician could incur by taking the bribe and the value of his moral qualms; the loss to the people who object to the zoning action—the important quantity for efficiency—never enters into the calculation. The developer's gain has to be large enough to satisfy the politician, not the public. So even if bribery were more widespread than it probably is, zoning would not be fungible in the right way.

2. Developer–Neighbor Agreements

Agreements or understandings between a developer and a property's neighbors are common, and they at least involve the right parties. Developers who want to obtain variances or special permits can sponsor Little League teams, give discounts and extend credit, contribute to charities, and generally act like good neighbors. Whether or not those involved think of it in these terms, to an outside observer one party appears to be paying the other not to object to what it wants. But such tacit arrangements fall short of the easy transactions that the Coase theorem requires. They may not work because the neighbors may be ungrateful or because people from another part of town may also be on the zoning board. The value given may not be the same as the value received, since cash is not changing hands. And the variance or special permit goes to the property, not the owner, and the owner cannot guarantee that he or she will not sell, retire, or die. Neighbors may like the owner of the corner

gas station very much because he gives free advice, lets kids use the air pump, and plows out some driveways, and they'd be inclined to support a variance so he could open a convenience store, but they may have doubts about how the next owner would likely handle the added traffic.

3. Noncash Developer–Government Agreements

The same sort relationship may exist between a developer and a local government. Developers may dedicate a certain amount of land to public use or simply give it to the government, renovate subway stations, build recreation centers, or make generous gifts to local charities. And the same sorts of problems bedevil these sometimes more explicit agreements.

The first kind of problem with these agreements is that their enforceability is dubious. What is known as the **reserved power doctrine** holds that contracts for police powers are unenforceable (see, for instance, Kmiec 1981, pp. 105–115). The U.S. Supreme Court enunciated this position most starkly in the 1880 decision *Stone v. Mississippi:* "The legislature cannot bargain away the police power of a State." While the exact implications of this doctrine are in flux now, as courts are issuing conflicting decisions, no one can feel certain that a contract for improvements in return for zoning will stand up in court. This uncertainty reduces fungibility, because in every deal, one party has to act before the other party has fully discharged her responsibilities. If the town lets the developer build her houses before she builds the public community center she promised, she may renege or walk away or declare bankruptcy after the houses are finished but before the community center's done, or she may build a chintzy community center. If the developer builds the community center first, before the permits are issued for the houses, the town leaders can renege or succumb to popular pressure or be voted out of office, and the developer will never get the permits. Deals without either contracts or cash are tough.

Residential development (as opposed to commercial or industrial) has an additional enforcement problem as well. After the new houses are built they will contain new voters, and these new voters will not be interested in compensating the old voters. The new voters can become a powerful enough force that the town government is no longer interested in acting on behalf of old voters to see that the old promises are fulfilled.

For instance, in New Jersey in the 1970s, a number of towns approved townhouse and condominium developments on the condition that the new residents would pay for their own garbage collection. Everyone else in these towns had their trash picked up by the municipality at public expense. Once the developments were built, their residents became active politically and fought, successfully in many cases, for the extension of municipal service. Old residents cannot commit the local government to look after their interests when they are no longer an overwhelming majority of voters. After several

years, new-development residents were a large enough group statewide that they could persuade the New Jersey legislature to rescind unilaterally all the private garbage collection agreements that had not already been abrogated locally.

Another problem with these agreements is that they are barter, not cash. This is an even more severe problem for developer-government agreements than for developer-neighbor agreements, because the commodities that can be bartered are more severely restricted. The residents of a town might really want help with their backyard gardens, but the developer can promise only to plant a rose garden in the park. In the 1980s, New York City frequently bartered additional bulk for the construction of public plazas around office buildings, but by the late 1990s these plazas were probably the most universally reviled feature of the Manhattan cityscape.

Notice that the barter problem hits hardest in the type of community where inefficiently restrictive zoning is most likely to be a problem—partially developed suburbs. Enforcement problems effectively ensure that public capital improvements are the only coin developers can use to compensate current residents, and the marginal benefit of public capital improvements may well be lower in these communities than in undeveloped areas or in older cities. Public infrastructure may be sparse in undeveloped areas, and so marginal increments would likely be more valuable there. In older, denser areas, infrastructure may be deteriorating, and residents may rely more on publicly provided services—parks instead of backyards, subways instead of cars, hydrants instead of backyard swimming pools, concerts in the park instead of television on the sofa, public trash collection instead of private. Partially developed suburbs are likely to have good, fairly new infrastructure and residents who are used to cars and backyards. The marginal benefit of a second community center in such a town can be pretty small, especially when most residents have stereos, swing sets, computers, and ninety-six-inch televisions at home.

4. Exactions or Impact Fees

An exaction (sometimes called an impact fee) is a money payment that a developer makes to a local government to cover additional capital costs—roads, hydrants, schools, and so forth—that a development will place on the municipality. Exaction amounts are not supposed to exceed those costs, but since the amount is hard to calculate, there is some room for negotiation. In 1985, about three-fifths of communities in the United States received cash exactions from developers. Four-fifths of the fastest growing communities received these fees (Altshuler and Gómez-Ibáñez 1993, p. 36). These proportions have probably risen since then.

Because they are paid in cash before the development takes place, exactions don't have the enforcement problems that plague agreements to build community centers or pick up garbage. But at the level of individuals—the

level that counts—these agreements are still a form of barter, even though cash changes hands at the government level. The individuals whom a project affects may never see better private goods for themselves (or even better ongoing public services); they only get to see better public capital. It really doesn't matter whether the contractor who builds the new community center gets his check from the town government or from the developer directly. (If the town illegally uses the exaction to lower taxes, the people most directly affected by the development will get some money, but so will taxpayers who are not affected at all, and the developer has no way to concentrate the money on those taxpayers who will suffer the greatest losses.)

More fundamentally, exactions are supposed to deal only with the harms that development causes by adding to the load on existing infrastructure (mispriced existing infrastructure, to be precise, since marginal cost pricing would obviate the need for these payments). Zoning, though, is supposed to protect the community against many other negative externalities—loss of sunshine, for instance. If you can pay for only part of something, the whole is not fungible. You can buy a train ticket to Boston, and getting to Boston is part of what you have to do to win the Boston Marathon, but that doesn't mean that Boston Marathon medals are fungible.

Thus, saying that zoning is not at all fungible is no doubt an overstatement. So is saying that winters in Canada are always cold and snowy and that cats never take showers. All of these statements, though, are useful approximations.

E. Sprawl and Zoning

Popular discussions about sprawl often ascribe special importance to zoning, but not in a completely consistent fashion. On one side, large-lot zoning is often portrayed as a villain contributing to sprawl; on the other, zoning in the form of tougher land-use regulations is often seen as the hero who is going to save us from sprawl. We need to sort out these claims systematically.

Do the inefficiencies in zoning contribute to sprawl? This question asks us to compare the actual spatial structure of metropolitan areas with the spatial structure they would have if they operated under some hypothetical efficient system that also addressed the real externality problems zoning is concerned with—say, fungible zoning with honest, responsive, nonthieving governments. It does not ask for a comparison of actual metropolitan areas with metropolitan areas as they would look without zoning; that's not the right question.

It's not the right question because we're not sure whether the complete abolition of zoning is a potential Pareto improvement. Remember that we want to reduce sprawl only to the extent doing so can potentially make people better off; reducing sprawl per se has no intrinsic value (if that's all you

wanted to do, the solution would be easy: outlaw cars and any residence with fewer than 500 people per hectare). Going from current zoning practice to a hypothetical efficient system is by construction a potential Pareto improvement, and the question we want to consider is whether this potential Pareto improvement would carry with it a reduction in sprawl. If it did, then this would be a reduction in sprawl that a reasonably strong argument would support.

Efficient, fungible zoning, we have seen, would probably be considerably less restrictive than actual zoning in partially developed suburbs, and maybe a little less restrictive in distant, largely undeveloped suburbs. In already developed older residential areas, zoning probably doesn't matter much, because little additional development is possible (although some economists have called for looser zoning in luxury areas like Chicago's northern lakeshore and Manhattan's Upper East Side).

In central business districts, I don't know whether current zoning is stricter than efficient zoning would be. The arguments advanced for the 1916 New York zoning ordinance and the Chicago height limitations lead to the conclusion that it is overly strict. At the same time, the degree of fungibility that has developed in these areas since 1916 is high, and political forces in central business districts are different from those in suburbs—construction workers, for instance, and utilities are reliable supporters of development in city centers. If CBD zoning were too strict, entire metropolitan areas would be losing the benefits of increasing returns to scale in production, but there is no evidence or strong argument that it is. For the remainder of this chapter, I'll assume that it's not too far from efficient.

The main effect, then, of efficient zoning would be fewer restrictions in partially developed suburbs. The consequences would depend on whether we are thinking about a closed city or an open city.

In a closed city, population by definition stays the same. If partially developed suburbs become more fully developed and more densely populated, population has to fall somewhere else—in the older, fully developed areas close to the CBD or in the largely undeveloped suburbs. How the population fall would be divided between these two areas depends on how good a substitute capital is for land, in terms of both production by home builders and consumption by residents.

If capital is a good substitute for land—if you can get the same kind of privacy and quiet in an apartment as in a house on a two-acre lot just by building thick-enough apartment walls and private-enough elevators; if you can get a similar gardening experience with a fancy greenhouse in your bedroom—then most of the population loss will come from the older areas. To see why, consider the two extremes. If capital and land can't substitute for each other at all—if every house has to have the same ratio of structure to land—then when there are more of these identical packages of land and structure in

the inner suburbs, the population of the outer suburbs has to fall. Vacant land in the inner suburbs has to be replaced by vacant land in the outer suburbs. This seems unlikely, however: capital is probably not a great substitute for land for these purposes (the marginal product of capital on a given piece of land diminishes rapidly). So most of the population loss would come from more distant, less developed areas. In a closed-city model, then, reducing sprawl by changing zoning would be a potential Pareto improvement. The edge of development would move in, and people would commute less.

This result does not hold in the open-city model—the better model for the long run. Relaxing zoning restrictions in partially developed suburbs increases population there but does not decrease population anywhere else. Holding the world's level of utility constant, whether or not a particular outlying place is inhabited depends only on its own amenities and the cost of commuting, not on any restrictions that might be in place elsewhere in the metropolitan area. The new population in formerly restrictively zoned suburbs is a net addition to metropolitan population. (Indeed, if the additional population drives up wages through the realization of increasing returns to scale, metropolitan-area population might grow even more and force the metropolitan boundary out beyond where it was with restrictive zoning.) Thus in the open city, increasing density is still a potential Pareto improvement, but decreasing the size of the metropolitan area is not. Restrictive zoning (like underuse of land taxes) makes metropolitan areas too sparse, not too big.

This conclusion, strictly speaking, is about the elimination of overly restrictive zoning in a single metropolitan area while nothing else changes in the rest of the world. Alternatively, suppose that overly restrictive zoning were eliminated everywhere simultaneously. Population cannot rise everywhere simultaneously, and so world utility would have to rise. In every metropolitan area, rents everywhere would have to fall, and so some distant areas would no longer be populated. Even in the long run, then, the boundaries of all metropolitan areas would move in if all metropolitan areas eliminated inefficient zoning at the same time.

What can we conclude? Whether zoning causes sprawl depends on what you mean by sprawl and what alternative to current zoning you are considering. Many plausible combinations—but not all—imply an affirmative answer. Eliminating excessively strict zoning makes a metropolitan area denser but not necessarily smaller.

F. Poor People

Whether zoning helps poor people or hurts them, like the question of whether zoning causes sprawl, is a question that needs to be phrased carefully. To be consistent, we'll begin by considering the same alternative we considered with sprawl—efficient, fungible zoning—to see the effect on poor people. For this

sort of question, only the closed-city model is worth considering, since in the open-city model, by construction, nothing can make poor people either better off or worse off. As we saw in the previous section, however, the closed-city model can give us a handy approximation of what would happen in a world with many open cities, all of which simultaneously changed their zoning policies.

Consider a closed monocentric city, then, in which some people are poor and some are rich. Every poor person uses the same amount of land as every other poor person, and every rich person uses the same amount of land as every other rich person. The poor-person land amount—the "poor lot"—is smaller than the rich-person land amount—the "rich lot." To be more specific, these are the amounts of land poor people and rich people would consume under optimal zoning, zoning that optimally controlled externalities. To keep things simple, assume no substitution between land and capital; people just curl up under a tree on the lot that they own. Everybody works in the CBD, and poor people have lower commuting costs per mile than rich people do.

We assume that the ratio of poor lot size to rich lot size is less than the ratio of poor commuting cost per mile to rich commuting cost per mile; poor and rich differ more in how much land they use than in how much it costs them to commute. So in the absence of excessive zoning, poor people would live close to the CBD and rich people would live far away; Figure 8.8 illustrates. Call M^* the boundary between rich and poor without excessive zoning.

It's fairly easy to see that removing excessive zoning under these circumstances will almost always make poor people better off, although whether rich

Figure 8.8 Rent gradient and location choice without excessive zoning.
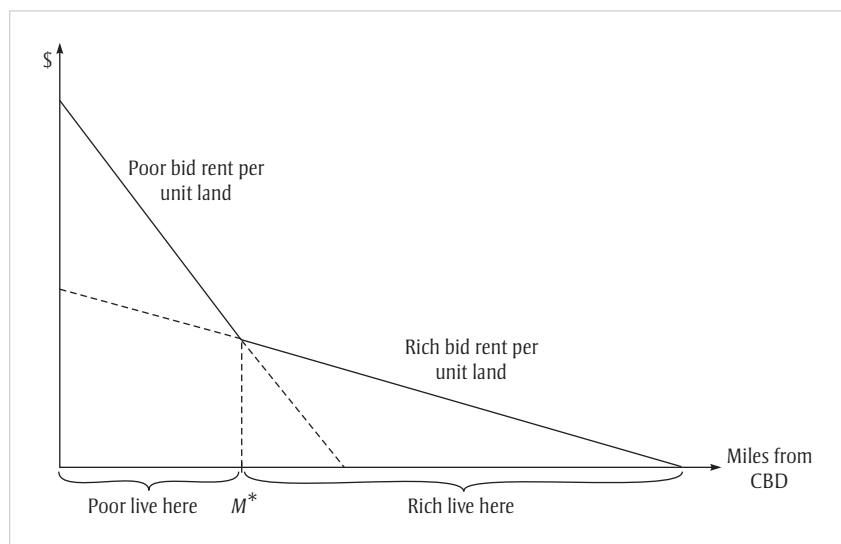
Figure 8.9 The effect
on rent and location
choice of removing
excessive zoning.
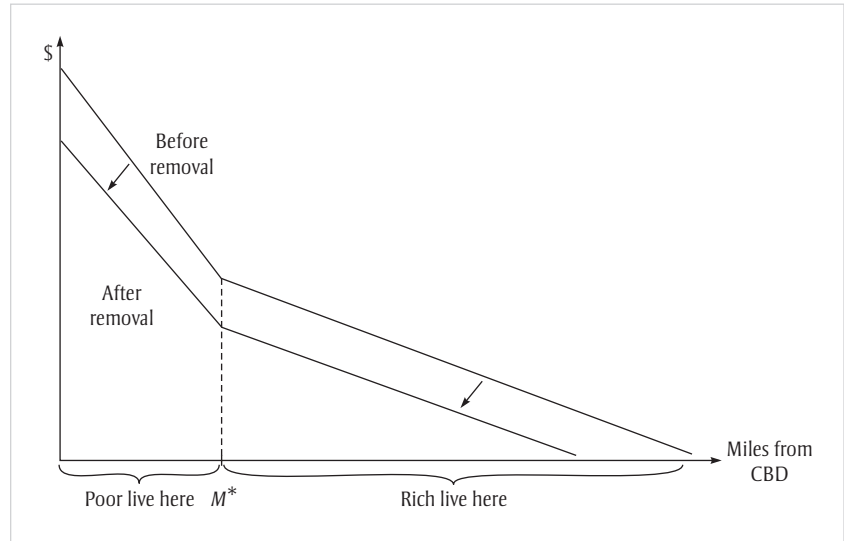
people also gain depends on exactly how the excessive zoning has distorted lo-
cation choices.

For instance, if zoning makes rich people use bigger lots than they other-
wise would, removing excessive zoning helps both rich and poor. Smaller lots
mean that the most distant rich person, who pays zero rent, commutes less
and so is better off. Since all rich people must be equally well off, this means
that every rich person is better off. The least distant rich person still lives at
M^*, and since she has the same commute as before, the rent she pays must be
lower (because she is better off). Poor and rich at M^* pay the same, and so the
poor person at M^* pays less rent, too. So all poor people pay less rent and thus
are better off. Removing excessive zoning helps all housing consumers, rich
and poor alike. Figure 8.9 illustrates.

Alternatively, removing excessive zoning could make poor people better
off and rich people worse off if the original zoning forced some poor people
to live beyond rich people. For instance, if zoning required rich lots inside M^*,
some poor people would have to live further out than all rich people, using
lots of either the poor size or the rich size. Removing excessive zoning would
let these poor people move inside M^*, and so they and all other poor people
would be better off. The farthest rich person has to move farther out, but her
rent decreases. If poor people are using lots of the same size before and after
the zoning change, the rent the most distant rich person pays falls by the
amount of the per-lot *poor* commuting cost for each poor person who moves
in (because this is what the most distant rich person had to pay to outbid poor
people when zoning was excessive). But the increase in commuting cost for
the most distant rich person is the amount of the per-lot *rich* commuting cost

Figure 8.10 Removing excessive zoning when it forces poor people to live far away. With excessive zoning (before removal), rich people live between *A* and *B*; poor people live between *O* and *A* and between *B* and *C*. Without excessive zoning (after removal), poor people live between *O* and M^*, rich people between M^* and *C*.
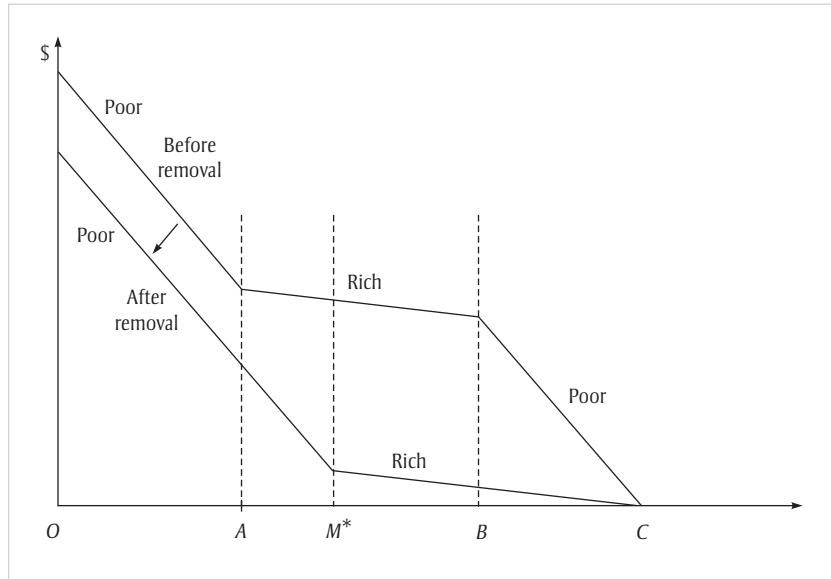
for each poor person who moves in. Thus the increase in commuting cost is greater than the decrease in rent for the most distant rich person. She and every other rich person are worse off. But since removing excessive zoning is a potential Pareto improvement, the gains to poor people are bigger than the losses to the rich. Figure 8.10 illustrates.

In summary, poor people would probably gain from the removal of excessive zoning, at least in the short run, to the extent that their residences are more centralized than those of rich people. What happens to rich people depends on details of the effects of zoning that are not well known.

G. Alternatives

So zoning can be improved, and as a result sprawl might be reduced and poor people might be helped. But how to do it? The simple fungibility change in the preceding examples is not really practical; it's not legal, and the altruistic political behavior of the governments involved in the buying and selling negotiations described is a little beyond realistic. The examples are not really concerned with any definite policy; they are more concerned about showing possibilities. Nevertheless, changing how zoning operates has been a frequent topic of debate since the 1970s, and many economists, lawyers, and planners have put forth suggestions. This section explores some of those policy options.

1. No Zoning

Before 1916 there was no zoning anywhere in the United States, and some areas still do not have zoning. Houston, for instance, the fourth-largest munici-

pality in the country, does not have zoning. Are places without zoning chaotic hellholes, with steel plants operating next to hospitals, bordellos next to high schools, and run-down saloons in the middle of posh neighborhoods? Hardly. Siegan (1972) looked at Houston and found no startling obvious differences with other cities. Businesses had self-segregated, and residences, while cheaper than elsewhere, were usually not bombarded by externalities.

Two legal instruments help people cope with low-level externalities in places without zoning: nuisance law and covenants. Under nuisance law, if somebody is doing something that harms or bothers you, you can go to court and get an order for them to stop. Covenants are agreements that are written into deeds and apply to anyone who owns the property covered by the deed. Through covenants, neighbors (or the original developer of a large parcel) can agree, for instance, to restrict landscaping, architecture, conditions of usage, or practically anything else that zoning laws typically deal with (covenants restricting the race of occupants were declared unenforceable by the U.S. Supreme Court in its 1948 decision *Shelly v. Kraemer*).

The major advantage of replacing zoning codes with nuisance law and covenants is that they both make it more difficult for people to place excessively stringent restrictions on their neighbors. With nuisance law, the check is a judge's interpretation of whether an activity is causing harm. You can't get a court to stop a proposed development because you think its residents will be tacky or because your kids like playing in the woods, or to prevent a rival store from expanding because it will be bad for your business. But you can do any of these things fairly easily under zoning if you can invoke "public health, safety, morals or general welfare" and you are in the majority. Nuisance law is also open to out-of-court cash settlements: if your neighbor's junkyard is truly an annoyance to you but his profit is much greater than the damage you suffer, he can pay you to settle the case, and no questions about impropriety will be raised.

Covenants prevent restrictions from being excessive because they are voluntary. If I want you to give up your right to build a gas station on your property, I must give you something in return that is at least as valuable—maybe I give up my right to build a gas station, too; maybe I give you cash. A covenant usually can't be created unless all owners who give up anything gain from it (or unless the original joint owner of all the properties gains).

Nuisance law and covenants may not be strict enough, however; they may permit activities that should be ended and that zoning codes would not permit. With nuisance law, one obvious failure might involve activities that in fact annoy greatly the people who must suffer them but do not bother judges. A town inhabited primarily by ethical vegetarians might find a butcher shop repulsive, but carnivore judges would not; a town composed mainly of families with small children might, in fact, find senior citizens' housing to be a great burden (they're lousy drivers, they complain too much about noisy kids). The high cost of litigation is also a problem: no one will go to court to

stop an activity if doing so is more costly than enduring it. Nuisance law also can't be applied prospectively; to argue that something is a nuisance, you have to be able to point to it and show what it's doing. You can't say, "If this happens, it's going to be a nuisance"; you can't head off a nuisance before it happens or prevent misguided and, therefore, wasteful construction.

Covenants share with nuisance law the problem of litigation cost. The only way to enforce a covenant is to go to court, and going to court is expensive. There may be a large numbers problem with enforcement, too: if I break my covenant by setting up an all-night rifle range on your block of single-family homes, everyone else benefits when you litigate to enforce my covenant, but you have no way of getting all the others to defray your expenses. Covenants also lack flexibility over time, and once again large numbers are the problem. Changing a covenant (so that it forbids satellite dishes, for instance, or permits working at home over the Internet) requires unanimous consent; it could take a lot of time and negotiating effort, or forever, to put through the change. Both problems could be mitigated, of course, if the parties to the covenant established a covenant-administering agency, financed by annual mandatory assessments, to enforce the covenant and perhaps to modify it with less than unanimous consent to meet changing times. But the further you go in this direction, the closer the covenant-administering agency gets to being a local government with zoning powers.

So while zoning has a tendency to be too strong, nuisance law and covenants have a tendency to be too weak. Which failing is likely to be more serious is an open question, with different answers in different places at different times.

2. More Zoning

The litany of burdens cited earlier as reasons why existing owners might oppose new development in partially developed suburbs should have sounded familiar. Except for the bigots and snobs sentence, these are the very reasons most ordinary people give for opposing sprawl. Thus it is easy to imagine that political agitation against sprawl could easily lead to a tightening of zoning rules in less developed suburbs—making them as restrictive as those in more developed suburbs. As we have seen, the economic consequences of such a policy would be more sprawl and more inequality.

An incident in Edison, New Jersey—a large, partially developed suburb—shows how the rhetoric of sprawl has the potential to be used to exacerbate sprawl. In 1999, a rumor spread through town that a twenty-six-acre sports club on a main street had been sold to a developer who was going to put up condos. Politicians and community leaders immediately attacked the plan as "sprawl"—even though increasing densities in suburbs like Edison is really the opposite of sprawl. The rumor turned out to be false, too (Margolin 1999).

Other zoning-based responses to sprawl are more sophisticated. One of

the most interesting is the idea of urban growth boundaries. A UGB is a line beyond which suburban or urban development is not permitted. In 1979 Oregon established a UGB for Portland, the first in the United States (technically, all Oregon metropolitan areas have UGBs, but Portland accounts for more than 90 percent of the state's metropolitan population). Other UGBs have been set up since then, particularly in California and Florida.

Theoretically, in a metropolitan area without any other form of excessive zoning, the closed-city response to a binding UGB is higher land prices. More people are bidding for the same amount of land, and so prices go up everywhere inside the UGB. Landowners inside the UGB gain; landowners outside the UGB and consumers inside it lose; and losers lose more than winners gain. In the long-run, open-city model, the population of Portland falls until prices are the same as they would be anywhere else in the world for land with similar amenities and proximity to jobs and shopping. Prices would go up relative to the rest of the world if people liked living near accessible countryside that the UGB provides; they would go down if the UGB kept Portland from realizing the economies of scale that larger population would provide. Since landowners outside the UGB lose in either case, people must really like being near the countryside a lot for a UGB to be a potential Pareto improvement.

This thumbnail theoretical argument against UGBs has a serious weakness, however: it assumes that all other prices match social cost perfectly. We have seen throughout this book—in discussions of transportation, for instance—that this condition does not hold, and we will see more instances later. The divergence of private from social cost is in fact the argument for UGBs: development far from Portland creates negative externalities that local zoning would not control because their baleful effects are spread over a much wider area than a single small town. Because politics keeps prices from being set right, people will want to develop where costs are in fact greater than benefits, and a device like a UGB is the only way to keep them from doing so. In this view, UGBs are a potential Pareto improvement.

This counterargument for UGBs faces two difficulties. The first is that it does not go far enough in looking for divergences between private and social cost. Such divergences exist outside Oregon, for example, as well as within it. The people the UGB prevents from living in Portland are going to live somewhere else—Oaxaca or Los Angeles, Mumbai or Seattle—and the social cost of their living there may also be more than the private cost. Even if the private benefit to me of living near Portland is less than the cost I impose on other people in Portland, it's still a potential Pareto improvement for me to move from Los Angeles to Portland if the excess of cost over benefit is less in Portland than in Los Angeles. By focusing only on Portland residents, we forget that people in Los Angeles might be willing to pay me something to get me to leave Los Angeles.

Moreover, even if we ignore the people in Los Angeles who want me to

leave, there is no guarantee that in the long run the political process that determines where the UGB is at any time will do a better job of weighing costs and benefits than whatever distorted market process is politically feasible.

What has the actual experience of the Portland UGB revealed about these theoretical arguments? Practically nothing. The UGB was drawn so far from the edge of development in 1979 that it's likely that it has constrained little development (Knaap 2000).

In 1979 the Portland UGB contained around 230,000 acres, of which 76,000 were vacant and buildable. Over the next two decades, about half of those buildable vacant acres were developed, but even at the end of the twentieth century approximately 38,000 buildable acres remained undeveloped within the UGB—equal to an area somewhat larger than the entire city of Minneapolis. The probability that a farm or forest parcel outside the UGB would convert to urban use did not fall when the UGB was adopted.

What about the future? By law, Portland's metropolitan planning agency must review the UGB every five years and expand it if necessary to make sure that it includes a twenty-year supply of buildable land. In 1998, in fact, the first major expansion was authorized—about 3,500 more acres.

Thus it's not surprising that some economists who have looked for UGB effects on land or housing prices have found practically nothing (see, for example, Phillips and Goodstein 2000; Downs 2002). Without a big cause, you don't have a big effect. But Portland housing prices rose rapidly in the 1990s, much faster than housing prices outside the West, and other economists (Fischel 2002) have argued that Portland looks good in the western context only because the rest of the West is so bad.

Of course, the open-city model might be more relevant here. Portland prices may not have risen much because of the UGB, but the city may be smaller. No one has tested for this possibility yet.

The reforms that Portland adopted in 1979, however, were not totally empty. Along with the UGB came the establishment of a metropolitan-area planning organization that has forced many suburbs inside the UGB to zone less restrictively in the hope of increasing density. But many of the desired density increases have not occurred (Knaap 2000), and whatever the new forms of governance have accomplished, they have not had much effect on even short-run housing and land prices.

3. Smart Growth

In the aggregate, a smart-growth policy is similar to a UGB. Certain parts of a metropolitan area are zoned for dense development—areas around mass transit hubs, for instance—and other areas are zoned for no development at all. Smart growth is like a UGB in that it sets aside areas for dense development and no development; the difference is that these areas are intertwined and mixed together; they're not clearly separated by a single UGB. Smart growth

has been championed, for instance, in the New Jersey state plan—which is more of an exhortation than a plan—and implemented in a few towns.

As an initiative that supersedes local zoning and applies across a metropolitan area, smart growth bears a close enough resemblance to UGBs that a separate analysis is unnecessary.

4. Inclusionary Zoning

Many proposals about zoning have concentrated on how it affects poor people, not just how it affects sprawl. Some of these proposals could cause perverse effects, but others, if carefully executed, could benefit poor people.

The basic premise of most of the attempts to change zoning to help poor people has been that exclusion is not a legitimate goal of zoning. This view has been perhaps most clearly expressed by the New Jersey Supreme Court in a series of decisions that invalidated certain zoning practices. Zoning, the court found, was a police power that the state had delegated to local governments to promote the general welfare—the welfare of all the people in New Jersey. The state did not delegate police power so that local governments could exclude people:

> When the exercise of that [police] power by a municipality affects something as fundamental as housing, the general welfare includes more than the welfare of that municipality and its citizens: it also includes the general welfare—in this case the housing needs—of those residing outside the municipality but within the region that contributes to housing demand within the municipality. Municipal land-use regulations that conflict with the general welfare thus defined abuse the police power and are unconstitutional. (*Southern Burlington NAACP v. Township of Mount Laurel,* 1983, 92 NJ 158, pp. 208–209)

Notice that this principle can be phrased naturally in property-rights language. The right to exclude poor people from living nearby is one of the sticks that may or may not be in the bundle known as owning a house in the suburbs. This stick could be in some other bundle as well; it doesn't matter which bundle it's in. The New Jersey Supreme Court decided that this stick did not belong in the bundle of suburban home ownership—just as the right to bash in your neighbor's skull is not included in the bundle called tire-iron ownership. This is a perfectly sound economic decision.

When you consider it in these terms, the most efficient policy response is obvious. The state would print certificates. Each certificate would allow the holder to prohibit construction of one unit of housing for low-income people for a period of years. A town owning such a certificate would attach it to a particular lot; anybody who wanted to could build housing for low-income people on a lot not covered by a certificate. The state would distribute the certificates originally to poor people—maybe randomly, maybe to the parents

of honor students or to workers in particularly difficult occupations, maybe to people with good political connections. They could sell the certificates to whomever they wished at whatever price they could get, or they could keep them. Towns wanting to exclude poor people would buy certificates; so might developers who wanted to keep them out of the hands of towns. This mechanism wouldn't be flawless, but it would effectively comply with the court's mandate, make poor people better off, and produce a potential Pareto improvement.

But the actual policy response has been much different. The predominant policy taken up is called **inclusionary zoning.** Under inclusionary zoning, developers are allowed build on a site at greater densities than the existing code permits, but they have to set aside a fraction of the units—often 15 or 20 percent—for sale to low-income people at below-market prices. Inclusionary zoning has been used extensively in New Jersey, California, Massachusetts, Oregon, and the suburbs of Washington, D.C., among other places; a version even applies in New York City.

You can place inclusionary zoning into property-rights terms and see its problems if you look at the sort of disputes in which it arises. Typically, a developer wants to build fairly expensive market-rate housing on a piece of land, and the surrounding neighbors and towns don't want him to. In other words, the developer is claiming that he owns the right to build market-rate housing in this particular place, and the town is claiming that it owns this right. The disputes are about ownership of this right.

When inclusionary zoning is used to resolve these disputes, the developer subsidizes some poor people. So it almost looks like inclusionary zoning gives the right to build market-rate housing to those poor people. After all, the owner of a right is the person who has to be paid if someone else wants to exercise that right (if I own a house and you want to live in it, you have to pay me). But this appearance is deceptive. If those poor people really had such rights, they would be free to sell them to the highest bidder.

Instead, inclusionary zoning is a way for the town to claim the right and to make the developer pay a lot to acquire it. The developer pays certain poor people, not the town, but from his point of view this difference is immaterial—whatever increases his costs, changes his behavior. The town could just as well require him to pay for famine relief in Africa, preservation of salmon spawning grounds in British Columbia, or the next scientific expedition to the South Pole. It could tell him to throw the money into the ocean. The greater the amount that the town (or the court) requires the developer to sacrifice in this fashion, the more progressive and generous it appears to be, but the more exclusionary it actually is. Making housing more expensive to build discourages housing.

In summary, then, inclusionary zoning is an exclusionary tool so far as people who don't qualify for low-income housing are concerned. So long as it

is not so exclusionary that no market-rate housing is built, it helps some lucky poor people. What about poor people who are not lucky? That depends on the way in which the housing market for rich and middle-income people is linked to the housing market for poor people. I will address this question in Chapter 14.

IV. Other Codes

Zoning laws, although they have received the greatest amount of attention from economists, are not the only regulations that govern city life. Many other types of codes are far older than zoning—blue laws, for instance, restricting commerce and other activity on Sundays, come from the Old Testament commandment to keep the Sabbath as a day of rest and were first established in colonial times, and the earliest colonial cities in America also had fire codes. Among Peter Stuyvesant's first acts after being appointed governor of New Amsterdam in 1647 was to ban the building of wooden chimneys and to enforce a literal curfew (from the Old French *cuevrefeu,* or to cover the fire): a time at which all fires had to be snuffed out in the evening (Burrows and Wallace 1999, p. 43).

Of these two other codes, blue laws are probably more like zoning in that they take advantage of homogeneity. Having a few people bustling about noisily is much more annoying when everyone else is in church or resting at home than when everyone else is out working. The marginal external cost of nonreligious activity on a Sunday changes a lot as the number of people engaging in such activity changes, and so regulating Sunday activity through rules rather than prices is appropriate. Blue laws are also easy to enforce through community reports.

Health, building, and housing codes resemble zoning in that they address externality problems that are sometimes not well defined. In the late nineteenth century, for instance, housing codes were designed to fight the spread of tuberculosis (TB) by ensuring that all rooms had adequate sunlight. This was enormously expensive and forced major changes in the way apartment buildings were built (courtyards and insets were created by the housing code, the same way setbacks were created by the zoning code). We know now that sunlight has little or nothing to do with the spread of TB, but we still have the window requirements. (Having more space per tenant probably did help reduce TB, and the housing codes also mandated bigger rooms. But rising wages and cheaper transportation probably contributed more to easing the space crunch than housing codes did.)

Codes have sometimes been quite Draconian. Quite recently, Singapore's strict laws on quarantine, isolation, and testing spared it from the severe acute respiratory syndrome (SARS) epidemic (Arnold 2003). In early twentieth-cen-

tury U.S. cities, people with serious communicable diseases were forcibly removed from their residences, and their furniture, clothing, and bedding were burned. And when the Black Death hit Europe in the fourteenth century, the despot of Milan, Archbishop Giovanni Visconti, "ordered that the first three houses in which the plague was discovered were to be walled up with their occupants inside, enclosing the well, the sick, and the dead in a common tomb. Whether or not owing to his promptitude, Milan escaped lightly in the roll of the dead" (Tuchman 1978, p. 108).

One curious aspect of city codes today is that although the severity of the externalities they address depends on the population density of the neighborhood in which they are being enforced, the codes hardly ever prescribe different standards for different neighborhoods. Perhaps this is an indication that existing codes are not the optimal ones a disinterested economist would propose.

Fire, plumbing, flooding, and electrical codes differ from zoning because they are more closely tied to services that municipalities or utility companies provide. They act as substitutes for these services. Requiring that tall buildings have water tanks and sprinkler systems allows a fire department to operate without extremely tall ladders and extremely high water pressure; it also need not be staffed to respond as quickly to fires in these buildings. Requiring that large developments and parking lots build retention basins for water runoff allows a city to achieve the same level of protection against flooding with smaller storm sewers. The electrical code reduces the probability of fire. The plumbing code—mandating indoor toilets, for instance, and specifying how a building's fixtures are linked to the municipal water and sewer systems—inhibits externalities directly and also moderates flows from the water system and to the sewer system; it also affects the content of sewage.

Most of these codes encourage activities that people would want to do anyway—I don't want a short circuit in my walls that starts a fire that burns my house down—but wouldn't necessarily do enough of because they don't bear the full cost when something goes wrong. My installing sloppy wiring hurts my neighbors indirectly because it makes them pay more for the fire department, or it decreases the protection they get because the firefighters are so often responding to my fires.

My faulty wiring hurts my neighbor directly, too. A fire in my house endangers hers, just as a flood from my parking lot can fill her basement. But like the codes more completely targeted on direct externalities, the codes concerned with municipal resources rarely differ in neighborhoods of different density. (An exception are the old regulations called "fire limits": many eighteenth- and nineteenth-century cities forbade wooden construction in their dense central districts. Such regulations are rare today, though.)

Codes like fire, flood, and electrical safety codes have an ancillary benefit involving risks that are usually covered by insurance. The company I buy fire

insurance from wants to know that my house has good wiring but can't take my word for it, or the word of an inspection company I choose and pay. The company could hire its own inspectors, but this would be expensive if its customers are spread out around the country, a few here and a few there. Better if all the inspections in one area could be done by the same inspector, someone familiar with the area's peculiarities, someone working for the local government. Thus, government regulations can facilitate private insurance contracts.

Both of these motivations—externalities and insurance—are strongly at work in the earthquake provisions in construction codes. When an earthquake occurs, emergency services will be overwhelmed. If ambulances, fire trucks, and rubble-moving equipment are responding to my building, they can't be responding to your building, and you may need them desperately. The cost of their responding to me—a cost that could have been avoided if I had built more carefully—is enormous and is borne by the (marginal) injured party who cannot be helped. Insurance companies, moreover, have no desire to specialize in Los Angeles earthquake coverage, and so are happy to delegate monitoring to the experts in Los Angeles city government.

Could these purposes be equally well served by Pigouvian taxes? Most of them probably could not, because this kind of regulation is mainly about how much each building draws on a common pool of capacity; marginal external cost is considerably greater as capacity is being approached than when it is not. But sometimes prices can work in circumstances like this—subways and airline flights are good examples—and so a blanket answer is a little hasty. Enforcement of these codes is often poor, but enforcement of Pigouvian taxes on, say, structural deficiencies might be worse. There may be many examples of particular parts of a code's purpose that could be better handled with prices.

As with zoning codes, these codes are often written wrong; politics produces less than optimal codes. Division of labor and intellectual specialization also work against optimal codes: fire codes, for instance, are heavily influenced by fire chiefs, who worry about reducing fire deaths, not about making houses less expensive. Academic discussions of these codes often allude to a different culprit not often cited in zoning discussions: producers, especially unions, who have some sort of monopoly on the ability to comply with the codes. If only licensed union plumbers can ensure compliance with the plumbing code, then the more elaborate the work the code requires, the greater the demand for their services and the more they can exploit the public. Since the cost of complying with these codes is high, some writers have cited this process as a major source of inefficiency (see, for instance, U.S. President's Commission on Housing 1982).

Not all studies, however, look seriously at the benefits of this kind of regulation, which are hard to measure. The conditions for monopoly exploitation, too, are often a little stronger than seem to be justified: producers must really be capable of keeping other people from learning their craft and from

migrating from other cities for it to be to their advantage to prop up demand through overly restrictive regulations.

Simple exploitation, however, is not the only motive producers might have in trying to influence the content of these codes. Probably a more plausible goal is resisting technological change that would render their skills and equipment obsolete. Even if they are earning just a normal (not a monopoly) return on their investment in the equipment and skills needed to work with cast-iron pipe, contractors and skilled workers could be hurt if plastic were to replace cast iron. Thus they have a strong incentive to make the plumbing code specify cast iron, even if plastic is better and cheaper; and they may prevail, since they may be more knowledgeable and better organized than the people who would gain from plastic. So codes may slow down the rate at which technological innovations are adopted, particularly in long-established towns where producers are well entrenched; and since the payoff to research comes only when its results are adopted, slow adoption discourages research in areas like plumbing and housing construction. Consumers probably pay more as a result.

But producers aren't the only parties who benefit from codes that are out of date and too restrictive—and they would be much more unlikely to succeed politically if they were. Owners of existing buildings gain, too. This works in two ways.

First, slowing innovation helps owners of existing buildings because it makes new buildings more expensive than they would otherwise be. If using plastic instead of cast iron for pipes makes it cheaper to build a house just as good as the one you already own, buyers aren't going to be willing to pay you as much for yours, because it's easier for them to get the same quality by building from scratch and not dealing with you. Owners of existing properties are better off if codes force new construction to be more expensive.

This gain from making new construction more expensive also explains why existing owners often support measures that make these codes tighter. Changes in code are usually grandfathered, which means they will apply to buildings built in the future, but don't apply to buildings that have already been built. If your assets become more difficult to reproduce, they become more valuable. It's nice to have the codes lax when you build but strict when people try to follow you.

The story for these codes, then, is similar to the story about zoning: both are often used to "pull in the welcome mat." Early arrivals use them to make it tougher for later arrivals to follow. Indeed, the most famous legal cases on the safety codes are just like the more recent legal cases about zoning—they are all about the exclusion of latecomers. These cases are known as the "laundry cases." In the early 1880s, several towns in the San Francisco area adopted regulations about night work and the use of wood-frame structures for laundries, ostensibly as fire-prevention measures. Violators, all of whom were Chinese,

were imprisoned, and they sued to overturn the regulations. Not coincidently, the same towns had also maintained ordinances prohibiting Chinese residency explicitly until the U.S. Supreme Court had struck those ordinances down. This time, though, the Supreme Court upheld the laundry regulations as a legitimate exercise of police power (Toll 1969, p. 27).

V. Sidewalks and Public Spaces

The hustle and bustle of sidewalks, plazas, parks, and transportation hubs contributes a lot to the vitality of traditional urban life. People like the "energy" of New York's Greenwich Village and Fifth Avenue, Mexico City's Zocalo, London's Piccadilly Circus; it attracts people and stimulates them. Public spaces are where people meet, deliberately or inadvertently, and also where they try to hide. Sidewalks are transportation corridors, too: the face-to-face interactions that lively central business districts are built around occur only when at least one party has traversed a sidewalk.

These spaces serve grander goals, too. Ellickson (1996, p. 117) writes: "To socialize its members, any society, and especially one as diverse as the United States, requires venues where people of all backgrounds can rub elbows . . . A liberal society that aspires to ensure equality of opportunity and universal political participation must presumptively entitle every individual, even the humblest, to enter all transportation corridors and open-access public spaces."

Supreme Court Justice Owen J. Roberts was even more emphatic about how public spaces were crucial for political liberty in the 1939 case *Hague v. CIO* (307 US 496, 515): "Wherever the title of streets and parks may rest, they have immemorially been held in trust for the use of the public and, time out of mind, have been used for purposes of assembly, communicating thoughts between citizens, and discussing public questions. Such use of the streets and public places has, from ancient times, been a part of the privileges, immunities, rights, and liberties of citizens" (cited in Ellickson 1996, pp. 1219–1220). The term *forum,* after all, comes from the Latin name for the public square or marketplace in an ancient Roman city; it's not just a metaphor.

Preserving and protecting these public spaces is therefore an important exercise of police power. Many activities in public spaces create positive externalities, but other activities are harmful. Street musicians, even those who aren't particularly skillful, entertain most passersby, but rude people and those with communicable diseases are annoyances, or worse. Some public spaces get congested—subway staircases right after a rush-hour train has arrived and many midtown Manhattan sidewalks at lunchtime, for instance. When this occurs, even the healthiest and most polite citizens are harming their fellow pedestrians just with their presence.

How well public spaces are managed also affects how activity is distrib-

uted within a metropolitan area. The kind of vital pedestrian life I am concerned with here occurs most often in the older, denser parts of cities, not in the newer, less dense, more automobile-oriented regions. Strolling home after dinner and watching the crowds come and go is something you do on the Loop in Chicago, not along an interstate in San Bernardino County, California. If public spaces are not managed well, downtowns will suffer: "Shoppers can switch to enclosed malls, employers can move to suburban industrial parks, and universities can shift activities to satellite branches" (Ellickson 1996, p. 1246). A metropolitan area where public spaces are not managed as well as they should be will sprawl more than it should.

(This argument holds for all uses of public space. A person prohibited from sleeping off his Saturday-night hangover on a bench in a public park may decide to give up on central-city life because it's no longer fun and move to the far suburbs or the country. Regulation can be too strict as well as too lax.)

No feasible technology exists for charging finely graded tolls for the use of sidewalks, parks, and plazas, and so pricing mechanisms are out of the question for now. (See Ellickson 1993 for a full discussion of this point.) Even if technology allowed it, pricing would probably still be inferior to direct regulation for many public-space issues. That's because homogeneity is often beneficial, and so the marginal external cost of some activities varies greatly. One peddler, his back to the wall on a busy sidewalk, slows foot traffic by creating a bottleneck and so does considerable harm; a second peddler next to the first, his back also to the wall, makes almost no difference, since he does not constrict the existing bottleneck. Yelling and screaming annoys no one when a parade to honor heroes of some sort is going on; it annoys almost everyone during a funeral procession. People who go to Hyde Park in London are not distressed to hear spirited speeches on politics and religion; those who flock to Disney World or to a concert in Central Park would be.

Regulations on public conduct, therefore, abound. (I do not count criminal laws that prohibit activities like murder and assault everywhere, even though public spaces can't operate at all if people feel disproportionately unsafe there; concerns about crime will be addressed in Chapter 16.) You must pick up after your dog; you can't litter, spit, or expose yourself. Your opportunities to solicit business for yourself are circumscribed in many different ways, and so are your opportunities to drink alcohol and smoke cigarettes. You can't ride bicycles on sidewalks in New York City, or on hiking paths in many parks. You have to keep the sidewalk in front of your house in good repair and shovel the snow off of it within twenty-four hours of a snowfall. The amount of noise you can make is also regulated.

As with the other police powers discussed in this chapter, benevolence and wisdom have not always been the hallmarks of local-government regulation of public spaces. Regulation of peddlers, for instance, is often heavily in-

fluenced by indoor (regular) merchants, who want as little competition as possible, and by licensed peddlers, who also want to keep the number of licenses as small as possible (so long as they are included). Once again, those who have arrived first get busy using regulations to pull in the welcome mat behind them.

(Does it matter that indoor merchants and current peddlers lose out when more peddlers enter the business? Yes, but not nearly so much as they claim, since their losses are really transfers to consumers. Everything else being equal, introducing more peddlers when their number is constrained drives down the prices that consumers pay; this lets consumers buy more and transfers rent from existing suppliers to consumers. More peddlers is a potential Pareto improvement. Consumers could reimburse the indoor merchants and existing peddlers for their losses and still be better off than they were without the added peddlers. Does it matter that indoor merchants have overhead costs? No, these are sunk costs.)

Public-space regulation, however, is more closely intertwined with the core political rights of expression and assembly than the other types of regulation we have been studying. The right to hold a rally to oppose the mayor is more fundamental to a democratic society than the right to use plastic plumbing (even though many people would pay more for the latter than the former). As a result, abuses of these police powers have been more spectacular than abuses of the other police powers, and judicial scrutiny has been considerably stricter. Laws that restrict political speech, for instance, must meet a higher standard than the substantive due-process requirement of a "reasonable relation" to some public policy goal; instead, the government has to show a "compelling interest" in the goal sought.

Before 1960 in U.S. cities, laws against loitering, vagrancy, and public drunkenness were the mainstays in local-government regulation of public spaces. These laws allowed the police officers who enforced them considerable discretion in deciding which activities to prohibit. In property-rights terms, the public authorities owned most of the sticks associated with street use. You could do something on the sidewalk only if the police let you. (The major exception was the right to assemble for political purposes; *Hague v. CIO* in 1939 had established that a city administration could not prohibit a demonstration.)

After 1960, a series of court decisions (both state and federal) and state legislative enactments shifted many sticks out of the control of public authorities and into the hands of citizens. The key Supreme Court decision was *Papachristou v. City of Jacksonville* in 1972. The Jacksonville police had arrested Papachristou, a white woman, for violating the city's vagrancy ordinance—she had been riding in a car with several black men, and had been unable to give a good account of what she was doing. The Court struck down Jacksonville's ordinance because it violated procedural due process; it was too

vague to apprise citizens of what was forbidden. This decision, and others like it on the state level, affected loitering ordinances as well as vagrancy ordinances. Laws against public drunkenness disappeared about the same time, as state legislatures (and police officers) came more and more to think of alcoholism as a medical condition rather than a criminal offense.

The presumption, thus, became that you could do something in a public place unless it was expressly forbidden, and it could be expressly forbidden only if the prohibition bore a reasonable relation to some public policy objective. Most cities, for instance, continue to have laws against obstructing sidewalks, because enabling people to walk from place to place quickly is a legitimate public policy goal. Legal controversy currently surrounds measures like curfews for youths and anti-panhandling ordinances because sometimes they are drafted in ways that interfere with First Amendment rights (restricting the ability of young people, for instance, to assemble peacefully for a political meeting) or leave too much discretion in the hands of public authorities (letting the police decide, for instance, which nighttime activities will be permitted and which will not). Certain varieties of these ordinances are clearly constitutional—prohibitions on panhandling near ATMs, for instance.

What's legal, however, is different from what's wise. Most of the recent debate about public-space regulations, unfortunately, has focused more on their legality than on their wisdom. Two lessons we have learned from other types of regulation seem to have been overlooked.

The first lesson is that segregation and homogeneity are often useful. Indeed, the advantages of homogeneity were my chief substantive argument for regulating public spaces rather than levying tolls. But most cities now draft public-space regulations that are the same everywhere. Ellickson (1996) has argued that this practice makes about as much sense as trying to enforce the same speed limit on interstate highways as on residential cul-de-sacs. He proposes that public-space regulation imitate zoning—have some areas where virtually anything goes (loud music, lots of peddlers, and public drinking, for instance) and some areas where strict decorum is enforced (no panhandling, no smoking).

One objection to such a scheme is that no property owners would ever agree to have their neighborhood zoned for the least restrictive public use. But if such zoning were sufficiently rare, the areas that had it would be crowded, and crowds mean money. In some ways Ellickson's proposal would re-create the old skid rows, where police informally ignored the vagrancy, loitering, and public drunkenness ordinances, and many business owners were glad that they did.

A more serious objection is that zoning of public spaces, like the existing zoning of private structures, is likely to end up being too restrictive. But it is in fact likely to be less restrictive than the one-size-fits-all regulation we now have.

The second lesson that seems to have been overlooked is that while

wholesale replacement of direct regulation is neither possible nor desirable, prices and fungibility can play a role. Instituting freely transferable peddlers' licenses, different fees for peddling in different places at different times, and licenses for minor nuisances like boom boxes, for instance, could probably create potential Pareto improvements. (In some cities, newsstands may be established on sidewalks only with permission of the adjacent property owner, and the newsstand owner can pay the property owner for this permission. But since the property owner does not bear the cost of the sidewalk obstruction the newsstand creates—most people passing by the building are headed elsewhere—there is no reason to believe that the right price is charged.)

VI. Conclusion

Sometimes regulations are better than prices for keeping people in cities from getting in each others' way too much. This is especially true when segregation—zoning—reduces the harm that an activity causes, as when you put all smokers in the same section of a restaurant, for instance, or a bunch of noisy factories in the same neighborhood. The regulatory systems we examined play an important role in making cities livable, a role that a price system probably could not play.

At the same time, direct regulation is a police power and police powers are not fungible, so there is good reason to believe that the regulations we have now could be improved considerably. One plausible political interpretation is that regulations tend to help those who are already somewhere, at the expense of those yet to come, and do so inefficiently. In several cases this might contribute to sprawl and make cities less dense than they should be.

Questions

1. John Hayseed, a farmer, owns a thirty-acre dairy farm at the end of Hardscrabble Road in Ong's Hat, New Jersey. If it were to continue as a dairy farm it would be worth $10,000 an acre. The land is currently zoned first residential. Under the Ong's Hat zoning code, the only permitted use for land in first residential zones is single-family homes on quarter-acre lots. Hayseed's farm is a "nonconforming use," which means it can continue as it is, but a new farm couldn't be started up on the property.

Hayseed is thinking of retiring to Florida and has been talking to Alfred Cityslicker, a developer. Cityslicker is planning to buy the farm, subdivide it, and build roads. The roads would take up five acres. On the remaining

twenty-five acres, he would build houses on quarter-acre lots and sell them for $250,000 each. The houses would cost him $200,000 each in construction costs. Roads and infrastructure would cost an additional $700,000 total.

Currently there are forty other households on Hardscrabble Road. They have heard about the proposal and have conducted a series of totally reliable traffic engineering studies. The studies show that each additional house built on Hayseed's farm would generate enough additional traffic to reduce the value of each of the original houses on Hardscrabble Road by $500. Since Hardscrabble Road runs directly to a major interstate, the households on Hardscrabble Road are the only households that would be affected.

a. What is the efficient solution?

b. If zoning doesn't change, what will happen?

The residents of Hardscrabble Road keep seeing Cityslicker visiting the Hayseed Farm, and they get madder and madder. Finally they get together and go to the Ong's Hat Planning Board to get all of Hardscrabble Road rezoned to agricultural use. (Their own houses would continue as nonconforming uses.) Politics in Ong's Hat is entirely honest and aboveboard. The Planning Board and the Zoning Board of Adjustment always respond perfectly to the desires of the majority of Ong's Hat residents. Ong's Hat residents never accept bribes.

c. What will the Planning Board do?

d. Is the result Pareto optimal? Why or why not?

Trying to reach a compromise, the mayor of Ong's Hat suggests that perhaps Hayseed might want to give a few acres to the town for use as a park. Because it's so remote, the park would be used only by the current and future residents of Hardscrabble Road. Fortunately, a detailed, absolutely authoritative study of the effects of a park happens to be available. Every acre that is devoted to parkland will increase the value of each of the new homes that might be constructed on Hayseed's land by $2,000. Every acre of park will also increase the value of every existing house on Hardscrabble Road by $2,000 (this is in addition to any effect from less traffic). But no matter what size the park was, five acres would still have to be devoted to roads and infrastructure.

e. Would Cityslicker develop a park on his own? If so, how big would it be?

f. What is the optimal size of the park?

The mayor says that she thinks fifteen acres is a good size for a park. If Hayseed agrees to donate fifteen acres to the town for a park, she thinks the Planning Board should support the rest of the development.

g. Would the residents of Hardscrabble Road prefer a development with a fifteen-acre park to no development at all? Show why or why not.

h. Would Cityslicker and Hayseed prefer a development with a fifteen-acre park to no development at all? Show why or why not.

2. Noisy City is thinking about regulating cell phones. People have complained about the nuisance and danger caused by other people talking on cell phones. You have been asked to prepare a recommendation that minimizes the expected value of social loss.

The private marginal cost of cell phone calls is constant at one. The social (external) marginal cost of a cell phone call is q, where q is the number of calls being made.

Because it's a relatively new technology, you are unsure of the demand (i.e., marginal benefit) of cell phone calls. You know that the (inverse) demand curve is

$$D(q) = 1 + A - q/10,$$

but you don't know what the value of A is. There is a 50 percent probability that $A = 14$, and a 50 percent probability that $A = 8$.

a. Find the optimal number of phone calls as a function of A. (That is, if you knew A, what would be the optimal number of phone calls?)

b. In general, if you ended up with q phone calls being made instead of the optimal number, q being greater than the optimal number, what would the social loss be as a function of A and Q?

c. Suppose you regulate the quantity of phone calls by fixing a number of permitted calls R. If $A = 8$, what's the deadweight loss (as a function of R)? If $A = 14$, what's the social loss (as a function of R)? What's the expected social loss (as a function of R)? What value of R minimizes the expected value of social loss? What rule should you make? If you make the best possible rule, what is the expected value of social loss?

d. Suppose, instead, that you set a tax on cell phone calls. Set the tax rate at t. Given A and t, how many calls will be made? If $A = 8$, what's the social loss (as a function of t)? If $A = 14$, what's the social loss (as a function of t)? What's the expected value of social loss (as a function of t)? What tax rate should you set? If you set the best possible tax rate, what's the expected value of social loss?

e. Should you use a quantity rule or taxes to regulate cell phones?

f. How would your answer change if the (inverse) demand curve were

$$D(q) = 1 + A - 10q,$$

instead?

3. Consider the Milanese way of dealing with the Black Death. Did it save lives on net? Given a choice between this policy and doing nothing, which would have been better? Are there other policies Milan could have used? If you were advising Archbishop Visconti when he was considering his policy, what would you say?

4. Mountain bikes damage most eastern hiking trails, especially after heavy rains. How should they be regulated?

References

Altshuler, Alan A., and Jose A. Gómez-Ibáñez. 1993. *Regulation for Revenue: The Political Economy of Land-Use Exactions.* Washington, DC: The Brookings Institution.

Arnold, Wayne. 2003. "In Singapore, 1970s Law Becomes Weapon against SARS." *New York Times,* June 8, pp. F5, F8.

Burrows, Edwin G., and Mike Wallace. 1999. *Gotham: A History of New York City to 1898.* New York: Oxford University Press.

Downs, Anthony. 2002. "Have Housing Prices Risen Faster in Portland Than Elsewhere?" *Housing Policy Debate* 13(1):7–31.

Ellickson, Robert C. 1993. "Property in Land." *Yale Law Journal* 102:1381–1387.

———. 1996. "Controlling Chronic Misconduct in City Spaces: Of Panhandlers, Skid Rows, and Public-Space Zoning." *Yale Law Journal* 105:1165–1248.

Ellickson, Robert C., and A. Dan Tarlock. 1981. *Land Use Controls: Cases and Materials.* Boston: Little, Brown and Company.

Eskeland, Gunnar S., and Tarhan Feyzioglu. 1995. "Rationing Can Backfire: The 'Day without Car' Program in Mexico City." World Bank Policy Research Working Paper 1554.

Fischel, William A. 1985. *The Economics of Zoning Laws.* Baltimore: Johns Hopkins University Press.

———. 2002. "Comment on Anthony Downs's 'Have Housing Prices Risen Faster in Portland Than Elsewhere?'" *Housing Policy Debate* 13(1):43–50.

Freund, Ernest. 1904. *The Police Power: Public Policy and Constitutional Rights.* Chicago: Callaghan and Company.

Glaeser, Edward, and Andrei Shleifer. 2001. "A Case for Quantity Regulation." National Bureau of Economic Research Working Paper 8184.

Kmiec, Douglas W. 1981. "Deregulating Land Use: An Alternative Free Enterprise Development System." *University of Pennsylvania Law Review* 130(1):28–130.

Knaap, Gerrit J. 2000. "The Urban Growth Boundary in Metropolitan Portland, Oregon: Research, Rhetoric, and Reality." *Public Investment,* December. Accessed on September 6, 2002, at *www.metro-region.org/growth/ugbursa/apa_article,html.*

Margolin, Josh. 1999. "False Rumor of Sports Club Sale Fuels Panic about Overbuilding." *Newark Star Ledger,* August 25, p. 27.

Phillips, Justin, and Eban Goodstein. 2000. "Growth Management and Housing Prices: The Case of Portland, Oregon." *Contemporary Economic Policy* 18(3):334–344.

Siegan, Bernard. 1972. *Land Use without Zoning.* Lexington, MA: Heath-Lexington.

Toll, Seymour I. 1969. *Zoned American.* New York: Grossman.

Tuchman, Barbara W. 1978. *A Distant Mirror: The Calamitous Fourteenth Century.* New York: Ballantine.

U.S. President's Commission on Housing. 1982. *Report.* Washington, DC: Government Printing Office.

Willis, Carol. 1995. *Form Follows Finance: Skyscrapers and Skylines in New York and Chicago.* New York: Princeton Architectural Press.

Chapter 9

Water, Sewers, Fire, and Garbage

This chapter is about the most well-known public health services that city governments generally provide (or arrange to have provided): delivering clean water, removing sewage, suppressing fires, and picking up solid waste. Together, they go a long way toward making a city a safe and vibrant place.

These services have more in common than this heritage, however. All of them involve significant capital investments: reservoirs and water pipes, sewer pipes and sewage treatment plants; fire trucks, hydrants, wide streets, and a trained and disciplined firefighting force; garbage trucks and some system of solid waste disposal. Partly as a result, **marginal cost** for each service is not closely tied to **average cost,** and unregulated market provision can go seriously awry—although, of course, government provision can also go seriously awry. (Marginal cost is how much producing one more unit of something costs; average cost is how much producing the average unit costs.) Each of these services involves external costs and benefits, but the policies for providing them are not based on Pigouvian taxes or Coasian property rights or rules. They exemplify a fourth approach to externalities: provision of free or cheap goods.

In this chapter we will look at these four services one after another. We'll examine why getting them provided right is hard, and why it's important. We'll look at how the public should be charged for each service, and what goes into calculating the right level of capacity. We will then examine a number of controversies that have touched these services, including the question of whether the way they are provided contributes to excessive sprawl.

To get the most from this chapter, you should be familiar with these concepts: average cost, elasticity, externalities, fixed cost, lock-in, marginal cost, moral hazard, natural monopoly, overhead, regressive, second-best, sunk costs, and variable cost. You can find these terms in the Glossary. Concepts from previous chapters include demand smoothing and ex post opportunism.

I. Water

A. Why It Matters

In 1975, when the prospect of bankruptcy forced the New York City government to develop a doomsday plan for closing down, the last city employees scheduled to be laid off were the water system operators. Human beings are about 70 percent water by weight; we lose water every day through normal activities, and so we can survive only a few days without replenishing that water. You can't have a city without some source of pure water.

The word *pure* is important because water can transmit a variety of diseases. Approximately 60 percent of the illness in the world is water-related (*Economist* 2003), and about 2.2 million people a year die from water-related illnesses (Water Supply and Sanitation Collaborative Council 2004). In waterborne diseases like cholera, typhoid, dysentery, and some diarrheas, the pathogen is in the water a person drinks. Other infections are known as water-washed: people get them from washing themselves or the food they eat with pathogen-infested or insufficient water. And diseases like malaria, yellow fever, dengue, and river blindness are spread by insects that breed in water or bite near water (Sah 1997, table 3.1). The World Bank (1993) emphasizes the multiplicity of ways that water affects health: "Water quantity is as important as water quality. Washing hands after defecation and preparing food is of particular importance in reducing disease transmission, but without abundant water in or near the home, hygiene becomes difficult or impossible. The lack of water supply and sanitation is the primary reason why diseases transmitted via feces are so common in developing countries."

Keeping people alive, healthy, and happy—city residents use water for water-balloon fights and swimming pools, too—takes a vast quantity of water, but it's still only a small proportion of the water that humans remove annually from lakes, rivers, and aquifers. The major uses of water in the United States are for cooling power-generating plants and for irrigating crops; these uses took up more than 80 percent of the water withdrawn in 1995 (U.S. Bureau of the Census 2001, table 351). Water for cities and small towns accounts for only about 11 percent of withdrawals, but we will concentrate on this use, since this is a book about cities, not about water. In general, per capita water withdrawals in the United States fell 24 percent between 1975 and 1995, largely because farms and power-generating plants produced more with less water, but per capita water withdrawals for cities and small towns rose 14 percent.

A few businesses and households in metropolitan areas have their own wells (for instance a lot of golf courses do, and a few bottling plants), but the vast majority of people and businesses get their water piped in from a central supply source. Such systems, with one source supplying many users, will be

the subject of this section. Thus we will often have to distinguish between water supply issues—questions dealing with the common source—and water distribution issues—questions dealing with the pipes that take water from the source to ultimate users.

B. What's the Externality?

The fact that water is necessary and that it's often expensive doesn't make water supply a public health issue. Many other nutrients—vitamin C and niacin, for instance—are also needed for human survival. We don't often hear arguments for municipal provision of vitamin C, and vitamin C played no visible role in the public health movement. What makes water different?

The answer you would expect in an economics book, of course, is externalities. The prominence of water supply in stories about improving health in cities would make you think there were externalities involved, but what are they? Are you ever better off because I use purer water?

Sometimes you are. The better my water supply, the less likely I am to contract a water-related disease. If I contract a water-borne disease and dispose of my excreta in a privy or a cesspool or even in a sewer that runs into a nearby lake, I put you and my other neighbors at risk. Even if I use a modern sanitary system, I still could spread my disease through food handling or some other form of accidental contact. Or the sewer system could be overwhelmed in a flood. So the higher the quality of the water *I* use, the lower the probability that *you* will contract a water-related disease, and thus you are better off.

This externality argument was clearly relevant for nineteenth-century cities in industrialized countries. Notice that the argument is not that you are better off the more pure water I consume (except for washing) but that you are better off the less impure water I consume. You don't gain when I fill my swimming pool. Supplying pure water to me isn't good for you in itself; it's only good for you to the extent that it discourages me from doing something that would be bad for you. Even in wealthy countries this is an argument that still has some force: there are very poor people even in very rich cities, and everyone in those cities would be worse off if the very poor sought out alternative sources of water.

Externality concerns are also still relevant today in developing countries. About 1.1 billion people lack access to clean water, and 2.4 billion lack access to basic sanitation (Water Supply and Sanitation Collaborative Council 2004).

There is another externality argument for a common water supply that takes the same form. In many areas of the world people can draw on naturally occurring sources of water—they can sink wells, for instance, and pump water out of an underground aquifer, or divert a stream, or lower a lake. Aquifers, streams, and lakes are in the short run exhaustible resources. The more water I

pump out of the aquifer, the deeper your well has to be drilled and the more likely it is that salt water will creep into it. You are better off if I draw on the public supply instead of drilling a well.

Notice that both of these externality arguments work on the principle of second-best. Start with activities—using unsanitary water or poaching from the aquifer—that cannot be priced or punished properly. Pure water is a substitute for these activities. Therefore, according to the second-best principle, the price of pure water should be lower than its marginal cost.

C. Why Are Governments Involved?

By themselves, of course, these two externality arguments aren't enough to make the case for a public water supply. But with water, externalities are not the only issue. The other important fact about water is that water distribution is a **natural monopoly** with a serious **lock-in** problem. (Lock-in means that once an investment has been made, some party has lost a lot of bargaining power.)

The cheapest way by far to deliver water is through a pipe. Bottling water takes bottles and energy; transporting bottles of water is tough work; and taking the water out of the bottles is a third source of expense. Slum dwellers in many developing countries get water from merchants who drive around with tank trunks, but this method is very expensive: people in the Kibeira slum in Nairobi, Kenya, pay on average five times as much for a gallon of water as Americans do (Water Supply and Sanitation Collaborative Council 2004). Turning on a tap and getting water from a pipe is much easier. With pipes, though, come **monopoly** power and lock-in.

Monopoly power arises because putting two or three different sets of pipes into a neighborhood would be expensive. It makes more sense to have one government agency or one company install a single set of pipes. Remember that there are increasing returns to scale in pipes—a pipe that carries four times as much water costs only roughly twice as much to construct, and less than twice as much to install.

Increasing returns come from demand smoothing, too. If ten of us on a street each wanted to be able to fill our 30-gallon washing machines in a minute, whenever we chose to, and we each dealt with separate water distributors, we would need a set of pipes thick enough to carry a total of 300 gallons a minute. But since it's unlikely that more than two or three of us would ever want to fill our washing machines at the same time, a single pipe with a 120-gallon-a-minute capacity would be more than adequate—and would cost only about twice as much to build and install as just one of the ten 30-gallon-a-minute pipes we would need if we used separate distributors.

Demand smoothing brings even bigger gains when a neighborhood has

hydrants for fire protection. Even a modest hydrant in a low-density residential neighborhood needs 250 gallons a minute when it's in use, but it's not in use often. So long as people don't all want to wash their clothes while their neighbor's house is burning down, the capacity needed to operate a hydrant usually isn't much less than the total capacity used by the buildings the hydrant protects. Whoever has installed the pipes for the hydrant has more than likely provided enough capacity for everyone else's water needs too.

It would be difficult to maintain competition among a group of different water distributors in the same neighborhood. If several firms started off competing against each other, the biggest could charge less and would soon become dominant, and the others wouldn't last (you can't really develop a niche market in water). New firms would be unwilling to enter neighborhoods where an existing firm was already operating; competition just wouldn't develop at all.

Lock-in makes the monopoly problem worse. In this situation, lock-in refers to the difficulty an individual customer or a community faces in switching to a different water distributor once the basic pipelines have been installed. People need water every day, and they need it in big enough quantities that transporting it any other way but by pipe is expensive. An individual household can manage on bottled water for a while, but not for long. New York City alone uses in a day about as much water as the entire American population consumes in bottled water in 100 days.[1] Once connected to pipes controlled by a particular distributor, a household or business has to pay whatever that distributor charges, up to the cost of bottled water. The consumer is pretty much at the distributor's mercy. A city is even more at the mercy of the distributor than a household is, since it can't just go to the supermarket and pick up a few bottles of water.

But lock-in works both ways. Once a private company has made a huge investment in pipes and reservoirs, it can't take them back and relocate to Indonesia—those pipes have no alternative use. And so the company is at the mercy of the local government just as much as the population is at the mercy of the company.

Monopoly and lock-in mean that an unregulated market probably won't induce people to consume the right amounts of water, even without the externality considerations. Monopolists can charge prices above marginal cost, since by definition they don't face competition, and so dissuade people from buying water from which they would derive benefits worth more than the cost to the distributor of delivering that water. If, say, the distributor is charging $4

---

1. New York City uses about 1.3 billion gallons of water a day. In 1999, per capita U.S. consumption of bottled water was 17 gallons (Murphy 2000). That means Americans drank 4.8 billion gallons of bottled water that year, or about 13 million gallons a day.

more per hundred gallons than marginal cost, then consumers are forgoing some uses of water that they value at $3 per hundred gallons more than marginal cost. Lowering the price of water to $2 more than marginal cost would therefore be a potential Pareto improvement. The public health externality makes the harm done by an unregulated water monopoly even greater: high prices may induce some people to use impure water or to poach on the aquifer.

By the same token, fear of expropriation may keep private water companies from expanding as much as they should. If the government can't guarantee that it won't expropriate the investment, that investment won't be made. This doesn't mean that private water companies are an impossibility: many people throughout the United States and the world get their water from private companies. But all of these arguments together imply that private companies, regulated or not, are likely to do a worse job of distributing water than of distributing, say, paper, and that a reasonable case for government provision can be made.

Is the same true of water supply? Probably less so.

There are many water supply technologies; different ones work best in different parts of the world. East Coast resort towns like Sarasota, Fort Myers, Boca Raton, Cape Hatteras, Newport News, and Cape May pump brackish water out of deep aquifers and desalinate it. Parts of Israel, Saudi Arabia, and some Caribbean hotels desalinate sea water directly. Most of Long Island's water comes from fresh-water wells. New York and many other eastern U.S. cities maintain surface reservoirs surrounded by large expanses of undeveloped land, many miles from the city boundary. Most cities on the Great Lakes pump water out of those lakes and purify it. Los Angeles and other southwestern cities transport water long distances from the snowy slopes of the Sierra Nevada Mountains, tap nearby groundwater, and divert some of the flow of the Colorado River. They also store water in underground aquifers. A final source of water is reclamation: treating wastewater and reusing it—generally for irrigation and other uses that don't demand great purity.

Most of these technologies have significant economies of scale and so are natural monopolies in regions that are not too densely populated, since transporting water is expensive. Tapping groundwater is usually possible on a small scale (as is harvesting rainwater), but a group of different water suppliers tapping the same aquifer will be working against each other, siphoning off a common resource. This will make production for each of them more expensive in the long run than if they had worked cooperatively. The advantages of cooperating in tapping a common resource will encourage mergers, and so monopoly is also likely in a region that relies on a single underground aquifer. The same is true for the watershed of a lake, river, or system of reservoirs. But a densely populated region with a wet climate may have access to several different aquifers and watersheds, and a producer in such a region may have some

ability to sell to more than one customer. So while water supply may sometimes be a natural monopoly, it is not always the case.

D. Pricing

How should water be priced? From the chapters on transportation, we know the answer by now: at marginal cost, except when people need to be discouraged from using impure water, and then it should be priced below marginal cost.

Water differs from transportation in many ways, however, and thinking about marginal cost in this case is quite a different exercise. The major difference is that water can be stored, unlike trips through the Lincoln Tunnel. Water that I don't use to wash the clothes today can be used to wash them tomorrow, but a trip I don't take through the Lincoln Tunnel or on the subway today creates no more tunnel crossings or subway rides tomorrow. The costs of using water today are not only the other uses I preclude or degrade today—the Lincoln Tunnel or subway case—but also the other uses I might preclude or degrade some time in the future.

These two kinds of costs of water correspond roughly to the two different physical systems—distribution and supply—and two different uses for the word *capacity*. Water *distribution* is like the Lincoln Tunnel—costs arise if my using water today precludes someone else from using it today, or degrades someone's experience today. With water *supply,* on the other hand, costs arise from future precluded uses or future degraded experiences, not current problems. Capacity, for water distribution purposes, is the maximum amount of water that can be delivered in an hour or in a minute—like capacity for a subway car or for the Lincoln Tunnel. For water supply purposes, capacity means the maximum amount of water that can be stored at any minute—quite a different concept, as we shall see.

Thus water prices involve two different questions—prices for water supply and prices for water distribution.

1. Water Supply

In this section we'll concentrate on reservoirs. First, think about a surface reservoir so big that even if water has a price of zero the reservoir will never run dry, even in a severe drought. Then no matter how much the reservoir cost to build, the marginal cost of water is zero (except maybe for chemicals to purify it and electricity to pump it), and so the price should be zero too. My using water from this reservoir prevents no one else from doing anything, and so it would be a shame if I couldn't take advantage of it to do something of benefit to me. Maybe I should even be paid to use water from this reservoir, if doing so makes me less likely to contract cholera or typhoid.

Next, to be slightly more realistic, think about a smaller reservoir that

isn't full now but will soon fill up and overflow. We know this for sure. If I use water from this reservoir now, the only consequence is that less water will spill out. No one will be any worse off. So in this case, too, marginal cost is zero and price should be zero.

So marginal cost is positive only with reservoirs that are subject to emptying in a drought. Now, to keep things simple, suppose we know how much water is in the reservoir now and we also know that it will not rain for another thirty days. If I drink a glass of water today, it could mean that someone will either have to go without some use of water later this month, or that water will have to be imported in tank trucks or manufactured through desalination. If tank truck importation or desalination is less costly than the most beneficial use that would otherwise be forgone, then the cost of replacing my glass of water that way is the marginal cost of my drink from the reservoir, and that's what I should pay. In this case the reservoir will empty before the month is out, but no one will notice.

However, if tank truck importation or desalination costs more than the maximum benefit somebody will have to forgo when the reservoir runs dry, then that maximum benefit is the marginal cost of my drink. In this case, price should be set so that the reservoir runs dry just when the rain starts to fall. That way every glass of water in the reservoir will be used, and none will

Box 9A

Evaporation and Interest

≡ In an analysis of costs of water supplied from a reservoir, evaporation and interest reduce marginal cost somewhat. If 10 percent of the water in the reservoir evaporates every month, and the reservoir will empty in a month, then my drinking a cup today deprives someone in the future of only nine-tenths of a cup of reservoir water. Similarly, if $1.00 saved today becomes $1.05 a year from now, I have to set aside only $1.00 today to compensate someone for $1.05 worth of added costs or forgone benefits a year from now, if that's when the reservoir empties. The effects of evaporation and interest work together. For example, if 30 percent of the water in a reservoir evaporates in a year, the rate of interest is 5 percent, and the certainty drought price is $1.50 a gallon, then if you knew for sure that the reservoir would run dry precisely a year from now, the marginal cost of a gallon of water would be $1.00 ($1.50 worth of water evaporates, so only 70 percent is left, or $1.05 worth; $1.00 today is enough to compensate somebody for losing $1.05 worth of water a year from now). In the text, I will for the most part ignore evaporation and interest questions in order to simplify the analysis.

Figure 9.1 Optimal price of water from a reservoir, given perfect foresight.
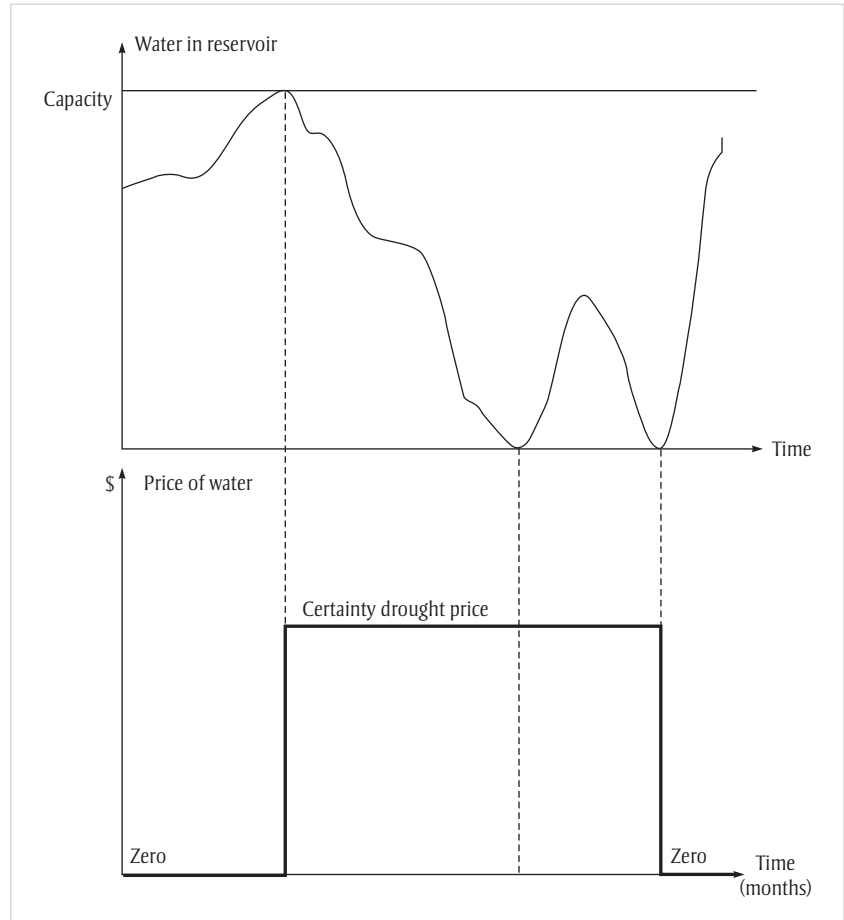
be used for a purpose less valuable than any use that was forgone. Every glass, for every day over the next thirty days, will have the same price.

Call the price that is charged in these circumstances—either the tank truck importation cost or the maximal forgone benefit cost—the **certainty drought price.** It's likely to be fairly high.

Pricing water in a reservoir is thus easy if we know the future perfectly. If the reservoir is destined to empty before it spills over, the price is the certainty drought price (corrected slightly, as described in Box 9A, for interest costs and evaporation). If not, the price is zero. Figure 9.1 illustrates how price should evolve over time in the case of a reservoir managed with perfect foresight.

But of course we don't know the future perfectly. We don't know how much it will rain or when, whether a major water main will burst next month, or whether a brewery using millions of gallons of water a day will slow down its operation. So at any given moment we don't know for sure whether the

price should be the certainty drought price—because the reservoir is going to empty—or zero—because the reservoir is going to spill over.

We can, however, guess how likely it is that the reservoir will empty. If, say, emptying is just as likely as spilling over, then it's just as likely that a glass of water I drink costs the certainty drought price as it is that it costs nothing. So I should pay half of the certainty drought price, because that's the expected cost I impose on the world. In general, the price of water should be the certainty drought price times the probability that the reservoir will empty before it spills over.

That means that the marginal cost of water will vary. When the reservoir is close to full and more rain is expected, the marginal cost of water is low. When the reservoir is close to empty and no rain is in sight, the marginal cost is high.

Day-to-day variation is governed by changes in expectations, and so price will change from one day to the next (when something surprising happens). On any given day we can be pretty sure that tomorrow's price will be different from today's, but if today's price is optimal, tomorrow's price is just as likely to go up as it is to go down. Unexpectedly heavy rains that fill the reservoir surprisingly full make running dry any time soon less likely and spilling over soon more likely; both effects reduce marginal cost. Unexpectedly heavy usage, or a leak that makes the reservoir low, will work in the opposite direction. (Thus with marginal cost pricing, a successful conservation campaign that persuades the public to consume less water rewards it with lower prices; with average cost pricing, conservation gets punished with higher prices.) Other prices vary from day to day—securities, vegetables, and airline tickets, for instance—and there's no deep reason why water prices shouldn't, too. (Customers worried about this variation could probably insure or hedge against it.)

The same basic reasoning applies to cities that get their water from wells and from rivers. You can think of an aquifer as a reservoir that happens to be underground and a river as a rainstorm that never quite ends, but varies in intensity.

With aquifers, the major new kind of cost to consider is **subsidence**—falling ground levels. The water in an aquifer supports the ground above it, and when too much of the water is withdrawn, the ground above sinks. Downtown Mexico City, for instance, dropped more than twenty feet in the last century because of subsidence (National Research Council et al. 1995). Subsidence is costly: buildings lose their stability, water and sewer pipes crack, and flooding becomes more frequent. The way to think about subsidence is to incorporate the costs it causes into the certainty drought price. When reservoirs get too low, water has to be imported or beneficial uses have to be forgone. When aquifers get too low, there is a third costly possibility—subsidence.

2. Water Distribution

Water pipes are similar to roads. Unless there's "congestion," their use should be free. Congestion in this case (as in every other) means that my use of the water pipes interferes with yours—I am using so much water at a particular moment that you can't get the water pressure you want when you step into the shower. In modern Western water systems, especially those designed to serve for fire protection, such events are rare. So use of water pipes should almost always be free.

E. Profitability

Will a reservoir with optimal capacity that charges optimal prices lose money, make a profit, or break even? Almost always, an optimal water supply system will make a profit. At the same time, of course, an optimal water distribution system will almost always lose money, since the price will be close to zero.

Optimal water supply systems make a profit for two reasons. The first is that capacity can be used over and over again as reservoir levels rise and fall—given that it does not either spill over or run dry. The amount of water that can be sold from a reservoir is in no way constrained by the reservoir's capacity.

To see this, consider a cycle between spilling over and running dry, like the one depicted in Figure 9.2. The reservoir almost empties out several times, but every time except the last, rains arrive in time and the reservoir almost fills up again. Several times the reservoir almost spills over, but it never does. Assume perfect foresight, to keep matters simple, so water should be priced at the certainty drought price throughout this cycle. During the cycle in Figure 9.2, the amount of water sold will be several times the reservoir's capacity, and thus the revenue collected will be several times as much as the revenue that

Figure 9.2 A cycle in which marginal cost pricing collects more than construction cost.
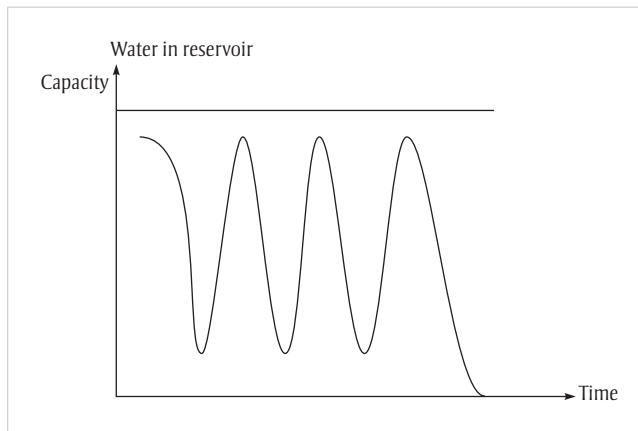
Figure 9.3 A cycle in
which marginal cost
pricing breaks even.
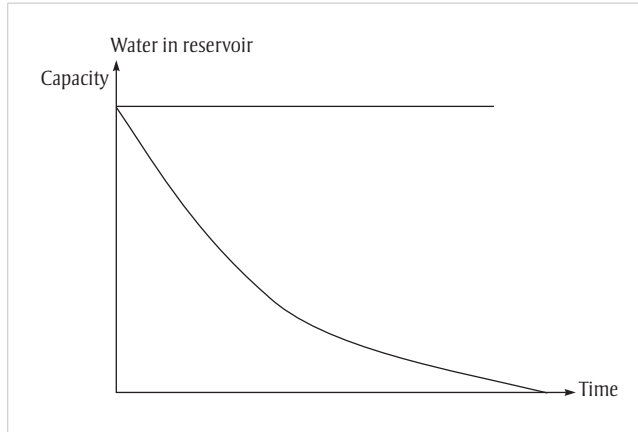
would have been collected if the supplier had sold *only* the reservoir's capacity at the certainty drought price.

When we think about optimal capacity, however, it doesn't matter how many times in a cycle the reservoir goes up and down. In the simple cycle shown in Figure 9.3, where supply falls without interruption, one more unit of capacity provides benefits worth only the certainty drought price. This is the exact same benefit that a unit of capacity provides in the cycle shown in Figure 9.2. So optimal capacity will be the same for both cycles.

When the cost of building each unit of capacity is the same (remember the condition required for breakeven with subway systems), optimal-capacity pricing breaks even if all cycles are like the unremitting downward trip in Figure 9.3. The easiest way to understand why is to suppose that this cycle is the only one that matters. The reservoir will be built (and won't be built infinitely big) only if the construction cost of a gallon of capacity is the same as the certainty drought price. The reservoir operator will collect the cost of construction for each gallon of capacity, and so will break even.

But few cycles are like Figure 9.3. Usually it rains sometime in between spillover and emptying, and so the reservoir sells more water than its capacity. This is one reason why optimal water supply systems make a profit.

One way to think about the extra rainfall is as a gift to the reservoir. Whatever capacity you build, the intermediate rainfall gives you more water. A more sophisticated view is that the relationship between the size of the reservoir (the amount of water it can hold) and its output (the amount of water that can be pumped out of it before it runs dry) is really one of decreasing returns to scale: doubling the size of the reservoir doesn't double the amount of water that can be pumped out of it. A reservoir a hundred times bigger than the current one won't produce anything like a hundred times more water, because almost all of the reservoir will be empty almost all of the time. De-

creasing returns to scale imply positive profits under marginal cost pricing: as the reservoir gets bigger and bigger, the amount of added output produced by each additional increment to size gets smaller and smaller; thus marginal cost is greater than average cost, and so profits are positive.

The second reason why optimal-capacity water supply systems are likely to make a positive profit is that there may be decreasing returns to scale in constructing capacity. The bigger the system to begin with, the more expensive it is to build another gallon of capacity. This phenomenon is more likely to occur with large systems. In small systems, the two-thirds rule may counteract other sources of decreasing returns to scale.

For big water-supply systems like those for New York and Los Angeles, added capacity is almost certainly going to be more expensive to build: it's going to be farther away and in a less desirable natural setting. You pluck the low-hanging fruit first, and so the more pieces of fruit you pick, the more difficult and expensive it is to pick the last one.

In contrast, optimal water distribution systems are likely to lose money, since they should rarely charge consumers anything. What happens when you combine water supply and distribution? For big cities in dry climates—Cairo and Los Angeles, for instance—the water supply component will be much more important that the water distribution component, and the supply profits are likely to outweigh the distribution losses. For wetter and smaller cities, the opposite could be the case. There's no simple, hard-and-fast rule about whether optimal water systems will make money or lose it.

For systems that make money with marginal cost pricing, there's also no hard-and-fast rule about what they should do with it. One possibility is to give it back to customers in payments that are not tied to water usage. For instance, each customer could start off the month with a three-dollar credit, be charged marginal cost for every gallon of water used, and have the credit increased the next month if she didn't use three dollars worth this month. Los Angeles inaugurated a pricing scheme something like this in 1993, except that instead of giving customers a credit, it gave them a price below marginal cost for all water used up to a certain cutoff point every month. So long as everybody reaches the cutoff point every month (which doesn't happen in Los Angeles), this is the same as a credit. The Los Angeles pricing scheme also uses different rates for winter and summer, and for drought and nondrought years (Hall and Hanemann 1996).

Similar systems are used in South Africa and Chile. Expanding water access for poor people was a major priority for the post-apartheid government of South Africa, and so it developed a system in which the first twenty-five liters of water per day are free; above that minimum, water is priced at marginal cost. In Chile, all water costs the same, but poor people get stamps that they can use to offset part of their water bill (*Economist* 2003).

Most cities in the United States do things differently. They base water

rates on a series of extraneous factors, like the size of the user's connection to the water main, and then employ what are called "declining blocks": the first x number of gallons cost so many cents per gallon, the next y gallons cost less per gallon, and so on. There is no good economic rationale for pricing water like this.

In many developing countries water supply and pricing systems have a different weakness. They charge prices well below marginal cost and confine service to only a few neighborhoods in a city, generally the wealthiest and most politically influential. According to a United Nations report, those urban Pakistanis lucky enough to have access to a public water supply were paying just 10 cents per cubic meter, while in comparison, people in Germany were paying $2.16 per cubic meter (Crossette 1999). Countries that charge marginal cost but make separate provisions for poor people, like Chile and South Africa, have been much more successful in expanding public access to water.

F. Externalities and Quality

What about the external health benefits of consuming clean water? Clearly our analysis of optimal pricing thus far has not taken them into account—we have simply assumed that water consumers receive all the benefits of their actions. But taking the health externality into account will not require major changes in the analysis.

Why not? Because people won't consume "dirty" water (water that may cause disease) unless its price is considerably below the price of clean water, and since water is a bulky commodity that's expensive to deliver except through a pipeline system, the authorities can prevent the consumption of dirty water by keeping it out of the pipes. Some people will forgo the convenience and cheapness of tap water to purchase bottled water if that bottled water is claimed to be purer or better for your health; hardly anybody would purchase bottled water known to be of worse quality than tap water. So as long as the optimal price of clean water is below some fairly high level, consumption of dirty water will present no problem and public health considerations won't enter into the analysis.

In developing countries, the major health issue is who has access to piped water. Policies that end up confining piped water distribution to the rich are thus doubly perverse.

In developed countries, where pipeline systems already exist, it is the quality of the water in the pipelines, not its price, that affects public health. Everyone on the same water distribution system gets water of the same quality: I can't order water with a few cholera germs in it unless my neighbors get water like that, too. Water quality can vary greatly; the distinction between "clean" and "dirty" is useful for thinking about the general issue of externali-

ties but not useful for thinking about how much chlorine to use, how many tests to run, or how large an area to close to development around a reservoir. Those kinds of questions involve difficult and complicated trade-offs between costs and benefits. Inherently they involve a degree of something like social engineering: different customers would give different answers for themselves if they could, but there can be only one answer for everyone.

In this sense, water purity is an example of a problem both created and solved by increasing returns to scale, in developed countries at least. It is only because large numbers of people want to live and work close to each other that water-borne contagious diseases are something to worry about; increasing returns to scale in both consumption and production are what draw them together. And increasing returns to scale in water distribution mean that they will all consume water of the same quality, and also that the water quality can be made high enough that almost nobody has to worry about water-borne contagious diseases. The cure for city life is sometimes more city life.

II. Sewers

A. Why They Matter

Sewer and water systems need to be studied together. Both combat the spread of water-borne diseases, and in that sense they are substitutes. But they are complements, too. You can't operate a modern Western toilet unless you have some source for the copious amount of water it needs to operate, and you can't deal with the huge quantities of water a modern Western water system sends out of your house unless you have an excellent sewer system.

Like water systems, sewers solve a problem that population density creates. If you are alone in the forest, where storm water runs off naturally into streams and is absorbed in the ground, the water that you wash with—the technical term is *sullage*—can run harmlessly back into the stream in which you're washing, and you can leave your excreta behind a tree where it will decompose naturally. At the slightly greater densities of a rural community, each household might want its own cesspool or septic tank for excreta, but sullage and storm water can still be allowed to find their way into streams and into the ground. Only at the kind of densities found in cities do all three forms of sewage become a problem. (*Sewerage* means the pipes; *sewage* is the stuff that runs through them.)

Cities use many different technologies for handling storm runoff, sullage, human excreta, and fluid industrial waste. The technology now used in Western cities is almost certainly inappropriate for developing countries, and it is probably inappropriate for Western cities, too, but it's too late to change now. The Western system uses far more water and far more capital than the best

systems available for developing countries today, and it likely uses more capital and water than the systems Western cities should install if they were starting from scratch today.

Western sewer systems are inappropriate primarily because they were designed to fight "miasma," not germs, and we know now that miasma doesn't cause disease. Eighteenth- and nineteenth-century physicians, however, did not understand how the four main urban killer diseases—tuberculosis, cholera, typhoid, and malaria—were transmitted. They attributed all of them to miasma, or a noxious atmosphere. One British official described miasma this way in 1847: "This disease-mist, arising from the breaths of two million people, from open sewers and cesspools, graves and slaughterhouses, is continual . . . [I]n one season it is pervaded by cholera . . . at another it carries fever on its wings. Like an angel of death it has hovered for centuries over London" (cited in Williamson 1990, p. 283).

Western sewers were designed to eliminate stench, not pathogens. The idea was to whisk excreta away from the population and to keep the water moving, not to prevent infection. In the 1849 cholera epidemic, for instance, London authorities systematically flushed the sewers to dispel the miasma. This action directed a flow of infected excreta right into the section of the Thames from which water companies were drawing their water, and accelerated the spread of disease (Sah 1997, pp. 22–23).

The miasma theory played a prominent role in two other major mistakes that nineteenth-century sewer builders made. The first mistake was to design systems that discharged into watercourses directly without any kind of treatment. Rivers and lakes were believed to have natural self-cleansing properties and to easily disperse the miasma. Even when the germ theory of disease became dominant, many engineers still argued that treating drinking water was cheaper than treating sewage.

Only after World War I and the complete triumph of the germ theory did the baleful consequences of dumping untreated sewage into watercourses become clear. The self-cleansing powers of lakes and rivers were really a lot weaker than the nineteenth-century engineers thought they were, and raw sewage affected not only the communities dumping the sewage and drawing water from these watercourses but also the people living and working downstream. The result in the United States has been close to a century of building ever more powerful and ever more expensive sewage treatment plants, a huge expense never contemplated when the nation's sewer systems were first being planned.

The second mistake was the construction of combined sewers—sewers that carry storm runoff as well as household and industrial wastewater—instead of separate storm and sanitary systems. This made building the later treatment plants even more expensive and difficult. As they were designed, combined sewers allowed storm runoff to hasten the movement of miasma-

bearing excreta away from the city, and they were cheaper to build and oper-ate—provided sewage didn't have to be treated before it was dumped. (Maximum pipe throughput quadruples when the pipe's diameter doubles, and so there are increasing returns to scale in the construction of pipe capacity; a single big pipe that carries both storm runoff and wastewater is cheaper to build than two small pipes.) But treatment is much more difficult with a combined system. You either have to build a treatment system appropriate for the sanitary sewer, but turn it off and let everything flow raw into the river when it rains—New York City was operating a system like this well into the 1980s, and Milwaukee still has a system like this—or you have to build a treatment system with a very large capacity that will go unused most of the time.

B. Pricing

These design mistakes have been compounded by the way sewer services are priced. Some services are essentially free, and other services are priced at average cost; one price is too low and the other is too high. As a result, city residents and businesses cannot derive all the benefits they should even from the wrong technology they have inherited.

The most notable free service is handling storm-water runoff. Big storms (or in places with low capacity, moderate storms) are the times when sewer capacity is fully used, and when additional inflows cause damage—either sewer backups in low-lying places or flows of untreated sewage into watercourses. Storms are for sewers what rush hours are for roads. Actions people take that increase the volume of storm-water runoff are therefore costly.

What actions are those? The most obvious one is paving. What matters for storm-water runoff is the amount of water that reaches a sewer in the first few hours after a storm—not the amount that reaches a sewer eventually. Rain that falls on an impermeable surface like a parking lot or a street has nowhere else to go; it flows into the sewer quickly. Some of the rain that falls on parks and lawns runs to the sewer immediately too, especially in a hard rain, but a lot sits in puddles or percolates into the ground. Some of this water may find its way to a sewer eventually, but not during or right after the storm, and that's what counts.

Paving a field is a costly endeavor—and so is keeping a field paved. My decision not to tear up a paved parking lot before June 15 imposes costs on a lot of other people if there's a severe storm on June 15, and so I won't make that decision correctly unless I bear those costs. Because the legal system doesn't allow people to sue parking lot owners when sewers back up or sewage flows into a watercourse, this is a cost that could be handled through billing by governments or sewer authorities. Not billing results in too much paving and too many sewer problems (or too much sewer capacity). This compounds one of the problems I noted in Chapter 4: the unduly favorable tax treatment

of employer-provided parking. Thus the handling of storm water is priced too low.

On the other hand, many jurisdictions charge for household and industrial sewer service at average cost—that is, rates are set to break even. The actual marginal cost of sewer use is quite a bit lower than this. Marginal cost is essentially only the **variable costs** of treatment—chemicals, electricity, sludge dumping, and wear and tear. The treatment plant is a sunk cost; so are the sewer pipes that have already been installed. Some cities don't charge residential customers for sewers; others charge an annual fee unrelated to usage. Considering actual marginal cost and the difficulties of metering, these last approaches have much to commend them.

In the short run, the consequences of overpricing are not especially severe for most residential sewer users. Residential demand is fairly inelastic; most people are not thinking about price when they stumble toward the toilet in the morning. Charging average cost is probably regressive—since the residential sewage that people produce doesn't rise too quickly as their income increases—but it's not likely to cause serious distortions in how people behave in the short run.

The same is not true for factories, however. Industries like breweries, leather plants, and paper mills discharge large amounts of fluid waste, much of it containing exotic chemicals. Often these companies are metered specially and billed according to the chemical content of their waste. They have many options: they can partially treat their sewage before they release it; they can change their production processes; they can build a "private sewer" and bypass the regular system altogether. Average cost pricing leads them to choose among these options incorrectly. They make investments in things like pretreatment and private sewers to avoid paying the difference between marginal and average cost, when everyone could be made better off if they didn't make these investments, stayed fully in the system, and paid something closer to marginal cost.

In the long run, when households and firms have time to make decisions about where to live and work, the consequences of average cost pricing may be more pernicious. Think of a family trying to decide between a house with an existing sewer connection to a treatment plant not often operating at full capacity, and a new house somewhere far from existing settlements that will require a septic tank. With the distant house, they will bear the full cost of their decision—they will need a septic tank and they will pay for a septic tank—nothing less, nothing more. But with the house that connects to the existing system, they will have to pay more than the costs they actually impose on the rest of the world, because they will pay some of the capital costs for a system that's already built and some of the **overhead** costs that would be incurred no matter where they lived. So average cost pricing of sewers (and gas and electricity) may induce some households (and firms) to move to more distant lo-

cations, even though everyone could be made better off if they lived closer to older, established cities.

Average-cost sewer pricing makes cities less dense than they should be, even when the alternative is not a septic tank. Compare the same tract of land in an open-city model under average-cost sewer pricing and under marginal-cost sewer pricing. Under average cost pricing, the more people are living on this plot of land, the greater the share of the capital and overhead costs the tract, and ultimately the owner of the land, will have to bear. Under marginal cost pricing, the share of capital and overhead costs that the tract's residents and owner bear does not depend at all on how many people live there. So average cost pricing makes owners and developers want less dense development; it's another reason for excessive sprawl.

Of course, if sewer systems are priced at marginal cost (or below it), they'll probably lose money, even with fees from owners of paved surfaces. How would economists make up this deficit? As with roads and mass transit, the answer is a land tax (although the surplus from a water system might also be used). The availability of a sewer system and sewage treatment capacity makes land more valuable, whether it's used for a tennis court or for an apartment building. Paying deficits with land taxes aligns landowners' political activity more closely with true net benefits, and has certain elements of fairness.

C. Capacity

Since 1956, sewer plants in the United States have had their capital costs subsidized, but by the federal government, not by affected landowners. Under the 1956 Amendments to the Water Pollution Control Act, municipalities received federal grants of up to 55 percent of the cost of building sewage treatment plants; Congress raised the ceiling to 75 percent in 1972 and lowered it back down to 55 percent in 1981. Obviously the federal subsidy has alleviated some of the problems average cost pricing causes, since it brings the "average cost" that sewer authorities must recover closer to actual marginal cost, but it also introduces new distortions. Because much of the money for construction has come from federal taxpayers, municipalities have not been as careful as they might have been in controlling construction costs. One study claimed that construction costs could have been cut by 30 percent if localities had handled federal money with the same care they usually exercised with their own (Congressional Budget Office 1985). So "average cost" was not reduced as much as the formulas implied.

More seriously, the geographic allocation of federal funds may have favored less dense parts of metropolitan areas at the expense of older, denser cities. Especially before 1977, new sewer plants were built in new areas, and so the federal subsidies made moving away from old places cheaper.

Since both pricing mistakes—charging too little for storm runoff and too

much for household and industrial sewage—encourage sprawl, you might think that if sewers were handled better American cities would be more dense, and that cities in developing countries would end up denser if they corrected Americans' sewer mistakes. But probably not.

On the one hand, there is little serious empirical evidence that sewer pricing and construction policies have contributed measurably to excessive sprawl. (That governments can stop new development in outlying areas by denying sewer permits is not evidence that sewers contribute to excessive sprawl. You could stop those developments by bombing them, too.)

On the other hand, the really significant problems with Western sewers may promote excessive centralization, not excessive sprawl. Remember that pricing is only one of the many sewer mistakes. Probably more serious is the centralized capital- and water-intensive technology—a nineteenth-century technology designed to combat the imaginary problem of miasma—that Western countries are now saddled with. Most of the new technologies—especially those that are appropriate in arid and semi-arid climates—are much more decentralized, essentially acting like Portosan portable toilets. Pipes seem like the best technology for delivering water, and probably for removing sullage and channeling storm water, but whether they should also handle human excreta is an open question.

III. Fire Suppression

A. Why It Matters

Before the middle of the twentieth century, huge fires periodically ravaged even the greatest and most up-to-date of cities. The Chicago fire of 1872 is the most famous conflagration in American history, but it was not a unique event. Most of London, for instance, burned in 1665, and a large portion of New York went up in flames in 1835 (the dollar value of damage done in this fire was greater than the entire cost of building the Erie Canal [Burroughs and Wallace 1999, p. 598]). Calgary was destroyed in 1886, and most of Paterson, New Jersey, burned in 1902. Major parts of Tokyo were destroyed by fire in 1872, 1881, and 1911, and the fires following the great Kanto earthquake of 1923 leveled most of the city (Shay 1994). You can't operate a flourishing business if you think your equipment and your inventory are likely to burn up after you've gone home; you can't go to sleep easily if you worry every night that your family could be killed and all your possessions destroyed by a fire.

Like sewage removal, fire suppression is a problem that density creates—and density helps to solve. If a barn burns to the ground, that's bad news for the farmer who owns it but not bad news for the farmers' neighbors, except in rare circumstances. It's difficult for fire to migrate more than a few meters where there are no intervening structures. Only when buildings are jammed close together—as they are in cities—do people need to fear rapidly spreading

conflagrations. (Forest fires are something of an exception to this principle, but small forest fires do little damage and big forest fires are pretty close to unstoppable.) In a city, when your house catches fire, all your neighbors have cause for concern.

Modern Western cities use high water pressure, hydrants, sophisticated fire-fighting apparatus, wide streets for maneuvering that apparatus, and a trained fire-fighting service to extinguish fires. We've seen that a water system's capacity is designed for fighting fires—the rest of us are essentially off-peak users who should bear no cost for capacity. For most water systems in developed countries, only when fires are being fought is "congestion" a problem—that's when there may not be enough water for everybody who wants it. A single fire hose uses 250 gallons a *minute;* I use about 50 gallons a *day.* It's not uncommon in New York City to have three fire hoses going at once to put out a structural fire, so that's a lot of water being drawn from the system. (Similarly, the road width in many areas is dictated by the needs of emergency vehicles, not ordinary traffic.)

It's not surprising, then, that the fear of fire motivated the historical development of water systems as much as the fear of disease did. Cholera forced roughly half of New York's population to flee to the countryside in the summer of 1832, and killed about 4 percent of those who remained in the city. But it was people who feared fire—warehouse owners and fire insurance companies—who were the strongest proponents of the modern water supply system that New York voters authorized in 1835 (Burrows and Wallace 1999, pp. 591, 594–595).

Conversely, fire insurance companies, in evaluating what their risks are in a particular city, look at the water system as much as they look at the fire department; both receive equal weight in their rating schemes (Brueckner 1981). This may overstate the social importance of water, however, since fire insurance pays for loss of property but not loss of life, and the rescue efforts of firefighters are relatively more important in saving lives than saving property.

B. Emergency Services

Fire suppression, then, requires a lot of capacity—water pipes, apparatus, wide roads, firefighters—most of which is almost always idle. In 1992, the average New York City firefighter was called to a structural fire about every other workday (Citizens Budget Commission 1993, p. 11). By itself this is not strange. Great musicians spend a very small part of their lives performing; expensive wines age for many years before anyone gets any value from them. So if people paid great amounts for the use of these pipes, skills, and apparatus for the little time they were used, fire suppression would be just like musicianship or fine wine. But they don't—you don't pay anything when the fire department puts out your fire.

This situation makes sense because firefighters, hydrants, and wide roads

can be extremely valuable if they are provided in a crucial five- to ten-minute period. The same is true for ambulances and some police responses. Extinguishing a fire that would have caused $100 million in damage is a service that's worth $100 million, but it's not a service you can provide whenever you feel like it. Arrive ten minutes before a fire starts and there's nothing for you to do; arrive thirty minutes after it starts and there's not much you can do. Once a fire is burning, every minute of delay is very costly and every minute saved is very valuable. But before a fire starts, time doesn't mean much. Fire suppression is an emergency service—time is of the essence—and costs and benefits of emergency services are slightly different from those of the routine services we have considered up to this point.

To begin with a simple and unrealistic case, suppose we know for certain that a fire will start at 420 West 118th Street at 11:37 AM on June 25, and that if it's extinguished within ten minutes it will cause no damage. Extinguishing it takes two companies of firefighters with their apparatus (and a road big enough for the apparatus) and a high-pressure source of water. (A "company" of firefighters means a squad that works together; it doesn't mean a business.)

The cost of extinguishing the fire is the cost of delivering the firefighters, apparatus, and water to the scene by 11:37. You have to make sure the firefighters have stopped whatever else they are doing and are standing ready to go at 420 West 118th Street at 11:37. They must have been trained. You need high-pressure water. The apparatus has to be right there, too.

Unless all the water capacity is going to be used again—say, because there will be a fire at 419 West 118th Street next week—the cost of building the additional capacity is a cost of preventing this fire (although if there weren't water pipes on West 118th Street before the fire, the consumer surplus realized by households that later used the pipes would count as a benefit of putting out the fire). If the benefits of putting out the fire exceed the sum of all these costs—and they probably do—assembling the task force to put the fire out is a potential Pareto improvement.

Time is the important element here. To get everything ready to go at 11:37 AM on June 25, you have to start getting things in place well before then. If you start looking for firefighters and laying the pipe at 11:36, you are not going to stop the fire. Most of the costs of fighting the fire have to be incurred well before the fire starts.

I don't mean to imply that actually fighting a fire is costless. Three hoses working for an hour use more water than I use in a year. Water-pressure drops may foreclose other valuable uses of water while the fire is being fought. And the greatest cost is to the firefighters, who risk their lives and often suffer serious injuries while working hard under extremely trying conditions.

The important point, though, is that however great the costs that must be incurred after a fire starts, there are other costs that are incurred before the fire. If you prepared for the fire and it didn't happen, it would still cost you quite a bit.

The timing of costs and benefits makes it difficult to operate an unregulated market in firefighting, even when the time and location of fires is known with certainty. Suppose there is a firefighting firm and an owner of 420 West 118th Street, who will bear all the cost of the fire if it is not extinguished. They negotiate well before the fire. The firefighting firm will incur its large up-front costs only if it can recoup them. If it arrived at the fire to find another firefighting firm standing there ready to go, the owner could play them off against each other. The low bidder would receive only a little more than marginal cost at that time—water and firefighting time—and it would not recoup its pre-fire costs. Therefore the firefighting firm will not agree to a contract unless it gets its money first, or has an assurance that it will be the only firm at the fire.

But if there's only one firm at the fire, the building owner has a huge ex post opportunism problem. If it's 11:37 and only one firm is at the fire, then it's clear what that firm will charge for putting out a fire that would otherwise cause $100 million in damage: about $99 million. There isn't even any time to dicker. The opportunity for ex post opportunism is present even if the owner had already paid the firefighting firm up front—the door to renegotiation can always be opened. So the only way the building owner can avoid paying $99 million to a firefighting firm is to contract with two or more firms, give them all their pre-fire costs up front, and hope they won't collude when they get to the fire. Not an inviting prospect.

(Where's the inefficiency here? Not in the firefighting process—the building will be saved if and only if it's cheaper to extinguish the fire than to let the building burn down. The inefficiency comes in making decisions about whether to build the building. Suppose the building is going to be profitable if the owner can collect rent for fifty years, and the building will in fact last fifty years if a couple thousand dollars are spent on fighting a fire in its tenth year. In this case everyone can be made better off if the building is in fact built. But if the owner can really collect rents until only the tenth year—because all profit after that will have to go to the firefighting firm—then the building won't be built.)

Avoiding ex post opportunism, moreover, is not the only problem in the organization of firefighting. Firefighters are people who run into buildings when other people are running out, and so they cannot be directly monitored by those who would be paying them. They have to make quick decisions—where to go, what to ventilate, where to direct the hose—under murky, trying, uncomfortable, and dangerous conditions. They have to make instant decisions about whether to try to save lives or try to save property, and about what chances to take with their own lives. When they fail to save lives or property, or fail to save what the owner most wanted to save, it's difficult to tell whether they're at fault or the cause was hopeless or too dangerous for them to pursue. It is hard to think of a good monetary reward system for firefighters.

The two organizational problems are solved the same way—by making

bravery important to the people who are firefighters, and making monetary gain an insignificant question at any working fire. Unless firefighters are reluctant to ask for bribes from the building owner and unless they evaluate their work among themselves, there is no good way of organizing their work. (The only alternative is for the firefighters to own all the buildings, but this would raise many other problems; moreover, firefighters protect lives as well as property, and it is not feasible or desirable for them to own other people's lives.)

The usual way to solve these incentive problems is to make firefighting a rewarding career, in terms of camaraderie and esteem as well as money. Firefighters who ask for bribes or who fail to make their best effort to extinguish a fire must be subject to loss of these career rewards. It must be more than worth firefighters' while to continue in their role; it must be something they would be very unhappy to give up.

C. Costs and Prices

The cost of suppressing a fire, then, is not so much the cost of what happens after the fire has started but the cost of being prepared—the cost of pipes, wide roads, training, a nearby firehouse, and creating an ethic that will keep firefighters from being extortionists. Putting all these elements in place is equally costly whether 420 West 118th Street burns or not. So these elements are not the marginal cost of putting out that fire. They are fixed costs.

Leaving behind the unrealistic assumption of a single fire at a known time and location will make the contrast between marginal cost and fixed cost more obvious. The water pressure, wide roads, and nearby, trained, conscientious firefighting force that were needed to extinguish the fire at 420 West 118th Street can easily do the same at 419 West 118th Street—all it takes is more water and more time. Whenever the owner at 420 West 118th Street has a firefighting force ready to extinguish a fire, so does the owner at 419. It's like a gym or a library; the weights, treadmills, and books are there even if nobody ever uses them, and since not everybody wants to use the same ones at the same time, one set can serve many people. The marginal cost of adding members to the gym, or properties to those covered by a firefighting force, can be small. And once a person is a member of the gym or of the group covered by a firefighting force, the cost of a visit is usually so small it doesn't make sense to charge per visit.

Charging for visits by a fire truck makes even less sense than charging for visits to a gym, for several reasons. The first reason is time: getting someone to agree to pay for the visit could take at least a few minutes, and a great deal of damage can occur in that time. The second reason is externalities: the building owner is not the only one who could be hurt if the fire were not extinguished quickly: everyone in the surrounding neighborhood would be in danger if the fire spread. An owner might be inclined to let a fire burn if the cost of extin-

guishing it were more than the private loss he would suffer, and yet the damage to the whole neighborhood could be far greater than the cost of extinguishing the fire. (This is especially true in the case of nonstructural fires—those in vacant lots, at railroad crossings, and in cars. In 2000, about half the fires that New York City firefighters responded to were nonstructural, and many of the structural fires occurred when no one was on the property who could have agreed to pay for the fire to be extinguished.)

The third reason that charging per visit does not make sense is **moral hazard:** most properties are covered by fire insurance, and so the fire insurance company, rather than the owner, would bear most of the cost of the fire. Making the owner pay for the visit would not eliminate inefficient visits, and charging the fire insurance company would not produce the efficient number of visits either, because much of the damage that a fire causes—loss of life, the pain of injury—is not fully covered by insurance.

In some ways, the second and third reasons are just variants of the first. There would be no inefficiencies if the owner, the neighbors, and the fire insurance companies got together in Coasian fashion to decide whether to call the fire department, after making appropriate side payments. But in the time that would take, the whole city could burn down. Once a fire gets started, transaction costs are very, very high.

Thus, in most cases fire departments don't charge for individual visits. In the United States there is one exception to this policy—false alarms—but what people are charged for a false alarm is punishment for bad intent, not an estimate of marginal costs. (Penalties for fire responses are more common outside the United States.) False alarms are the inevitable product of a system in which getting the fire department to visit you is free. People will induce the fire department to visit them for all sorts of trivial reasons: to amuse themselves, to enliven an otherwise dull day at school, to respond to a dare from juvenile friends. In 2000 the New York City fire department responded to 56,000 false alarms and only 58,000 fires. Punishing severely the small number of people who can be proved to have called in an alarm for flimsy reasons is one way to reduce some of the problems that free service causes.

Although charging for fire suppression once a fire has started is not a good idea, charging beforehand for expected fires can create some potential Pareto improvements. A fire company—and the associated apparatus, pipes, wide roads, and firehouse—is like an empty subway car, the kind of congestible resource I talked about in Chapter 5. Maintaining a house or a factory or opening up a store in the area the fire company responds to is like riding in the subway car. With the subway car, the right charge for riding in it was the marginal cost—the additional discomfort and lost time you imposed on your fellow passengers—and the same is true for fire suppression.

What's marginal suppression cost? Partly, of course, it is the marginal cost of an actual fire, should one occur—the water and the firefighters' effort

and exposure to danger. The more buildings a fire company covers, the greater the expected number of fires, and the greater the size of these costs. But water and effort aren't the only costs in suppressing a fire. If the fire company is responding to your house, it can't respond to my house, too. (This is the major cost of false alarms.) When your property is on fire, I am at greater risk because responding to a fire on my property would take longer and water pressure would be lower. (In most urban areas, response wouldn't be impossible; it would just take longer because a company from a neighboring region or town would have to respond.) A longer wait for a response could be very costly.

Therefore buildings that are more likely to have fires should pay more—wooden buildings rather brick or stone ones, abandoned buildings, barbecue restaurants and restaurants that use a lot of grease. The cost that an actual fire imposes on other properties also matters. Buildings far from the firehouse require more travel time; more travel time means that the company is tied up longer, and so the other buildings lose more protection. More distant buildings are more expensive to protect.

The number and characteristics of the other buildings a firehouse covers also affect the cost of a fire in an individual building. If a firehouse covers only one building other than mine, it's unlikely that the other building will burn at the same time as mine, and so my fire will not much increase expected response time to that building. If there are 10,000 other buildings, the probability that my fire will hinder response to at least one of them is much greater, and so the cost of my fire is much greater. This is just like congestion on a subway car: with one other rider I don't get in anyone's way, but with 200 other riders I get in the way of a lot of other people. Similarly, the more valuable the other properties—and lives are very expensive—the greater the cost of a delay caused by my fire. So the marginal suppression cost for a building in a big, densely populated fire district is greater than for one in a less populated area.

D. Capacity and Preparedness

A firehouse that covers a small, sparsely populated area and charges marginal suppression cost will lose money; one that covers a large, densely populated, valuable area will make a profit. What does this tell us about the optimal arrangement of firehouses and fire companies?

The answer is much like the answer we derived for trains in Chapter 5. Think of capacity as the number of identical firehouses in a metropolitan area. Increasing capacity makes paying firefighters, buying apparatus, and generally maintaining fire suppression efforts cost more, but it reduces the costs that the public suffers from delayed response. If the conditions in Chapter 5 are met—continuously variable capacity, constant per-unit cost of capacity, response delay cost to the public a function of the utilization ratio alone, and

no other distortions—then at optimal capacity a fire department that charges marginal cost will break even.

Economists who have studied the costs of fire suppression have concluded that these conditions probably do not hold, however. The problem is the utilization ratio: per capita costs of fire suppression seem to be lower the larger the population covered, holding density constant. Two companies covering two equal-size districts do better than one company covering one district (Brueckner 1981). The reason is probably the demand smoothing I talked about in Chapter 2: having one company available to back up the other makes it a lot more likely that someone will be able to respond to a fire. Two companies together can also fight bigger blazes better.

Thus a fire department that uses marginal cost pricing will probably lose money; marginal cost is less than average cost. The empirical estimates suggest that this loss will not be great, but it will be a loss nonetheless. Land taxes, as usual, are not a bad way to cover this loss—all land in a fire district is more valuable because of the firehouse, even if it is undeveloped land, since the option to build is more valuable when protection is present.

How does the way firefighting is actually organized and paid for in the United States affect sprawl? Within a fairly homogeneous fire district, average cost pricing through the usual tax on real property probably has the same effect that average cost pricing of sewer services does—it encourages sprawl by discouraging density. But since the difference between average and marginal cost is not large, this effect is almost certainly small.

IV. Solid Waste

A. Why It Matters

The problems of solid waste are like the problems of sewage. They are both urban problems: on a farm or in the woods you can burn solid waste or let it decompose, and no one will be affected. Similarly the argument for public provision of solid waste disposal is the second-best principle: it discourages people from engaging in behavior that produces negative externalities. Your putting out three bags of garbage instead of two doesn't make me better off in itself. I like your garbage to be picked up because if it is picked up, it's not being burned or dumped in my backyard or putrefying next to my kitchen window. The externality comes from what you don't do, rather than what you do.

Solid waste not properly disposed of causes a variety of problems. It "can attract and promote the breeding of undesirable and potentially disease-transmitting or disease-causing insects, rodents, and pathogens," promoting the transmission of schistosomiasis, South American trypanosomiasis, and Bancroft filariasis, for example, diseases that are closely associated with poor

solid waste practices in developing countries (Bloom and Beede 1995, p. 22). In 2000, the World Health Organization estimated that these diseases caused the loss of about 8 million "disability-adjusted life years"—not a huge burden compared with, say, tuberculosis, but still more than leukemia or ulcers (World Health Organization 2002).

Solid waste can also clog open drains and sewers; this creates breeding grounds for mosquitoes that transmit malaria and dengue, and also leads to flooding (Mensah and Whitney 1991, p. 163). Waste can catch fire, too: trash fires on the tracks are one of the leading causes of delays in the New York City subway system. And garbage is ugly and smelly.

Solid waste differs from sewage in several important dimensions. For one thing, there's a lot less of it. The sewage that the average American produces on an average day—mostly sullage—weighs hundreds of times as much as the solid waste he or she produces. But solid waste is harder to move—liquids flow through pipes, but solids don't. Wherever solid waste goes, somebody has to carry it there.

Because solid waste doesn't move on its own, a more sustained effort has to go into collecting it than into collecting sewage. This rule holds internally within households as well as externally, for whatever agency transports solid waste from the household (or the street) to its ultimate destination. Flushing a toilet is a lot easier than bundling up two weeks' worth of newspapers, and it's hard to think of any household task that's less onerous than letting the shower water run down the drain. Conversely, antisocial disposal of solid waste is often easier than antisocial disposal of sewage. Because it's much more common and socially acceptable while strolling home from the movies to eat a cup of ice cream or munch on a candy bar than to wash dishes or defecate, you are more likely to find napkins, plastic spoons, and candy wrappers on a busy sidewalk than sullage or human waste.

Finally, solid waste is much more heterogeneous than sewage. In part this is definitional: sewage is the waste from processes that use a lot of water, and solid waste is everything else people want to get rid of. So solid waste includes old tires and wedding invitations; furniture and batteries; clippings from the garden and discarded clothing; half-eaten hamburgers and kitty litter; plastic packing peanuts and almost-empty ketchup bottles; cereal boxes, broken dishes, defunct light bulbs, yesterday's newspaper, and the first three drafts of this book.

B. Pricing

Because illicit disposal of solid waste is so easy, handling solid waste the economically obvious way—charging for it to be picked up—is probably not an optimal solution. As a result, for most people garbage disposal is a free good—they pay for it either in municipal taxes or in annual fees, but neither method

punishes them for putting out more garbage or rewards them for putting out less. People probably put out too much garbage as a result; putting out less, and having more money and maybe less pollution, would be a potential Pareto improvement. People also probably buy goods that are less durable, retire them sooner, and hold fewer garage sales than they would in a world in which they had to pay for disposal.

To be sure, a few communities in the United States have instituted fee-for-bag programs (about 200 out of 35,000 municipalities and townships). In these communities, households generally buy either special program bags or stickers to attach to their own containers. Only containers with this special identification get picked up. Fullerton and Kinnaman (1997) studied these communities and estimated that the fee system reduced the weight of garbage collected by 14 percent (people packed it into bags a lot more tightly) but that about a third of that reduction was the result of more illegal dumping. If the social cost of having a kilogram of garbage dumped or burned illicitly is more than twice the social cost of legal disposal, then the Fullerton and Kinnaman result implies that pricing garbage collection in this fashion is not a potential Pareto improvement; the reduction in the amount of garbage to be collected would not be worth it.

The heterogeneity of solid waste also has important economic implications. Since different kinds of waste can be processed differently, separating them from each other is valuable—but not infinitely valuable. Yard waste can be recycled as compost, but batteries cannot—and neither can a mixture of yard waste and batteries. The whole is less valuable than the sum of the parts. The questions about separation are when it should occur—whether, for instance, households should keep their yard waste separate from their batteries and never mix them, or whether the two should be sorted out later at some kind of processing facility—and whether it should occur at all. The distinction between sewage and solid waste already recognizes the idea that households ought to be responsible for some separation—no one thinks you should flush newspapers down the toilet or drain dish water into a garbage can—but it doesn't make sense to ask households to separate ink from paper when they finish reading the newspaper.

If the same person bears all the cost of separating and gets all the benefits from separating—a mass of garbage that is more valuable after the separation than before—then the right decisions will be made, in the sense that no other set of decisions could produce a potential Pareto improvement. Sometimes the market can encourage a great deal of recycling, especially when labor is cheap relative to the commodities that recycling produces, since separating different kinds of garbage is often a labor-intensive activity. In Jakarta, for instance, "an estimated 30,000 to 60,000 scavengers extract recyclable materials such as glass, paper, cardboard, metals, wood, rubber, bones, and textiles from the waste stream. They sell the materials to small-scale entrepreneurs who

sort, clean, bundle, and sell the materials to other middlemen who specialize in particular materials, which they in turn transport and sell to recycling factories" (Beede and Bloom 1995, p. 43). In the 1980s, more than 2,500 scavengers lived and supported themselves in Mexico City's dumps (Oster 1989).

But if households can dispose of their garbage for free, there will be too little recycling, especially of items that households can separate most cheaply themselves. When households do not pay for the garbage they dispose of, they will gain nothing for the time and effort they spend separating recyclables themselves. Free garbage disposal implies that there is too little recycling.

Because of this, governments have instituted a variety of policies to encourage recycling. Generally these policies subsidize recycling—by funding pickups of recyclable material, subsidizing the purchase of recycled commodities, or mandating that governments or government contractors use recycled materials. Some policies also penalize households that don't recycle, but enforcement is usually weak; garbage police are not a common sight in most cities. Governments, then, encourage people to separate recyclables by making it easy to do so, and by appealing to their civic and ecological values.

Recycling subsidies alleviate one kind of problem, but they don't alleviate it completely, and they don't address other kinds of problems. I *should* recycle my plastic soda bottle if and only if the social cost of doing so—the cost of my washing it out, storing it for two weeks separate from the number-two plastics, bringing it out to the curb at the right time, having it picked up, transported, and reprocessed, minus the value of the recycled output—is less than the social cost of having it transported to a landfill and dumped there. Most likely I *will* recycle my soda bottle if and only if I'm a civic-minded kind of guy or I worry about what my neighbors think or I live in irrational fear of the garbage police. Some recycling that should get done won't get done, and some that shouldn't will. In this sense subsidies and campaigns don't alleviate the recycling problem well. But some recycling that should happen does happen, and in that sense the problem is alleviated some—maybe close to as much as is possible.

Recycling policies totally ignore another question, however. They try to alter what I do once I have the soda bottle in my house, but they don't try to affect whether or not I acquire the soda bottle in the first place. If I had to pay the full cost of getting rid of the soda bottle, either by recycling or by regular garbage collection, maybe I would have decided to buy orange juice in a paper container or drink water from the tap instead (or buy a larger container of soda and take advantage of the two-thirds rule). In this case, my forgoing the soda bottle would be a potential Pareto improvement, whether or not I intended to recycle it. The problem again is free disposal, regardless of whether it's called recycling.

A different kind of policy tries to make these potential Pareto improve-

ments. It is called an **advance disposal fee** (ADF). An ADF is a tax on the sale of material that is likely to enter the waste stream (this is like making the sewer fee proportional to the amount of water a household consumes). Because I will eventually dispose of it, I have to pay to take the soda bottle out of the store, whether or not I recycle it. An ADF doesn't affect what I do once I have the bottle; it affects whether I buy the bottle. It does a reasonably good job of alleviating the problem that recycling subsidies don't address, but doesn't do anything about the problem that recycling subsidies do address.

Obviously, putting both kinds of policies together is a good idea. You get to pluck a higher percentage of low-hanging fruit if you pluck from two different orchards instead of just one. The most popular policy that combines the two approaches is called a **deposit/refund.** People who buy soda bottles have to pay a tax, like an ADF, but they get a subsidy in the same amount if the bottle is recycled. "A deposit/refund ensures that the least-cost method of reducing disposal is used, whether it be through source reduction—reducing disposal through less production and consumption—or through recycling" (Palmer, Sigman, and Walls 1997, p. 2).

Not surprisingly, the deposit/refund method reduces the amount of waste going to landfills and incinerators much more cheaply than either ADFs or recycling subsidies. One study estimated it would take a $45-per-ton deposit/refund to reduce regular garbage disposal by 10 percent, and that the same reduction would require an $85-per-ton ADF or a $98-per-ton recycling subsidy (Palmer, Sigman, and Walls 1996).

Deposit/refund systems are not confined to bottles and cans. New Hampshire and Maine tax the sale of new car batteries and give rebates to people who turn in old ones. Oklahoma charges a one-dollar tax for a new tire and offers a fifty-cent rebate for each old tire that is disposed of properly. Quite a few jurisdictions operate deposit/refund systems for pesticide containers. Many European countries operate the equivalent of a deposit/refund system for lubricating oils: they tax virgin oil and subsidize oil recycling. Europe recycles about 65 percent of waste oil, as opposed to 15 percent in the United States.

But deposit/refund is not a panacea. For one thing, if recycling costs money, then the refund is too high. It should be only the difference between the cost of regular disposal and the cost of recycling. Second, the administrative costs of most existing deposit/refund systems are high, possibly high enough to wipe out any other potential Pareto improvements. That is probably why only a few goods have been subject to deposit/refund systems. How you would credit people for recycled scrap paper is a little hard to me to understand. But it's good to remember what has happened with tolls on highways—the type of technology that seemed like science fiction twenty years ago is being installed today.

C. Final Disposal

There are a plethora of options for disposal of solid waste, once it's picked up. Unfortunately, most of the financially cheap options, like dumping in the ocean, which was New York's original method, are environmentally costly.

Incineration and landfills—burning garbage and dumping it in the ground—are the most common disposal options used in developed countries today. With incinerators, the great advantage is that little residue accumulates and sometimes energy can be produced as a by-product; the great danger is air pollution from particulate matter and incomplete combustion products, such as carbon monoxide, nitrogen oxides, chlorinated hydrocarbons, acid gases, and metals. With landfills, the great advantage is simplicity and the great disadvantage is water pollution from leachate—liquid that picks up dissolved or suspended materials as it trickles through the old garbage, and then seeps out of the landfill and pollutes aquifers and surface water—as well as odor and disease. Landfills also contribute to global warming; they emit about 6 percent of the methane that enters the earth's atmosphere every year (U.S. Department of Energy 1993). Both incinerators and landfills attract a steady stream of big, foul trucks, and the farther those trucks have to travel to their ultimate destination, the greater the damage they cause.

Landfills use land, too, but so does everything else. Most land, as we have already seen, is empty. All the world's annual output of solid waste, if put in landfills to a reasonable height of 100 meters, would occupy less than 29 square kilometers. At current generation rates, it would take 100 years for the entire planet's landfill to cover an area the size of Rhode Island (Beede and Bloom 1995, p. 79). There is an opportunity cost to using accessible land for landfill rather than, say, apartments or golf courses, but the price of land represents that opportunity cost.

The disadvantages of both methods can be overcome—it just costs money. Modern incinerators can generally eliminate most of the pollutants through high furnace temperatures, electrostatic precipitators to remove particulate matter and metals, and scrubbers to remove acid gases. Installing an impermeable clay liner in a new landfill area can reduce leachate damage, and covering each day's garbage with a layer of clean dirt can help mitigate problems of odor, insects, disease, and wind-blown trash. Some landfills collect the methane emitted and burn it off to produce energy.

How far these expensive designs and techniques should go in reducing the negative externalities of waste disposal differs according to different circumstances, such as how cheaply reductions in flow to landfills or incinerators can be achieved through source reduction or recycling, and how costly it is to transport waste (it may be cheaper to move the landfill twenty miles away and then not have to worry about waking up neighbors). The better method, too, will differ with circumstances. Landfills, for instance, are a bad idea where the water table is high, and high-temperature incineration is difficult in develop-

ing countries because the waste stream tends to have a high moisture content (wet stuff is hard to burn). Dijkgraaf and Vollebergh (2003) find that in the Netherlands landfills are considerably cheaper socially than incineration (provided the methane is burned to produce energy), but they can't say how far their results can be generalized.

In developed countries, what almost all of the environmentally acceptable disposal methods have in common is a large fixed cost. In the United States, for instance, the average cost per ton for operating a sanitary landfill that handles 2,700 metric tons a day is only about 30 percent of that for a landfill that handles 200 tons a day (Beede and Bloom 1995, p. 21). So, as with sewage, the marginal cost of solid waste disposal is less than the average cost. But since landfills and incinerators almost never try to break even by charging households directly the average cost (or more), overpricing is not as serious an issue with solid waste as it is with sewage.

Because ultimate customers are not charged average cost, the pricing of solid waste disposal does not contribute to sprawl the same way that the pricing of sewage disposal does. But the way solid waste collection costs are generally spread out is not totally innocuous. Properties with more frontage increase the cost of garbage collection but bear no cost for doing so. If my property has a thirty-meter frontage rather than fifteen meters, it takes the garbage truck twice as long to go by it, and yet I will pay the same as for any property with the same value that has a fifteen-meter frontage. This is an incentive for sprawl, but it's probably a small one. Everyone in a town can end up with more frontage and higher garbage collection costs than they want; everyone would be better off if everyone had less frontage and garbage collection cost less.

V. Conclusion

A city requires a source of water, a way of suppressing fires, and means to recycle or get rid of both sewage and solid waste. All of these activities involve externalities for which no Coasian solution is apparent, and so public policy decisions about how they should be priced are probably unavoidable. Whether these services should be delivered by public or private entities is a separate question, and one I have not addressed in this chapter. Currently there are almost as many different arrangements of public, semipublic, and private services as there are cities, and they seem to be becoming more complicated rather than less. The general principles I sketched in Chapter 5 are probably just as applicable to these environmental activities as they were to mass transportation.

All of these activities have some degree of increasing returns to scale, and all of them mitigate problems that do not exist in rural settings. They are also inherently spatial: you can't get a gallon of water or a fire truck by fax, nor can

you e-mail your solid or liquid waste to a virtual dump. You need pipes, roads, fire engines, hydrants, and garbage trucks. These heavy, solid, long-lived assets also imply that history matters a lot, for good or for ill. You can't decide to move New York City tomorrow, or replace it with Phoenix. The existing infrastructure of New York is a great gift that residents receive every day, one that they could not bring with them to Phoenix. But the existing infrastructure can be a curse, too, as we saw in the case of the absurd sewer systems Western cities are saddled with.

None of these systems is perfect. Rich and middle-class people probably use too much water, and poor people probably use too little, especially in developing countries. Sewer services charge too much. People have weak incentives to reduce storm water runoff, to make their property less flammable, to recycle or to throw out less garbage. The pricing structures of most of these systems probably encourage sprawl, although the size of this effect is likely not large. I have outlined some ways that prices and investment policies could change to create potential Pareto improvements, but even with such changes the delivery of these services would still likely be far from optimal.

Questions

1. In spring 2002, rainfall in New Jersey faltered and reservoir levels fell to around 40 percent of capacity. Experts believed that, unless heavier than normal rains arrived, these reservoirs would run out of water in about fifty days. The state government reacted by declaring a state of drought emergency and promulgating a series of regulations prohibiting certain uses of water. For instance, prohibitions included car washing (but not by commercial carwashes), lawn watering (except for newly installed turf), watering of athletic fields (except for no more than 45 minutes at night), power washing (except by firms in the business), and serving water in restaurants. Watering clay tennis courts was permitted.

    If you had been advising the governor of New Jersey at this time, what would you have done? How would you have explained to the public why you chose policies other than those that were actually used?

2. During the 2000 drought, the Maplewood Country Club in New Jersey continued to water its greens. This was legal because it didn't use reservoir water for this purpose; it drew water from its own wells. The country club put up a sign to this effect to prevent complaints from passers-by.

    Was a public policy that permitted the Maplewood Country Club to act this way wise? Why or why not? What would a better policy look like? What

does the Coase theorem say about whether any change in government pol-
icy should have been called for in this situation?

3. In Thirsty City, daily demand for water is given by the function

$$D = 100 - p/2$$

where demand is measured in millions of gallons per day (mgd) and p is
the price of a million gallons of water. If the regular water supply runs out,
bottled water can be imported into Thirsty City endlessly at a cost of $80
per million gallons.

   Thirsty City's reservoir has a capacity of 350 million gallons. Every
Sunday it rains the same way: enough rain falls that 490 million gallons
could be deposited in the reservoir if it didn't spill over. The Sunday rain
storms are absolutely certain. The rate of interest is zero, and no water ever
evaporates.

   What's the optimal price of water on each day of the week in Thirsty
City? (Decide whether the optimal price is different on different days.) How
much water is consumed every day?

4. The Seton Leather plant deposits about a million gallons a day of industrial
sewage into pipes controlled by the Passaic Valley Sewerage Commissioners
(PVSC). The PVSC is a large operation that treats the sewage and bills
Seton and many other customers for the service. According to the latest
PVSC financial reports, its daily costs are:

Debt service	$20,000
Salaries (executives, watchmen, etc.)	20,000
Chemicals to treat sewage	20,000
Electricity for sewage treatment	20,000
Sludge removal	20,000
Total	$100,000

PVSC treats 10 million gallons of sewage a day. PVSC is required by law and
bond covenants to set its price to ensure it raises enough money to cover its
debt service and operating costs. It doesn't discriminate among customers.

a. What does PVSC charge per gallon? What does Seton Leather pay per
   day?

   A treatment equipment salesperson visits Seton Leather and explains
how Seton can buy its own treatment equipment. It costs $10,000,000, and
Seton can borrow at a 0.03 percent daily interest rate. The Seton plant is
right next to the Passaic River, and the treatment equipment is powerful
enough that if Seton installs it, its effluent will be just as clean as the PVSC
effluent and so the environmental authorities will have to give them per-
mission to dump it directly into the river. Of course, they'll have to pay
daily costs of $1,000 for electricity, $1,500 for chemicals, and $600 for re-
moval of sludge.

b. If you were advising the CEO of Seton Leather, what would you do? How would you explain your recommendation to the stockholders?

c. Is this efficient? Why or why not?

d. If you were running the PVSC, what would you do when you heard what Seton Leather was considering? How would you balance the books? How would you explain your approach to the public, including PVSC's other customers?

5. Sewer City is built along one side of a long road. Everyone in Sewer City has to be connected to the sewer plant, which is at the end of the road. Everyone in the world is identical. If they live elsewhere, they have to pay $10,000 for sewer and housing costs. Otherwise, Sewer City is just like every other place in the world.

Sewer City is laid out in a series of 100 lots, each fifty feet wide. The sewer connection has to be at the end of the lot farthest from the sewer plant, for technical reasons. Pipes come only in one standard size, and never fill up.

On each lot, the owner can build as high as she wants, and can fit as many people as she wants to into the resulting house. Taller houses, though, have higher per unit operating costs. In particular, if q people live in a house, the total cost is

$$C(q) = 500q^2.$$

Every resident of a house produces ten units of sewage.

The sewer plant costs $400,000 to build, no matter how many people live in Sewer City. Once the plant is built, processing a unit of sewage costs nothing.

a. What is the optimal price that should be charged for sewage in Sewer City?

b. If the optimal price is charged, what will rent per person be? How many people will live on each lot? How much profit will each owner make?

c. Suppose that the deficit that results from the optimal sewage price is funded through a land tax. How much total land tax will have to be collected? How much from each lot? What is the after-tax profit of each owner? Should Sewer City stay in business?

d. Suppose that Sewer City decides that the sewer operation should pay for itself and therefore adopts average cost pricing. If the population stays the same, how much will each person have to pay for sewage disposal? How much rent will each person have to pay?

e. The lot owners know when they decide to build the houses that the sewers will be subject to average cost pricing. How many people will live on each lot? (Note: fractions are acceptable; if there are multiple equilibria, the one with an increasing marginal cost of building is the one that will be stable.)

f. Will Sewer City be more densely populated with optimal pricing or with average cost pricing?

g. What will after-tax profit per owner be? Is this more or less than it was with optimal pricing?

References

Beede, David N., and David E. Bloom. 1995. "Economics of the Generation and Management of Municipal Solid Waste." National Bureau of Economic Research Working Paper 5116.

Brueckner, Jan. 1981. "Congested Public Goods: The Case of Fire Protection." *Journal of Public Economics* 15(1):45–58.

Burrows, Edwin G., and Mike Wallace. 1999. *Gotham: A History of New York City to 1898.* New York: Oxford University Press.

Citizens Budget Commission. 1993. *The Performance of the New York City Fire Department: Recommendations for Improvement.* New York: The Commission.

Congressional Budget Office. 1985. *Efficient Investments in Wastewater Treatment Plans.* Washington, DC: U.S. Congress.

Crossette, Barbara. 1999. "For the Poor, Water Is Dirty Yet Costly, Experts Find." *New York Times,* August 8.

Dijkgraaf, Elbert, and Herman R. J. Vollebergh. 2003. "Burn or Bury? A Social Cost Comparison of Final Waste Disposal Methods." Nota Di Lavoro 46.2003, Fondazion Eni Enrico Mattei, Milan.

The Economist. 2003. "Priceless." Vol. 368 (July 17), no. 8333:3–14.

Fullerton, Don, and Thomas C. Kinnaman. 1996. "Household Responses to Pricing Garbage by the Bag." *American Economic Review* 86:971–984.

Hall, Darwin C., and W. Michael Hanemann. 1996. "Urban Water Rate Design Based on Marginal Cost." In Hall, ed., *Advances in the Economics of Environmental Resources: Marginal Cost Rate Design and Wholesale Water Markets.* Greenwich, CN: JAI Press. Pp. 95–122.

Kinnaman, Thomas C., and Don Fullerton. 1997. "Garbage and Recycling in Communities with Curbside Recycling and Unit-Based Pricing." National Bureau of Economic Research Working Paper 6021.

Mensah, Joseph, and Herbert A. Whitney. 1991. "Some Third World Environmental Perceptions and Behaviours regarding Urban Waste: A Survey of Techiman, Ghana." *The Canadian Geographer* 35:156–165.

Murphy, Dean E. 2000. "Tap or Bottle? Contamination Fear Puts New Slant on Old Question." *New York Times,* September 17, p. 41.

National Research Council, Academia Nacional de la Investigación Científica, A.C., and Academia Nacional de Ingeniería, A.C. 1995. *Mexico City's Water Supply.* Washington, DC: National Academy Press. Accessed at *http://lanic.utexas.edu.*

Oster, Patrick. 1989. *The Mexicans.* New York: Harper and Row.

Palmer, Karen, Hilary Sigman, and Margaret Walls. 1997. "The Cost of Reducing Municipal Solid Waste." *Journal of Environmental Economics and Management* 33(2):128–150.

Sah, Rohit. 1997. "Sanitation in Developing Countries." PhD diss., Columbia University.

Shay, Ted. 1994. "The Level of Living in Japan, 1885–1938: New Evidence." In John

Komlos, ed., *Stature, Living Standards, and Economic Development: Essays in Anthropometric History.* Chicago: University of Chicago Press. Pp. 173–201.

U.S. Bureau of the Census. 2001. *Statistical Abstract 2001.* Washington, DC: Government Printing Office.

U.S. Department of Energy. 1993. "Emissions of Greenhouse Gases in the U.S.: 1985–1990." DOE-EIA 0573. Washington, DC: Government Printing Office.

Water Supply and Sanitation Collaborative Council. 2004. *Progress Report.* Accessed on May 25, 2004, at *www.wsscc.org.*

Williamson, Jeffrey. 1990. *Coping with City Growth in the British Industrial Revolution.* New York: Cambridge University Press.

World Bank. 1993. *World Development Report: Investing in Health.* Washington, DC: World Bank.

World Health Organization. 2002. World Health Report 2001. Accessed August 1, 2002, at *www.who.int/whr/2001/main/en/pdf/annex3.en.pdf.*

Chapter 10

Education

*C*ities are the greatest education tool ever devised. But harnessing the power of that tool is hard.

For some of the important inputs to education, increasing returns to scale are important. It takes no longer, for instance, to prepare a lecture to be delivered to 100 students than to prepare one for 10. Some cities specialize in education—Ann Arbor today, for instance, or Paris in the Middle Ages—but even in those cities that don't specialize, supplying inputs to education is an important amenity, like sanitary sewers or city buses. Everything else being equal, schools, sewers, and buses are all more productive in cities than in rural areas, and they should be one of the reasons why city living is better.

In this chapter, we'll see how urban economics affects getting kids educated. We will concentrate on kids—early childhood, elementary, and secondary education, not college—because many of the stumbling blocks to getting a good education have been eliminated by the time people reach their late teens. Eighteen-year-olds are reasonably competent in making decisions about their own lives; they understand themselves, to a considerable degree; they can drive, and can protect themselves on sidewalks and on public transportation; and they can easily live apart from their parents or guardians. Five-year-olds have none of these capabilities. That makes finding good systems for educating five-year-olds a whole lot harder than finding good systems for educating eighteen-year-olds.

We'll look first, in section I, at why various people would want someone to be educated—the person herself, her parents, her future children, her peers, and the rest of the world. In other words, we'll enumerate the benefits of education, being careful to distinguish the parties they accrue to and the aspects of education that matter. This last distinction is particularly important: the

---

To get the most out of this chapter, you should be familiar with these concepts: closed-city case, Coase theorem, consumer surplus, diminishing marginal returns, increasing returns to scale, monopoly, open-city case, second-best, and the two-thirds rule. You can find these terms in the Glossary.

benefits that accrue from my learning to speak Mandarin, a Kansas school-child's learning the biblical theory of creation, and an African American child's learning about George Washington Carver's accomplishments are all different.

In section II we'll examine the problems that markets run into in supplying education. Section III then looks at actual policies that attempt to provide education better than a market system would. The array is familiar: rules (compulsory attendance laws), subsidies (child care subsidies and a series of tax preferences), and direct provision (public schools). We will look at the advantages and disadvantages of each approach, and the empirical evidence for their results. Section IV examines a number of reform proposals.

I. Benefits of Education

A. For Children Who Learn

The most basic reason children like learning things is curiosity; humans want to know more, even if that knowledge has no practical application. It feels good to know that Millard Fillmore was the fourteenth president. Education also helps children perceive themselves as closer to being adults, something they find valuable; babies don't know who Millard Fillmore was.

Children also like learning things that have a direct application in their life. Learning the days of the week helps you plan your own time; learning the rules of baseball helps you play the game; learning arithmetic helps you go to the store. Reading and writing open up large new worlds for you. Moral and interpersonal skills also have direct applications: they help you get along with other people (and yourself) and achieve ends that can only be achieved with others' help.

Other school activities are fun, too. School is a place to meet people, make friends, and engage in activities with your friends. There are sports, games, recesses, and social activities. These activities are a major vehicle for learning how to get along with other people.

Education also provides benefits to children that will be realized in the future. Educated people earn more money, and they live longer. In most developed countries today, one more year of education generally increases your wages by an average of 7 to 10 percent, everything else being equal. Most of the evidence indicates that the reason why educated people earn more is that education makes them more productive.

There are many other stories about the possible correlation between education and income—for instance, that more able people get more education, or that employers use education to screen workers even though that education is worthless to them—but the data don't appear to support these stories. If you compare identical (homozygotic) twins who have different educations, or

if you look at people who were forced to complete more schooling because of compulsory attendance laws, you'll still find about the same rate of return to an additional year in school. And so far as we can tell, the rate of return to education no longer varies by race.

The effect of education on health and longevity also appears to be causal. People with more education live longer than less educated people with the same income, even if the additional education was forced by compulsory attendance laws (Lleras-Muney 2002a).

Children, of course, may not fully understand or fully appreciate these benefits when they are very young—or even when they are teenagers. But they're often aware of some glimmer of a connection between success in school and success later in life, although they may not have much idea of what "later in life" feels like.

Schooling, however, is not without drawbacks. Most kids would often rather be somewhere else than in school, doing something else, concentrating on some other problem. Older children could be making money or helping their parents (or raising their own children).

B. For Parents

To the extent that parents value their children's well-being, parents gain everything that children do. And parents' better understanding of the world and probable greater concern for the future often help them make decisions for the child that are better for her (or her future self) than those the child herself would make.

The interests of parents, however, are not always perfectly coincident with those of their children. Parents may want to brag about their children's verifiable accomplishments, for instance. They may also value the time that children spend away from home because it lets them pursue their own work or recreation activities.

Parents may also have concerns, in ways that children do not, about moral and religious development. They may want their children to be good and wholesome, kind and generous. They may also want children with whom they can share religious and cultural traditions (anything from fly-fishing to Orthodox Judaism), and who will be kind to them in their old age. Children themselves, on the other hand, may have little conception of what kind of people they want to be. Of course, to the extent that children value their parents' well-being, any conflicts will be mitigated.

C. For Other Children

Other children also gain, in several different ways, when a child they associate with learns something. One way is when the child who knows teaches the others; you don't lose knowledge by sharing it with other people. This ap-

plies for something the others don't know. Learning something the others already know also helps: it makes conversations with them easier and more productive, and it reduces the frequency of interruptions and digressions in classes.

Learning may also have a demonstration effect on other children. Figuring out who you are and what you can (and cannot) do well is an important part of growing up. If one of your friends, someone you think is quite similar to you, masters a skill, that should lead you to think that you're more likely to succeed, too, if you apply yourself to mastering that skill. Such thinking makes you more likely to try, and trying makes you more likely to succeed.

The demonstration effect may also work with teachers and other people who assist in learning. The more that teachers think students can achieve, the higher the performance they are likely to ask for and the tougher the standards they are likely to impose. (Because of this, some children may lose from the demonstration effect as well.)

Considerable empirical evidence supports the proposition that students learn more when they have better classmates. Zimmer and Toma (2000), for instance, show that students in classes with higher test scores at the beginning of a school year increased their test scores faster during the year than students in classes with lower beginning test scores. This result held in a number of different countries and institutional settings. This effect, called the peer group effect, is stronger for students from weaker backgrounds. Purely in terms of test scores, weak students gain more from being around strong students than strong students lose from being around weak students. Summers and Wolfe (1977); Argys, Rees, and Brewer (1996); Hoxby (2000b); and Hanushek et al. (2001) come to similar conclusions, but the latter two papers don't find different-size peer effects for students from weaker backgrounds. Peer effects are pretty big: improving an average student's class one standard deviation (from, say, a class performing at the median to one at the bottom of the upper third) is the same as accomplishing another quarter of a year's work.

Classmates and friends affect each others' moral and cultural development as well as the development of skills that can be measured on tests (the so-called cognitive skills). Gaviria and Raphael (2001) show that having classmates who drink alcohol, use illicit drugs, attend church, or smoke cigarettes makes a student more likely to engage in each of these activities herself.

D. For Unrelated Adults

The most obvious benefit of education for adults not related to the children in question is that it keeps kids out of mischief during the time they're either in school or studying. Property crime goes down when school is in session (Jacob and Lefgren 2003). Unrelated adults gain this benefit no matter what kids learn (if anything). Similarly, taxpayers gain when children's time in school al-

lows their parents to work more and thereby increase the taxes they pay or reduce the welfare payments they receive.

The other kinds of benefits for unrelated adults are tied more closely to particular kinds of learning. Some benefits come from general moral and cultural learning. The more children believe that crime and lying are wrong and that generosity and trustworthiness are good, the better off everyone else is.

Similarly, the more children know the conventions the rest of us know—how to tell time, what the days of the week are, when right turns are permitted, what language to speak—the more easily and effectively we can all carry out our lives. The rest of us also probably gain if children are more patriotic—if they are more willing to sacrifice, if necessary, to advance or defend the interests of some collective that we're all part of.

Not all adults feel the same way about these benefits. For some adults, it's very important that children learn to respect and care for nature; for others, that they learn the value of sexual purity. Some adults want children to believe SUVs are bad; others, to believe that abortions are bad. Some are better off if more children speak Spanish; others care more about cultivating the next generation of Boston Red Sox fans. For some adults, the collective they want children to be ready to defend is the white race or the state of Texas; for others, it's the United States or the Catholic Church. Because adults have different views on morality and live in different cultures, benefits to one adult can easily be losses to another.

Unrelated adults also benefit from the cognitive skills children acquire and the consequent increased earnings and better health they'll enjoy. Because of income taxes and social welfare programs, some of the benefits of their increased earnings will show up in higher taxes paid or lower welfare payments, and other taxpayers will be better off as a result. Third-party medical insurers, whether governmental or private, will gain some of the benefits of educated adults' better health. When schools increase skills that can be used in legal job markets more than those that can be used for crime, they reduce the threat of crime. People with higher wages also have more to lose from a year of incarceration, and so should be more deterred from crime by the threat of prison sentences. Lochner and Moretti (2004) find that education reduces crime significantly (it's not just that less criminally inclined kids stay in school longer) and that the social benefits from reduced criminality are large—about a quarter of the size of private benefits from higher wages.

Politically, a society in which almost everyone is educated enough to be reasonably well off may work better, too, because rich people will have less to fear from expropriations and the deadweight losses that are likely to accompany them. Education appears to increase voting in the United States, but not in the United Kingdom (Milligan, Moretti, and Oreopoulos 2003). More educated people read newspapers more and support free speech more (Dee 2003). These are all causal effects.

Changes to the supply of workers with various skills may also have broad ramifications in the market. If more people learn how to be computer programmers, for instance, and demand for computer programmers depends on wages, then the wages for computer programmers will have to go down to absorb the new entrants. Some programmers will be worse off, and so will some people whose skills are good substitutes for those of programmers. Other people may be better off because programming is cheaper—producers of computers, for instance, and video-game consumers.

Exactly who wins and who loses varies from case to case (and is generally fairly hard to tell), but there will be external winners and losers. The person who gets educated doesn't necessarily realize all the output gains that her education causes.

An important example of this phenomenon is innovation and entrepreneurship: innovators and entrepreneurs don't often realize all the benefits that they create. Even an inventor with an iron-clad patent can't extract all the consumer surplus an invention generates—people would be willing to pay something for the privilege of being able to buy it at the price at which it's being sold—and so the social return to the inventor's education is greater than the private return. For innovations that can be imitated more easily—ideas for types of TV shows, for instance, or new ways to arrange supermarket floor patterns—the proportion of the gains accruing to the innovator may be substantially smaller.

Lastly, there are external benefits that are political—and again, losers as well as winners can be found among unrelated adults. Kids learn specific propositions that influence their future behavior as voters and as participants in democratic debate—for example, "Communism is bad," "The Holocaust occurred," or "The Mexican-American War was fought to expand slavery." To the extent that government policies are molded, propelled, or constrained by the beliefs of the citizenry, today's learning shapes tomorrow's policies. People who care about tomorrow's government policies therefore care about what kids learn today.

E. For Future Children of the Children in School Now

Children of more educated parents learn more, and so the education that today's children receive helps the next generation of kids. We don't know how big this effect is; for contrasting views, see Black, Devereax, and Salvanes (2003) and Oreopoulos, Page, and Stevens (2003).

II. Why Is Good Education So Hard to Achieve?

Getting kids educated the right way clearly matters, since so many different kinds of people are affected and, in some cases, the effects are profound. Other

goods, like food and clothing, are also very important, and we rely on markets, or markets modified by small amounts of taxes and regulation, to provide them in some reasonably good fashion. In this section, I'll discuss why similar institutions might not work well for certain types of education, and why they might work reasonably well for other types.

A. Increasing Returns to Scale and Monopoly

Many of the inputs to education have increasing returns to scale in many relevant ranges. Time spent teaching is probably the most important of these: one person's listening silently to a lecture or reading a book or using computer software doesn't detract from anyone else's learning through the same medium. Another student may even help, if he asks helpful questions or adds to private discussions on the material. Classroom space is also a source of increasing returns to scale, since it's governed by the two-thirds rule.

But there are limits to increasing returns to educational inputs. Students can be disruptive, ask silly questions, and generally get in each others' way. The more students in a class, the more likely it is that a few of them will be disruptive. Some teaching tasks, moreover, may require individual attention, such as grading papers, answering a student's idiosyncratic questions, getting a child to hold a baseball bat right, writing a letter of recommendation. These tasks may be quite difficult to separate from those that don't require personalized attention. Finally, different students bring different backgrounds, different learning styles, and different goals to the classroom so the lecture or book that's optimal for one student is not likely to be optimal for another. Homogeneity helps.

Of course, we've seen all these problems before, in the discussion of mass transit (Chapter 5). Buses, like classrooms, have some increasing returns to scale, but they're subject to congestion, and they don't usually have origins, destinations, and departure times that are precisely optimal for most of their riders. In Chapter 5 we saw how good pricing schemes for buses may be determined, and good scheduling rules, too, but we also saw that finding an organizational form that would implement these rules was hard. The same is true for schools.

Empirically, the effect of class size on cognitive achievement has been studied extensively, and no researchers have found evidence that reducing class size below the level of fifteen to eighteen students in elementary school improves cognitive achievement. Whether having classes bigger than this hurts achievement is controversial: Krueger (1997) finds that students in smaller classes, especially poor and minority students, do several percentiles better on standardized tests than students in regular-size classes (twenty to twenty-five students), while Hanushek (1997) finds that generally class size has no effect. Krueger uses data from a randomized experiment in Tennessee,

while Hanushek summarizes the results of many studies over the years. More strikingly, reports by Hoxby (2000a) and Angrist and Lavy (1999) both work with variation in class size induced by administrative rules and rounding problems, and they reach different conclusions: Angrist and Lavy, looking at Israel, do find class-size effects, but Hoxby finds none, based on her data from Connecticut. Card and Krueger (1992) find that men who were in smaller classes in school in the first half of the twentieth century earned more than those who were in larger classes, everything else being equal, but the range of class sizes was much greater then than what is common now.

Why such confusing and contradictory results about a relationship that seems to be common sense? Most explanations hinge on the difficulty of holding constant or even observing all of the inputs to learning. Hoxby and Hanushek, for instance, emphasize that schools have limited incentive to act efficiently, and so smaller class sizes may just allow teachers and administrators to live easier lives or concentrate on other goals (this is the explanation we will find in Chapter 16 for the weak relationship between numbers of police and index crime). Within schools, stronger teachers may be assigned to larger classes. Across schools, less disruptive students in bigger classes can function just as well as more disruptive students in smaller classes. (Lazear 2001 emphasizes these last two explanations.) Students in larger classes might decide to exert less effort on learning, since they can learn with less effort, and parents, too, may decide that they won't work as hard on helping their kids with homework.

To use the bus analogy, then, there's no congestion with fewer than fifteen to eighteen riders. Above that, congestion is probably slight (not everyone can sit by a window) and its existence is controversial. When you get above maybe forty or fifty riders, congestion gets to be a real problem. These results apply only to regular classes, to elementary schools, to the teaching techniques of the late twentieth century, to cognitive achievement, and to the degree of homogeneity that was the norm in schools at that time. Gym classes can handle more students, piano classes fewer. Younger students, many of whom have not mastered self-control or the art of putting on your own winter coat, probably require smaller classes, while older students can often manage in much larger classes—college students routinely attend lectures with hundreds of students. New ways of teaching based on information technology may radically change these relationships, but they haven't done so yet. Thus it's pretty certain that there are increasing returns to scale for classroom teaching, which is an important part of learning.

As with buses, the implications of increasing returns to scale depend on population density. Small children are expensive to transport: they don't drive, and, if they're young enough, they need adults to accompany them. Thus it's expensive to gather a group of small children in the same place every day if they don't all live close by. Older children, who can walk faster and by them-

selves or ride school buses or bikes, can be gathered in somewhat larger numbers, but it still takes a large populated area to fill a normal elementary school class in most metropolitan regions.[1]

The costs of gathering enough children to take advantage of increasing returns to scale are mitigated if households choose where they live based on where their children attend school, and to some extent they do. But choosing a house on this principle is costly, too: it means the household must give up the advantages (less travel time to work, for instance) that living some place farther from the school would afford them. Elementary school students (unlike many college students) live with family members who have other destinations to worry about.

The problems of travel time and increasing returns to scale are exacerbated by differences in the type of schooling children and parents are looking for. Only a small proportion of students want to learn to speak German or play the accordion, and so gathering enough students for a class of fifteen in either subject can require a large area, unless population density is high. The same is true for those desiring a Muslim or Montessori education, or for those students for whom special education facilities or more advanced classes are desirable. When population density is low, it's hard to find a bus that departs at a time that's convenient for you.

Thus we have another advantage of city life: it makes it easier to produce cheap education that matches more closely what people want (just as it's easier to have a better bus system in a city).

Concomitant with this advantage is the opportunity for escaping monopolies. If a town can support only one school, that school is going to have monopoly power that it can use to charge high prices or produce low quality or both. (It couldn't charge incredibly high prices, because that would drive people to leave or set up home schools, but it can charge more than marginal cost without inducing entry by a competitor; monopolists are powerful, not omnipotent.) In contrast, in a large city—say, Manhattan—enough schools probably could reach a decent size that real competition would be possible.

In most of the United States, however, and for many types of schools, population densities may be low enough that monopoly would be a serious problem with unregulated schools. This is probably not the case for the youngest and oldest students: college students can reduce transportation costs by moving close to campus, and very young children require too much personal attention for increasing returns to scale to have much power. But for elementary and high school students in many places, monopoly might be a problem.

1. About one person in 68 was ten years old in 2000, and so if you want 25 ten-year-olds for a class, you need a population of 1,700. No state has, on average, more than 1,000 people per square mile, and many suburbs are no more dense than 3,000 to 4,000 people per square mile; thus you need to draw from an area of about half a square mile to get enough kids to fill a single classroom.

Monopolies can be treated in many ways—rate regulation, ownership by customers, benign neglect—but each of these ways has costs. Monopoly is another reason why we can't expect education markets to work flawlessly.

B. Finding the Beneficiaries

We saw in section I that many unrelated adults gain (or lose) from different aspects of children's education, and so for children and their parents to have the right incentives regarding education, these aspects need to be subsidized (or taxed, if unrelated adults lose). The problem is finding someone to pay the subsidies.

Coasian bargaining won't work, because each child's education affects many adults and children who haven't been born yet. I may be a tiny bit better off if a youngster in Los Angeles is more patriotic or law abiding, but so are millions of other people, some of them not born yet, who will be safer when that youngster grows up. The youngster's patriotism and law-abidingness are a public good. All of the millions of us would rather get the others to pay for it, and so we will not rationally voluntarily contribute toward it nor even honestly report our benefits, if doing so is going to cost us money. Therefore, government intervention in subsidizing the types of education that benefit large numbers of people—teaching social conventions as well as morality—is likely to be a potential Pareto improvement.

C. Verifying the Output

Physically, learning probably consists of the rearrangement of synaptic connections in our brains. We don't know precisely what sorts of rearrangements correspond to what sorts of learning, and it's not clear if we'll ever know. Observing synaptic rearrangement is often difficult to do without sacrificing the subject. Thus verifying whether someone has learned something is extremely difficult. This is especially true for such attributes as "good moral character," "ability to get along with people," and "deep understanding of microbiology"—attributes that people often care more about than knowing the capital of Bolivia.

As a result, contracts that make payment contingent on someone's learning something are almost impossible to enforce and so are almost never written. This isn't unusual—for instance, employers can't write contracts about "hard work" either, but they still often manage to elicit it. For education, substitute contracts can be written, but since they aren't about what people really want to agree on, they won't be as effective as contracts that exhaust all potential Pareto improvements.

Tests can sometimes be useful proxies for some kinds of learning, but a person who wants to learn something would be foolish to offer a contract to a teacher in which the teacher's payment depended on how well the learner per-

formed on a test that had no other implications for the learner. Realizing that the learner could deliberately flunk the test and owe him nothing, no teacher would accept a contract like that. Similarly, while a student may embark on a course of study in order to increase her future earnings, tying payment to teachers to those future earnings would distort her career choices, inclining her more to interesting or rewarding jobs with low pay instead of high-paying jobs with terrible working conditions.

Third parties—parents or unrelated adults—who care about someone else's education have the same sort of problem. They can hire teachers and tutors, but they usually can't observe well what those teachers and tutors are doing; this is especially true for a citizen with a full-time job who cares about the progress of all of the town's—or the nation's—kids. Even if a third party does observe what's going on in a classroom, she may not be able to evaluate it intelligently; perhaps she doesn't know the material being taught, or understand the teacher's pedagogic technique, or comprehend the allusions to TV shows she's never watched. Third-party adults can't observe what teachers and tutors produce, either, even if they have informative scores on good tests, because success or failure on a test could be attributable to the student's lack of effort or capability as much as to anything the teacher might do or fail to do.

What's true of education by teachers is also true of education by parents. Unrelated adults have a huge stake in what parents teach their children but almost no way of rewarding them or sanctioning them for it.

D. Making Children Matter

Monopoly, large numbers of beneficiaries, and problems of observability and verification are not unique to education, although the severity and conjunction of the difficulties they impose would by themselves probably make education an area where government involvement was fairly common, like fire suppression or police, even if nothing else about it were special. What makes education special is the crucial role of children.

Children are the ultimate actors in their education. It is they who either learn or don't. Children have to decide whether to pay attention, what to pay attention to, how hard to try, how much to practice, what to care about, whom to believe. Parents help, teachers help, computers help, books help—but the only thing that matters is what happens in the child's own brain and central nervous system.

Children, though, are lousy decision makers, despite all the decisions they have to make. Part of the job of educators is to make them into better decision makers, but many crucial decisions have to be made before this job is even partially complete. Language acquisition after puberty, for instance, is much more difficult, and learning a language without an accent almost impossible; yet few ten-year-olds understand enough about the world to make an

informed decision about which languages are likely to be most useful to them when they are adults (or which languages' literature would lose most in translation).

Because nature and society give kids too many decisions to make before they're ready to make them, human societies take away their right to make many of those decisions (but not all of them). Children are also small and weak, and so adults can push them around and force them to do things. Thus children lack many rights that adults take for granted: the right to choose their religion, for instance, to choose their own food or their own schools, to play bingo or gamble at a casino, to drink alcohol, to decide about their own medical care, to sue, to borrow, and to enter into contracts.

Not being able to enter into contracts is the key loss; it makes the conditions for the Coase theorem fail. Only through some form of altruism can the gains and losses of children matter for decisions that others make. If children lacked other rights but could still enter into contracts, they would be worse off but they could influence events by offering bribes, and so all potential Pareto improvements would be realized. Because kids can't enter into contracts, Pareto improvements will go unrealized.

Being unable to borrow (while only a special variety of the inability to write contracts) also has important implications. Many of the inputs to a good education cost money: high-quality prenatal care; nutritious food; an environment free of toxins, such as lead; a peaceful and secure existence in the first years of life (high levels of stress hormones in the blood in these years produce changes in brain chemistry that make paying attention and learning difficult); good medical care; stimulating adults present for many hours every day; toys and games; decent housing, with good opportunities for sleep and self-expression; knowledgeable and concerned teachers; books, computers, and software. If kids could borrow to finance these investments, in many cases they would—and the lenders would be rewarded handsomely. This would be a Pareto improvement. But if kids can't borrow, it won't happen.

Someone has to make the decisions children don't have a legal right to make, or can't competently make, and in most societies those decisions are divided up between parents (and other guardians) and the government. Plausible arguments can be made that each of these decision makers will sometimes act in a child's best interest, but no one can argue that either entity will always do so. Even the most loving parents may have strongly held ideas about what sort of person they want their child to become—a secular humanist or a Muslim or a firefighter or a mensch—and those ideas can in no way be said to be the child's.

We shouldn't think of children as their parent's property, and we can't think that parents should make untrammeled decisions about their children's future, even if those decisions affect no one outside the family. Nor can

we think that children should make untrammeled decisions about their own future.

III. Policies

The preceding two sections offer plenty of reasons why governments should get involved in the education business—monopoly, externalities, observability, and the incompetence, both legal and substantive, of key decision makers—or at least reasons why there is no prima facie case against government involvement. In this section, we will look at some government education policies, see who gains and loses from them, and ask whether they might be improved (perhaps by being scrapped). As with other activities involving externalties, education policies can be divided into those that use quantity rules, those that use subsidies or taxes, and those that involve direct government provision. I'll follow the same organizational scheme in this section.

A. Rules

Rules are a good idea when the marginal external benefits or costs vary greatly with the amount of activity. This was one of our conclusions in Chapter 8. A number of inputs to education seem to work like this, and so rules are prominent in some areas.

Parents, for instance, are required to provide their children with a certain minimum level of food and clothing. The parents are the decision makers, and the marginal benefits to children (and those indirectly affected by the children's development) from avoiding malnutrition and cold and shame are great, much greater than the benefits of a third hamburger at lunch or a sixth pair of sneakers.

More controversial are the regulations on mandatory school attendance. Children between certain ages (generally five and sixteen, but it varies from state to state) must attend school a certain number of days a year (generally about 180) for a certain number of hours a day, whether or not they want to go to school or their parents want them to. Not only must they sit inside a particular type of building, but they also must follow a curriculum that includes certain mandatory items (these vary significantly). (Homeschooling exceptions allow some flexibility on the type of building kids have to sit in but still mandate a curriculum and a rough idea of time.)

Three separate arguments can be advanced in favor of these rules. The first is just a repetition of the child protection argument: the benefits to the child of exposure to this minimum curriculum are great, whether the parent recognizes it or not—much more than the marginal benefits of additional education. (In the one dimension in which this argument has been tested, it has

received some support: so far as earnings later in life are concerned, the rate of return to early years of schooling appears to be greater than the rate of return to later years, according to Psacharopoulos 1994.)

The second argument makes the same sort of claim about the benefits that unrelated adults gain from a child's attendance at school. In the early grades a child learns patriotism and morality, the basic conventions of our society, and enough job-related skills to survive and raise a family without public assistance. These also are years when kids would otherwise have to be watched to keep them from getting into trouble. Additional years of training bring external benefits, but none so large as those of earlier years: the citizenship component is considerably less, subsequent earnings are subject to taxes but are above the level that would necessitate public assistance, and kids are mature enough that they don't have to be watched.

The final argument, more popular with philosophers than economists, is that children need a certain amount of schooling to assure that they have their "rights to an open future." The idea is that adults have certain "autonomy rights"—the right to form and express their own political opinions, for instance, and the right to choose their own religion (or none at all). Children should have these rights, too, but are precluded by their natural developmental limitations from exercising them. It is in the interest of future adults—the adults that today's children will become—that they be allowed to enter into adulthood with as many options still open to them as possible. A certain amount of schooling is needed to keep those options open. If you decide when you're eighteen that you want to be a surgeon or you want to study the Koran but you've never learned to read, it's probably too late. (See Feinberg 1992 for a more detailed exposition of this argument.)

This third argument for compulsory schooling can be seen in economic terms if we note that a child just becoming an adult might reasonably be expected to be willing to pay a lot of money, if she could borrow it, to have more options open to her. Her parents might be unhappy about her having some of those options, but since those options are "normally" welcomed by "normal" parents in our society, and since it is their children's lives that are involved, rather than their own, the amount of money that parents would need to assuage their unhappiness is probably less than what their children would be willing to pay, if they could borrow, to have those options. Thus, requiring the child to attend school in order to preserve her options is a potential Pareto improvement.

All three of these arguments rely on a kind of rough-and-ready cost-benefit analysis: they hold that certain kinds of education are valuable (to the recipient, or to unrelated adults, or to the potential adult the child might become), and that the cost is rather low. Each specifies certain types of schooling as mandatory, but not others. Very early childhood schooling is not worth mandating because considerable individualized attention is required at that

age, and very young children are incapable of the more valuable types of learning; late-teenage schooling is not worth mandating, because older teenagers can earn more when they're not in school, they don't need to be watched, they've mainly learned the valuable things by this time, and they're capable of making decisions for themselves.

We should expect the limits of compulsory schooling to change as these costs and benefits change. Higher wages for mothers, for instance, which make minding very young children in school on net less costly, and scientific findings about the importance of early childhood experiences to children's later development strengthen arguments for reducing the age at which mandatory schooling should begin. The rapid fall in recent years in the relative (and absolute) wages of high school dropouts argues for raising the school-leaving age or for increasing the length of the school day or the school year. (Between 1979 and 1996, the earnings advantage of high school graduates over high school dropouts rose from 19 percent to 40 percent; see Krueger 1998, p. 38.)

Compulsory-schooling laws rely on an implicit, hypothetical, and often counterfactual weighing of costs and benefits (asking, for example, how much a child would pay, ten years from now, to have the option to become a surgeon, if he could borrow, even though he won't be able to borrow and by that time will have neither the inclination nor the skills to become one). Hence these laws must either specify only those cases where benefits hugely exceed costs for almost everyone—and so fail to mandate many other potential Pareto improvements—or risk great controversy. American laws have generally taken the former approach, giving up potential Pareto improvements in an effort to avoid controversy, but they haven't wholly succeeded in avoiding controversy. The controversies have mainly involved religious groups, the Amish in particular, for whom extensive education in secular matters threatens the fabric of the community they are trying to maintain: children in Amish communities are raised to be simple, pious farmers, and the idea of raising them to be anything else is extremely repugnant to their parents, just as classes in child pornography would be extremely repugnant to most parents. These difficulties, of course, are a general feature of all rules: they mandate some instances of an activity that are not potential Pareto improvements, and they fail to mandate other instances that are.

Child labor laws—laws restricting the hours and conditions under which children can work, and forbidding work altogether in many circumstances—are often seen as a natural concomitant of compulsory education laws. If parents find it more difficult to send a child to work, they are more likely to send him to school. In other words, child labor laws cut the costs to parents of complying with compulsory education laws. Work itself, particularly hazardous work or bonded labor, may also be detrimental to children's development, and so child labor laws may be desirable in themselves, not just as a means of promoting school attendance.

But the case for child labor laws, especially of the more general and draconian kind, has some weaknesses. As a practical matter, compulsory attendance at school is much easier to enforce than compulsory nonattendance at work. As Basu (1999, p. 1090) puts it: "A child's presence in school is easier to monitor (and thus ensure) than a child's absence from work." School and moderate amounts of work, moreover, are not incompatible. For poor families, the income a child brings in contributes to his own well-being—in better food and clothing, in more school supplies—as well as to that of other family members. The option to put children to work is also a safety net for poor families, something to fall back on when times are rough (Dehejia and Gatti [2002] find that countries with better financial systems have less child labor, all else being equal). Earning money may also increase a child's influence in family decisions, even if only by giving them a credible threat of running away. Basu (1999, p. 1115) notes that "there are worse things for a child than having to work."

Because both compulsory attendance laws and child labor laws require parents to do things most of them want to do anyway, and because few resources have been devoted to enforcing them, there is little strong evidence that they actually make much of a difference. Lleras-Muney (2002b) finds that compulsory attendance and child labor laws modestly raised school attendance among whites in the early twentieth century in the United States; they had no effect on African Americans. Scholliers (1995, p. 208) studied child labor laws in Ghent, Belgium, and concludes that, by 1850, "the number of [employed] children under 12 had diminished substantially . . . without any legislative intervention." Moehling (1998) studied the employment of children in U.S. manufacturing between 1880 and 1910, a period that saw both considerable activism against child labor and a large drop in the employment of children. By comparing states that enacted laws at different times, she finds that legislation had little effect; the same reduction would have occurred even if no laws had been passed (Basu [1999, pp. 1090–1091], however, questions her methods). On the other hand, Brown, Christiansen, and Philips (1992) find a modest effect of legislation in explaining the decline of child labor in the U.S. fruit and vegetable canning industry between 1880 and 1920; and Bolin-Hort (1989) finds that laws played a big role in the decline of child labor in the cotton mills in Manchester, England.

B. Subsidies

Most laws on what parents do to educate their children thus seem to require less of most parents than they would do on their own. And since we've seen that what parents would do on their own is probably less than what they should do, a case can be made for subsidizing education—using a carrot as well as a stick. Subsidies also have the advantage of allowing individual varia-

tion—only those parents with the lowest costs of providing additional educa-tion will do so. When the government pays parents to send their kids to school instead of forcing them to go, it doesn't have to contend with special groups like the Amish (although it does have to contend with taxpayers).

The advantages of subsidizing education are apparent in Latin America, where Brazil and several other countries pay poor families to send their kids to school instead of trying to prohibit child labor. Israel has experimented, with some success, with paying rewards to students who complete the Bagrut, which is roughly equivalent to a high school degree (Angrist and Lavy 2004). And schools in some U.S. cities, such as Chicago, give in-kind prizes, like tick-ets to sporting events, to high school students who have good attendance.

In the United States, most subsidies for education come from the federal government and are concentrated on students who are not covered by com-pulsory education laws—very young children and those in college. States, cit-ies, and the federal government (largely through the deductibility of property tax payments and the exemption of imputed income from houses) also pay for schools that governments run, but we'll consider these in the next section on direct provision. I wish to concentrate here on the subsidies for very young children; college education does not fall within the scope of this chapter. No-tice, though, that I've identified both early childhood schooling and college as areas where monopoly is less likely to be a problem. Subsidizing purchases from a monopolist has obvious drawbacks.

The federal government gives subsidies to families for early childhood education in two chief ways: by granting tax credits of 30 percent of the cost of child care, and by giving out certificates that can be used to pay for a por-tion of child care. The tax credits are open-ended—any taxpayer can claim them—and are claimed mainly by middle- and upper-income households. The certificates are available only to poor families seeking child care in facili-ties that meet state standards. The proportion of child care expenses that the certificate pays for depends on the family's income. More families are eligible for the certificates than receive them; the number who get them is limited by congressional appropriation. The federal government also gives some direct aid to child care centers, mainly in the form of free food.

Because increasing returns to scale are unimportant in the child care market—a large fraction of children are cared for in unregulated homes that serve three or fewer children—this subsidy arrangement gives parents many options in arranging care for their young children, and great freedom in de-ciding how much to spend. Providers that parents don't like can easily be forced out of business. Parents can walk away from any provider who charges too much or gives too little, as they see it. Parents who don't want care for their children don't have to place them anywhere.

Concomitant with great parental power in this market has been great un-happiness, especially among education experts, with how parents have used

this power. Most child care appears to do far too little to develop children's cognitive skills; many child care workers appear to be underskilled and underpaid. Rigorous studies have shown that well-trained teachers and staff working with small numbers of children in a well-structured environment can make a difference in children's cognitive development (for a review, see Currie 2001), but many early childhood facilities fail to do so.

One important reason why early childhood centers don't provide the kind of quality that would make education experts happy is that parents are more interested in other kinds of quality. According to Kisker and Maynard (1995, p. 130), "Parents appear to judge the quality of child care according to (1) whether it offers a safe and healthy environment—many parents express considerable concern about potential child abuse; (2) whether the environment promotes learning—a concern that is especially prevalent for older children; and (3) convenience, including location within a 10 to 15 minute radius of home or work and hours that mesh with the mother's work schedule." Sonnenstein (1991) asked a group of mothers to rate their most recent child care arrangements on each of fourteen characteristics, and also to indicate their overall level of satisfaction. The mothers said that they cared about the child-staff ratio and specialized training of the providers, but when it came to predicting their overall level of satisfaction, these attributes didn't matter. The most significant predictors of satisfaction were the convenience of the hours and the location, and the number of workdays the mother missed because of child care foul-ups. Walker (1992) found that child care providers with better child-staff ratios and better-trained staff weren't able to charge higher fees.

Why don't parents spend more money on the developmental aspects of child care? One reason, obviously, is that parents are human and some worry more about themselves and their convenience than their children's future grades in school. But a deeper reason may be that parents have little good information on which to judge the developmental contributions of the child care they use, and even less to judge the developmental contributions that could be made by the child care they don't use. Parents spend little time at child care centers watching what's going on, and their children are too young to provide reliable reports. They spend no time at child care facilities where their children aren't enrolled. They aren't experts in early childhood education and so don't know what skills to expect their children to develop at what ages, or how to test for those skills. They have few other children, and so have little experience in child-rearing for comparison. Healthy children differ from one another immensely, anyway. And since most child care operations are small and many go out of business after a little while, few have an incentive to advertise or develop a strong reputation (Walker 1991, p. 66, found that, for all types of child care, most users learned about their providers through word of mouth, and no more than a quarter learned about them through advertising or referrals from formal sources).

When you don't know about some aspect of a service, that aspect can't figure heavily in the decision you make. Parents who can see for themselves how convenient and reliable a child care provider is will rationally pay much more attention to those aspects of the service than they will to its developmental contribution to their child, even if they care more about the developmental contribution.

This is not to say that parents don't care a lot about the development of their young children's cognitive skills. Almost all do. Nor are parents totally ignorant of the contributions child care can make to cognitive development. State licensing procedures help some in identifying good providers, although these are mainly concerned with physical safety and sometimes with simple measures such as minimal staff training and maximal student-teacher ratios. Voluntary, nongovernmental accreditation procedures and the attempts of a few national chains to establish reputations also help to improve early childhood education.

But the way subsidies for early childhood education work, they don't give parents enough of an incentive to seek the best provider for their child: the subsidy is the same whether they use it for their own convenience or for furthering their child's cognitive development. The benefits that *other* people gain from their child's cognitive development—the child's future classmates, because of peer-group effects; the general public, because of their future better citizenship and greater taxpaying; and the children themselves— nowhere enter into the parents' decision. Parents' incentives would be more closely aligned with the true social implications of their actions if, for instance, the subsidies depended in part on measures of their child's cognitive development on entering regular schooling. Whether such measurement is feasible is an open question.

C. Direct Provision

The child care situation shows that subsidies may not work well when the activity that should be subsidized is difficult to observe or to verify. Increasing returns to scale and the possibility of monopoly add further difficulties for subsidy schemes. Under these conditions, direct provision by the government is an attractive alternative to subsidization: the government can take charge of the activity itself without having to write a contract that a judge can enforce, and it can operate on a large scale without worrying whether that scale will allow private parties to charge the wrong prices.

Libraries and schools are the two most important educational inputs that most governments provide, but since the expenditures for schools are much greater, we'll concentrate on them.

The argument for publicly provided elementary and secondary education is that the socially optimal sort of schooling for kids between five and

eighteen is different from the sort that parents would buy on their own. Writing subsidy contracts that accurately specified the type of schooling that would be paid for would be too difficult. And of course public provision doesn't imply that every child has to attend public schools. Just as you can choose to buy bottled water instead of using the tap or to go to a bookstore even when a public library is just around the corner, you can pay for private schooling instead of using the public schools, which are free. Compulsory education laws require only that your child attend *some* school. The public schools are the ones that are cheapest.

In fact, U.S. public schools almost always have the same price: zero. Since the short-run marginal cost of schooling is likely to vary greatly (depending, for instance, on congestion and the students' disruptiveness), as are the marginal external benefits (depending on peer group effects, for example, or the subject being taught), this means that the price is almost always wrong. Capital market imperfections matter, too, in determining the optimal price: everything else being equal, a student from a poor family who is not likely to have cash available should be subsidized more than a student from a rich family (this is a second-best argument). Whether optimal prices would be positive or negative (you'd get paid to go to school, as in Brazil) is something I don't know, because no one has ever tried to charge or even compute optimal prices. But they're not always all zero.

Uniform pricing of public elementary and secondary schools creates several perverse incentives for parents. First, it induces them to devote less effort than they should to educating their children while they're away from school. Kids learn both at school and at home, and the more they learn at home the better they perform in school as classmates and mentors. But since uniform pricing keeps parents from being rewarded for helping their children's classmates (or even for helping their own children), they do too little teaching at home.

Second, uniform pricing makes all parents want to send their children to precisely those schools where they will be most costly to their classmates by having the most negative peer-group effect. All else being equal, you want to be in that school where your classmates are the best students, and that school is precisely the one in which you are the biggest impediment to their performance.

Finally, just as with cars and buses, uniform pricing means that parents have insufficient incentives to avoid creating and participating in classroom congestion. My entering a crowded class hurts everyone else in it, even without any peer effects, and with uniform pricing the losses I cause are of no concern to me.

The poor incentives that uniform pricing creates aren't the end of the story. Those who believe they might be hurt by parents' choosing schools under the influence of these incentives react by protecting themselves in ways

that make more problems (in part because they can't react by using better pricing). To understand the form this reaction takes, we need to begin by understanding what is known as a **school sending district.** A school sending district is a connected geographic area (you can walk from any place in it to any other place in it without leaving it) in which all the children of the same age who go to public school go to the same school. Since transporting kids is costly, it makes sense that every school should have its sending district and only students from that district should attend the school. With optimal (that is, nonuniform) pricing, people would sort themselves into districts voluntarily: they would eschew crowded schools because of the harm they did to other students, and they would go to schools with better students only if the private benefits to themselves outweighed the peer-effect losses to those students, and they wouldn't make longer trips to school than they had to.

With uniform pricing, the current residents of school sending districts with good schools have an incentive to turn those districts into fortresses to prevent others from imposing costs on them. The first step in creating a fortress is a residency requirement: students can't go to the school unless they live in the district. Relative differences among schools then show up in relative rents and housing prices among school sending districts; Black (1999) documents this phenomenon empirically in Massachusetts, Figlio and Lucas (2000) in Florida, Downes and Zabel (2002) in Chicago. But residency requirements alone aren't sufficient protection: they can be circumvented by families willing to live very densely in order to send their kids to good schools. Thus, to turn good school districts into true fortresses, residency requirements have to be supplemented by restrictive zoning codes.

In a world where school sending districts are fortresses, the richest districts will have the best schools. There are several reasons for this. Better-educated people tend to be both richer and better teachers of their children, and so wealth and well-prepared classmates tend to occur together. Certain cognitive skills are somewhat heritable across generations, and those cognitive skills are also (weakly) correlated with income. Richer parents consume more of everything, including in-home preparation of their children. For all these reasons, richer communities produce better classmates, on average. Richer communities, moreover, are more willing to spend money collectively on every good thing, including schools, and so their schools are likely to have more care lavished on them. Richer parents and taxpayers, moreover, are more likely to have the skills to monitor that spending better, to make sure it's being used to foster education rather than to hire administrators' otherwise unemployable cousins.

This outcome is inefficient. Some poor students, if they or their parents could borrow against their future earnings, would be willing to pay the students and families in richer districts for the opportunity to attend those schools, and the richer students and their families would be willing to accept

what the poor students would pay. But this deal can't happen. Potential Pareto improvements are being missed. Unrelated adults, too, might be willing to pay to improve the education of poor kids at the expense of that of rich kids, since many of the external benefits of education are realized when a lot of people reach moderate levels of proficiency, rather than when a few reach high levels.

Ironically, an idea that was supposed to promote opportunity for poor kids—uniformly free pricing of elementary and secondary schools—turns out to be a prescription for privilege: the way you get a good education is to have parents with enough money and enough concern for you that they buy a house in a good school district. Uniform pricing sacrifices efficiency but produces no equity.

The problems with direct provision of schooling have not gone unnoticed. The predominant response has been to spend more money—generally raised from statewide taxes—on schools in poor districts. This response makes a lot of sense. The external benefits of education, particularly at basic levels, affect adults in a very large region, not just those in the immediate vicinity of a school. So the return on basic levels of achievement in poor districts is great, and is something that adults in a wide region should be willing to contribute toward.

The results of increased state funding for schools in poor districts, however, have not always pleased the advocates of such measures. Test scores have not increased markedly, and parents from these districts still resort to elaborate ruses to get their children into schools they perceive as better.

There are some simple economic explanations for these disappointing results. Some poor districts may spend less of their own money than they would have in the absence of large amounts of state aid (or their towns may charge the schools for services, like garbage collection, they wouldn't have otherwise charged for). Parents may teach their children less at home when their learning opportunities are better at school (see Houtenville 1996 for a discussion of this relationship). **Diminishing marginal returns** to a particular input may be part of the story, too: learning requires a combination of student effort, parental assistance, peer group example and pressure, and the things that schools do. Improving one of these inputs without improving the others is not going to get you far. Making meals takes cooks and raw food, but if you don't have any more raw food, adding more cooks won't help much, after a certain point.

An alternative explanation that has received much more attention is that many schools just don't do their job. This, too, is a simple economic story: direct provision means that the people who run schools have only weak incentives to educate children. They don't get more money if they educate better, or less money if they educate worse, and so since educating better is harder, we shouldn't be surprised that they educate worse. Since private schools are a lot more expensive than public schools and are unavailable in some regions or to

parents of certain religious faiths, public schools have captive customers who can't demonstrate their displeasure by withdrawing their children.

This reasoning applies to schools in both rich districts and poor districts—indeed, it applies to direct provision generally: to police protection, public libraries, and garbage collection, as well as to schools. Under some conditions, however, public schools have less monopoly power, and when they have less monopoly power you would expect them to perform better. Indeed, Hoxby (1994, 1998) finds that more competition, either from private schools or from other public schools, induces better performance: students do better on tests. (Schools in smaller administrative districts face more competition from other public schools because parents can move out more easily.) Competition could work either by giving parents a more credible threat of pulling their children out of the school they're in, or by giving them an opportunity to compare schools and demand improvements.

Schools that poor kids attend are likely to enjoy less of this healthy kind of competition than schools that rich kids attend. Private schools may be totally unaffordable for poor parents. Moving to a better public-school sending district may also be tough, and not only for reasons of zoning. Moving costs money, and a better housing location for school purposes may be worse for jobs and adult friendships. Project-based housing subsidies, too, like public housing, also make it harder for poor families to move and easier for school officials to enjoy a soft and easy life.

IV. School Reforms

The current system of direct provision thus seems to be preventing two potential Pareto improvements from occurring: some students could gain from access to better classmates and better schools, and some students could gain if school employees worked more assiduously and more creatively. I'll call the former the fortress problem, the latter, the school incentive problem.

Many reforms have been proposed recently to address these problems, and in this section I will analyze some of these proposals (as well as a few ideas I don't think anyone has proposed). I'll treat the problems as distinct, although often in general discourse the notion seems to be that solving the school incentive problem is equivalent to solving the fortress problem. This equivalence appears in the idea that somehow incentives could be arranged so powerfully that the schools that poor kids attend would be staffed by teachers and principals so dedicated, energetic, imaginative, and masterful that all other deficiencies of classmates, parents, and history would vanish like Silly Putty under a steamroller, and there would be no reason for poor kids to go to rich kids' schools. Even if such a magic elixir of incentives could be concocted,

sending some poor kids to rich kids' schools might still be a cheaper way of improving their education. How kids get sorted among schools would still matter.

A. Testing

Proposals for more use of standardized testing are directed at the school incentive problem. These proposals (some of which were implemented nationally as part of the Bush administration's education package of 2001) require schools to administer nationally designed tests to all students in certain grades. The distribution of test scores is publicized, both for the school as a whole and for ethnic groups within the school. If the results aren't good enough, or aren't improving fast enough, the school can be reorganized in ways that hurt the top administrators financially, or at least disgrace them and hurt their chances for career advancement.

Even if the test results produce no formal action, they may provoke parents and other interested adults to lobby for improvements or to leave the district. Levy and Murnane (1998, p. 119), for instance, tell the story of Zavala Elementary School in East Austin, Texas, a school whose students come mainly from low-income families. Zavala students were in fact developing little in the way of cognitive skills, but "few parents were aware of their children's low skill levels because their children received grades of A and B. Teachers gave high grades for poor work because they thought the children were not capable of better work." When Texas instituted a standardized testing program, parents found out for the first time where their children really stood. "When parents learned of their children's poor performance, they were outraged. Teachers were stunned; parents at Zavala had never seriously questioned the quality of their children's education." This outrage touched off a process that led to a vast improvement in test scores; within five years Zavala students were scoring better than the state average.

Testing thus is one way to alleviate the parental information problem I talked about earlier, when we looked at parents' decisions about early childhood education. Testing also can help give individual students more incentive to improve their cognitive skills. Students who do poorly on tests are sometimes forced to attend summer school or to repeat a grade, both of which are unpleasant, and high school graduation is sometimes made contingent on satisfactory test performance.

Testing programs, however, have several drawbacks. First, by providing more accurate, credible information than was previously available about students' cognitive achievement across schools and school districts, such programs may exacerbate the fortress problem. Letting a student from a poor family (and therefore likely to lower test scores) attend school in a wealthy district will make the school's administrators and teachers look worse, even if the

new student's presence has no effect on the children already there. Parents will be willing to pay more to live in good districts, since their assessments of the schools will be based on more solid information, and so every homeowner will have more to lose if students from less wealthy backgrounds attend that district's schools. Segregation may become more stark: parents who chose their location based on some factor other than school quality (maybe they didn't think there were big differences) may discover that their choice is hurting their children's cognitive development and decide to move. Testing programs may also make schools with poor kids less attractive to administrators and teachers: why go to a school where a streak of bad luck with test scores can turn you into a pariah, when you can go to a better, less challenging school and never have to worry?

Second, testing may distort teachers' and students' efforts and make them devote too much attention to the wrong things. In educational circles, this problem is called "teaching to the test" or, more colorfully, "drill and kill." The subset of cognitive skills that standardized tests can measure are only a small part of what students, parents, and the general public want schools to teach. Moral development, critical thinking, artistic and cultural refinement, physical fitness, athletic achievement, and just plain fun—all of these are also valuable outcomes of school. The greater the reward for teaching testable cognitive skills, the less time, resources, and energy will be devoted to the other goals schools have. This can cause resources to be misallocated: the first hour of music instruction may be more valuable socially than the eighth hour of arithmetic, but since the school gets rewarded for arithmetic proficiency, not music, the likely outcome in a testing regime is eight hours of arithmetic and no music. Teachers and administrators may also have incentives to encourage poor students to drop out of school.

In summary, then, testing and standards can probably do a decent job of modestly strengthening the weak incentives of some public schools and some students. But the cost of providing anything beyond weak incentives to a few schools and a few students can be considerable.

Empirically, almost all studies of testing programs find that incentives lead to gains in student skills—as measured by the test scores on which the incentives are based (see, for example, Jacob 2002; Lavy 2002a, 2002b; Glewwe, Ilias, and Kremer 2002; Richards and Sheu 1992). Some gains are modest, some dramatic. Hanushek and Raymond (2004) also show gains on large-scale, independent tests. But some studies that have tried to find out how the gains in test scores are achieved have concluded that they come primarily from "teaching to the test" and administrative manipulation, not from enhancing true learning. In Chicago (Jacob 2002) and in Kenya (Glewwe, Ilias, and Kremer 2002), scores on alternative tests (without incentives attached to them) administered at about the same time showed little or no improvement, and the Kenyan students lost everything they had apparently gained within a

year after the testing program ended. The Kenyan teachers did not alter their pedagogic methods or improve their rather low student attendance rate, but they did conduct review sessions after school geared specifically to the test on which their pay was based. To help improve their test scores, Chicago schools increased the number of students they placed in special education, preemptively retained students in lower grades, and took teaching time away from low-stakes subjects like science and social studies. In Virginia, some schools improved their test scores by serving highly caloric lunches on the days that tests were administered, because a "sugar buzz" helps test-takers (Figlio and Winnicki 2002). And when all else fails, at least 5 percent of teachers cheat (Jacob and Levitt 2003). Hanushek and Raymond (2004), on the other hand, found no increase in special education placements.

B. Vouchers

Vouchers also address the school incentive problem, but they try to create much finer and more nuanced incentives than the crude, coarse ones that testing programs use. Many different voucher schemes have been proposed, and a few have actually been tried. The common element is that all of them center around subsidies for some parents who choose to send their kids to private schools that meet certain standards. (Thus I consider tax deductions and tax credits for private school tuition to be vouchers.) None of the voucher schemes that has been implemented proposes the elimination of public schools (the same holds true for most of the programs that have been studied); instead they foresee public and private schools operating side by side. I'll follow the same approach—although logically nothing would preclude a system entirely without public schools.

Voucher plans differ according to which children are eligible, which private schools they can attend, and how the subsidy is related to the tuition parents pay (no voucher scheme links the subsidy to anything else).

1. Vouchers and Incentives

Vouchers provide strong incentives for both public and private schools to please parents. Private schools that don't satisfy enough parents will go out of business, and those that satisfy many parents well will be able to charge high tuition or expand greatly with many good applicants, or both (depending on the rules of the voucher scheme). Public schools that don't satisfy enough parents will similarly suffer embarrassing shortfalls of students, and will come under pressure from taxpayers to shape up or close down. Administrators and teachers in these schools will suffer career damage. On the other hand, public school administrators and teachers who perform well will see their enrollments rise and their worth validated to the people above them and to taxpayers; private school entrepreneurs will recognize their talents, too, and try to entice them away with attractive offers.

Various kinds of private school behavior can also provide incentives for students to work hard, and for parents to help their children. Private schools can require certain levels of academic achievement and moral development as criteria for admission and retention, and they can ask parents to participate in various ways in their children's education (beyond selling raffle tickets). Or they can vary their tuition, charging less for students who bring more to the school. Private schools have an incentive to act this way—internalizing the external benefits of students' and parents' academic effort—because better classmates make a school more attractive to prospective students and their parents. Not all voucher schemes allow private schools to act like this, but they all could be designed to permit this kind of behavior.

Notice that voucher incentives don't promote the same kind of misallocation of resources that testing program incentives do. If the first hour of music instruction leads to greater parental satisfaction (at the same cost) than the eighth hour of arithmetic, then a school will have the hour of music, even if it leads to lower test scores.

Vouchers do create a serious incentive problem, however. They induce schools to try to please parents and no one else. Thus even if it's really in children's best interest that they learn Mandarin or C programming, for example, or in the general public's best interest that they learn about Guadalcanal or the Holocaust, but parents instead want them to learn French and the Stations of the Cross, parents will prevail, even if teaching Mandarin and the story of Guadalcanal would be a potential Pareto improvement. This wouldn't be a problem if kids were private goods that parents owned, but they're not. Indeed, if kids' education were a concern of parents alone, there would be no reason to subsidize it with vouchers or in any other way. Just like testing, vouchers can lead to a misallocation of resources by overemphasizing one set of goals at the expense of others.

Vouchers are like early childhood education subsidies in this regard. In fact the experience with those subsidies is a massive voucher experiment. Some differences are relevant, however. Parents of older children receive more accurate reports and are more likely to concentrate on the development of cognitive skills, especially if testing programs give them more complete information about some of those skills. And laws could easily require that all schools accepting vouchers meet some minimum standards. Thus there's no reason to think that voucher plans would experience the controversies over unaccredited providers that engulfed child care in the latter part of the twentieth century.

On the other hand, older children are supposed to learn a greater variety of material than younger children, and the subtleties in the material may be greater, too. The divergence between what today's children might be expected to learn and what their parents learned a generation ago is also greater at higher grade levels. Thirty years ago, just as today, five-year-olds needed to be toilet trained and know how to count, things people expect from early child-

hood education; but a high-school graduate today is likely to have a greater need for Spanish, Mandarin, and modern molecular genetics than her parents did. (Levy and Murnane [1998], for instance, emphasize how parents, by using their own outdated experiences, consistently underestimate the skill demands of the contemporary workplace, and so are satisfied with education they shouldn't be satisfied with.) Thus child care subsidies may in fact teach a fairly accurate lesson about the effects vouchers would have.

The child care experience indicates that accreditation standards may not be an efficient or effective way of protecting the interests of children or the general public when those interests diverge from the interests of parents. Such standards have to be verifiable in court and they have to be easy for monitors to observe; that's why the child care standards rely on simple counts of chairs, square footage, fire exits, staff, and sometimes credentials. The general public cares about such things as children's future earning power, industriousness, patriotism, congeniality, generosity, and propensity to obey the law; children (or the adults they will become) worry about the fun they have now, their future earnings, and the options they will have open to them when they become adults. Writing verifiable, easy-to-monitor criteria that would protect these interests would be extremely difficult.

Nor is it clear that the best way to encourage activities that benefit children and the general public, not just parents, is to set a single "pass-or-fail" standard. You might want to encourage good schools to do better—or discourage poor schools from giving up.

To be sure, today's private schools seem to stack up quite well against public schools in the dimensions that children and the general public care about. Careful studies of Catholic schools, for instance, show that they modestly improve students' test scores (relative to the scores those same students would have achieved in public schools) and noticeably improve their chances of graduating from high school and attending college (Ladd 2002). Students in private schools, particularly those in Catholic schools, are no worse than their public-school counterparts in terms of their tolerance for others and knowledge of American civic values, and in some cases their responses are better (Teske and Schneider 2001, p. 622). They also seem to behave no worse (Mocan, Scafidi, and Tekin 2002). Students who had been randomly assigned to a voucher program in Toledo were found to be more generous but less confident (Bettinger and Slonim 2004).

These results, however, deal with averages and the relatively small number of private schools that now exist in the United States. In Chile, widespread use of vouchers led to the establishment not of more Catholic schools, but of secular schools that were somewhat less effective than public schools in promoting cognitive achievement (McEwan and Carnoy 2000, p. 227). Whether all new private schools in the United States would promote civic virtue properly remains an open question, and one that governments are not likely to

leave to chance. In a California referendum, voters rejected vouchers in part because of stories about schools run by witches' covens.

2. Increasing Returns, Variety, and Monopoly

Since elementary and secondary education has increasing returns to scale, adding private schools must force either higher transportation costs or some sacrifice of these returns to scale, or both. Even if private-school households segregated themselves so that the spatial pattern of schools and journeys to school was the same with vouchers as it was without vouchers, nonschool transportation costs would increase, because households that moved in order to be closer to schools would be farther from other locations that mattered to them, like work and relatives.

Of course, in return for these sacrifices, students get greater variety and the opportunity for better matches with their capabilities and interests. In this regard, public schools are like buses, vouchers like cars. Vouchers let parents substitute longer physical journeys to school and higher per-pupil school costs for better matches with the type of school they'd most like to have. This explains in part why vouchers are more popular today than they were, say, fifty years ago. People are richer, and variety is a normal good. Transportation is cheaper, too, especially since elementary school kids no longer go home for lunch. And American society today is much more diverse in many dimensions.

Vouchers thus work better in densely populated cities than sparsely populated rural areas. In cities, the sacrifice in travel time and economies of scale needed to create any additional increment of educational variety is less. Cities can support educational variety in schools the same way they can support mass transit systems with frequent service. In Chile, vouchers led to much bigger private school enrollment gains in cities than in rural areas (Hsieh and Urquiola 2002).

Population density also reduces the monopoly power that private schools might have. In all but possibly the densest cities, however, some monopoly power is likely to be present, because of increasing returns to scale and travel cost. How many holistic Afro-Caribbean Montessori schools can a medium-size city actually support? Monopoly, of course, can be addressed by regulating fees, but regulating fees is meaningless unless quality is regulated, too. We have already seen that private schools financed by vouchers will have to comply with minimum standards to protect the interests of third parties and the children themselves; now we see that they will need minimum standards in almost all cases to protect the interests of parents, too. (We have also seen that minimum standards may not be terribly efficacious.)

Voucher policies may therefore be good ways to reduce sprawl. They make population density more attractive by exploiting some of its real advantages that have not hitherto been fully exploited. Moreover, if monopoly

power of schools has been stronger in central cities than in suburbs because many good substitutes for any particular suburb are available, the improvements in public schooling will also be larger in more densely populated portions of metropolitan areas.

3. Vouchers and the Fortress Problem

The final set of issues with vouchers concerns the fortress problem and in general the allocation of children among schools. An obvious argument for vouchers is that, by severing the link between residences and schools, they destroy the foundation on which fortresses are built—the ability to exclude students from a school because their parents don't live in the school sending district. But since voucher systems don't require public schools to accept students from outside the district, any assault on the foundation would have to be indirect.

Nechyba (1999) outlines one scenario under which such an assault could take place. Currently, wealthy districts include some wealthy families who live there and pay high housing prices only to gain admission to the public schools. Under a voucher system where schools are independent of residence, they could move to a cheaper house in a poorer neighborhood, perhaps a neighborhood closer to work or to cultural attractions, and still find a good school—in this case, a private school that vouchers would pay for. In a closed-city model, this would lower prices in the rich neighborhood, raise them in the poor, and so induce some movement of less wealthy students from worse to better school districts. In an open-city model, rich-district house prices would stay the same (except for the eventual drop in the school's peer quality), but the poorer district house prices would still rise, and some movement would still be induced.

Under some circumstances, moreover, the new private schools that would spring up to serve wealthy families in poor neighborhoods might also enroll some poorer students from those neighborhoods. This would happen if poor students were supplied large enough vouchers, the new private schools were reasonably close, and the students' cognitive skills and behavior were pretty good. Thus, in Nechyba's model, there is no frontal assault on the fortress of public schools in rich districts but there is a slight chipping away, from the fall in relative house prices and the development of alternative good schools that might be accessible to some poor students. Rich families who can live in more convenient locations gain, poorer kids who get into the new private schools gain, and renters (those renters whose children don't change schools) in the poorer neighborhood lose.

Still, it's not clear that this scenario is a potential Pareto improvement. Since the new private schools in poorer neighborhoods are likely to take the best students from that neighborhood's public schools, the remaining students in those schools will be worse off because of lower peer quality. This is a

problem that Manski (1992) emphasizes. With vouchers, private schools (assuming they don't exercise too much monopoly power) will be under competitive pressure to charge each student marginal cost—the right price—but public schools are still prohibited from charging the right price. Good students who leave schools in poor districts don't properly take into account the losses they impose on the students they leave behind.

This is another example of the problem of second-best pricing. If two goods (public and private schools, in this case) are substitutes, and one of them is stuck with the wrong price, then marginal cost is not the optimal price for the other. This was the argument, for instance, for mass transit fares below marginal cost when driving is not subject to optimal tolls. Just as too many commuters will drive instead of taking subways, unless subway fares are below marginal cost, so too will too many good students leave public schools, unless private schools charge them more than marginal cost. This adds another regulatory concern. (Interestingly, according to Nechyba, different kinds of voucher schemes have different effects in counterintuitive ways. It's only because some rich families take advantage of vouchers that some poor kids have a chance to go to school with them. These benefits wouldn't be realized if vouchers were restricted to lower income families. Voucher schemes like this would create peer group losses for poor kids left behind, without the big, offsetting gains to those who go to school with richer kids.)

Thus, without regulation private schools will probably give tuition discounts to more able students. A hierarchy will be created: different private schools will specialize in different segments of the ability spectrum, and the best students will leave the public schools. Epple and Romano (1998) describe in detail what this equilibrium would look like. Epple, Figlio, and Romano (2000) find that even today private schools give significant discounts to the most able students, with a substantial fraction of even the wealthiest students paying little or no tuition. Ladd (2002) and Hsieh and Urquiola (2002) show that large-scale use of vouchers in Chile and New Zealand produced the increased stratification that Epple and Romer predict (although both countries had restrictions on the pricing strategies private schools could use).

4. Summary on Vouchers

Vouchers are likely to make school personnel work harder to please parents, give parents and students a wider variety of educational offerings to choose among, revitalize some poorer neighborhoods, and provide some opportunities for students from poorer families to attend better schools. But incentives remain wrong—just wrong in different ways. These wrong incentives lead to neglect of the interests of kids and the general public, greater overhead and more travel time, and excessive loss of positive peer influences for students who remain in public schools. Vouchers will work better in cities than in less densely populated locations and so may reduce sprawl.

C. Charter Schools and Magnet Schools

One conclusion from the previous section is that private schools that receive vouchers should be fairly heavily regulated—to reflect the interests of kids and the general public, to offset monopoly power, perhaps to keep them from enticing too many of the best students out of public schools—but that it's unclear how adequately this regulation could be accomplished. Considering the difficulties of regulation, it might be that most of the advantages of vouchers can be better secured within the framework of public schools, without privatization. This argument leads to such reforms as charter schools and magnet schools—schools that are public in the sense that they are subject to some discretionary direction by public officials, but private in the sense that they have more autonomy than traditional public schools and are not confined to drawing their students from traditional, exclusive school sending districts.

In the arguments for vouchers in the previous section, only the strong version of school incentives required that the new schools in question be private. Public agencies can create enhanced variety (witness the contrast, for instance, between the formal, almost entirely concrete, open setting of Washington Square Park, essentially a plaza in the middle of a traffic circle, and the almost totally undeveloped wooded wilderness of the Staten Island Greenbelt, both of which are administered by the same public agency, the New York City Parks Department), public agencies can ignore geographic boundaries, public agencies can charge the right prices, and public agencies can compete with one another. Thus appropriately reformed public schools should be able to accomplish many of the same improvements as vouchers—and of course cause many of the same new problems.

The big difference is in incentives. Charter schools and magnet schools have weaker incentives to please parents and stronger incentives to please bureaucrats. But some of what bureaucrats want reflects the interests of kids and the general public. Private schools supported by vouchers have weaker incentives to serve those interests.

Charter schools and magnet schools are not the only possible ways to blend "private" and "public" aspects of school decision making. In France, for instance, private schools receive public funds, but the amount depends on how much control they relinquish to the government. As I emphasized in Chapter 5, running a school is not a simple question of public versus private; it's a long series of questions about individual decisions and particular incentives that can be answered many, many different ways.

D. Zoning Changes

Zoning—nonfungible exclusion of poor families from rich neighborhoods— was the key link in the fortress problem, and we saw in Chapter 8 that an egal-

itarian distribution of rights to exclude could alleviate many of the difficulties this type of zoning causes. Could something similar work for schools?

Obviously, since school congestion and peer group effects are just two more examples of the negative externalities that we considered in Chapter 8, salable rights-to-exclude would cause the same sort of efficiency gains—provided parents represented perfectly the interests of kids and the general public. Because they don't, any sales of rights to exclude would have to be taxed.

Something similar could work if rights to attend all the schools in a state were distributed randomly to all the children in the state, and communities that wanted to exclude certain students were forced to buy back those rights. Again, transactions where above-median-performance schools wanted to exclude below-median-income families or below-median-performance children would be taxed, since parents don't represent all the relevant interests. An alternative to simple taxation would be to require that all payments go to the school the parent's child will be attending, but the transaction would have to be approved by the parent. This is taxation in that the value of what the parent gets—the added payment to the home school—is less than the payment from the excluding school, and it may help the child and the general public.

While probably exceedingly complicated, a system like this would begin to move schools in the direction of marginal cost pricing.

E. Federal Income Tax

The largest single federal activity in the field of elementary and secondary education is a series of income tax preferences. The state and local taxes that pay for public schools are deductible (for itemizers) and the interest paid on bonds to pay for school capital projects is not counted as income by its recipients. Together these items amounted to roughly $16 billion a year in the late 1990s.[2] In addition, that part of the imputed income from homeownership that derives from school availability and the capital gains on houses that derive from school improvements are also untaxed. These are discussed in more detail in Chapter 13, on housing. By way of contrast, direct federal outlays on elementary, secondary, and vocational education were about $18 billion a year, not counting food programs (U.S. Bureau of the Census 2000, tables 538 and 242).

2. This figure is a rough estimate. For 1999, tax expenditures for the property tax deduction were $21.2 billion; for other deductible state and local nonbusiness taxes, $37.7 billion (U.S. Bureau of the Census 2000, table 539). In 1996, state and local expenditures on elementary and secondary education amounted to $279.4 billion, which was 23.5 percent of the total direct expenditure of $1,189.4 billion (table 494). Using the same percentage for the total tax expenditure of $58.9 billion implies a tax expenditure of $13.8 billion on education. Similarly, the tax expenditure for bond interest was $22.8 billion (table 539). Out of a total state and local indebtedness of $1,169.7 billion in 1996, $130.7 billion, or 11.2 percent, was for elementary and secondary schools. This implies an additional tax expenditure on education of $2.5 billion.

There are, of course, many reasons why the federal government has an interest in subsidizing elementary and secondary education. Using the tax code in this way, however, strongly discourages marginal cost pricing, since the federal government subsidizes payments that are not tied to the cost of educating a student, but not those that are. The federal tax subsidies are also greatest for the richest and best school districts, and smallest for the poorest.

V. Conclusion

Education is tough for economists to think about because it makes people who they are; usually we take our people ready-made. We can think of many helpful analogies—schools are like buses because of congestion and increasing returns to scale, like garbage trucks because governments provide them free lest people do worse with substitutes, like open space because some towns overprotect them with zoning, like specialty map stores in that they are best located in cities. Few other topics have such a vast array of affected parties; and few others deal so directly with the intractable question of who we want to be.

But tough doesn't mean impossible. We've seen that parents can't be totally trusted with their children's education, and so compulsory schooling laws, subsidies to early childhood education, and public schools are probably justified. The right way of educating children keeps changing, since it depends on kids, their parents, and technology. A number of trends—better transportation, more educated parents, greater diversity, greater wealth, and better information technology—have made traditional neighborhood public schools less attractive than they were a generation ago, but vouchers are not the only alternative. Greater parental control of education is likely to decrease sprawl, but all increments to parental control are not necessarily potential Pareto improvements.

Our analysis of education, however, has been missing one crucial piece—race. African Americans and Hispanics end up less well educated than Americans from European and Asian backgrounds. Understanding the causes and consequences of this disparity is one of the goals of the next chapter.

Questions

1. The evidence and the arguments on the relative merits of traditional public schools, charter and magnet schools, and vouchers are drawn primarily from the early and middle 1990s. As information technology improves and people learn how to apply it better to education, the relative merits of these three different kinds of schools will surely change. How do you think they will change? How do you think optimal class size will change?

2. How would the arguments about these three types of schools be different in a large city in a developing country, like Beijing?

3. The text showed some of the difficulties that ensue from uniform pricing of the services of public schools. Outline a plan for marginal cost pricing of public schools. Will the plan increase money for the public school system or lose money? How will problems like "teaching to the test" be mitigated? Will residency continue to be a requirement? How will children in poor families be treated? Relative to the current system, is your plan regressive or progressive?

4. Suppose that in any minute there is a 2 percent probability that each student will disrupt a class. In any minute in which one or more students disrupt a class, no one learns anything—it doesn't matter whether one student is disrupting or ten.

   a. In a random minute in a class with twenty students, what is the probability that none is disrupting? What is the probability in a class with thirty students?

   b. Each class lasts 100 minutes a week, and after a semester, a student's test score equals the number of uninterrupted minutes she experienced in an average week. What's the expected test score in a class with twenty students? In a class with thirty students?

   c. Suppose you have one class with twenty students and another class with thirty students. What's the marginal cost (to the receiving class), measured in test points, of having one student transfer from the small class to the large class? How many points does that student lose? What is the marginal benefit to the remainder of the smaller class? Does this move increase or decrease the average score over the two classes? In this model, if the goal is maximizing the average score, would the two class sizes be equalized?

   d. Why would maximizing the average test score be an appropriate goal? Are there any other goals that would be more appropriate? Would those goals imply equalizing class size?

5. Suppose that there are two kinds of students, quiet and noisy. The probability that a noisy student disrupts a class in any given minute is 2 percent; the probability for a quiet student is 1 percent. Teaching and test scores work in the way described in question 4. There are twenty students of each type.

a. Suppose the school employs "tracking"—it assigns all the quiet students to one class and all the noisy students to the other. What's the average test score in the school?

b. Suppose the school decides to have identical classes instead of tracked classes. What's the average test score?

c. Suppose that instead of maximizing the average test score, the school wants to maximize the number of students who pass the exam. The passing score is 70. There are only two teachers and two classrooms. How should students be assigned?

6. Means-tested vouchers have received considerable attention. The idea is that only children with household income below a certain level would be eligible to receive vouchers. Most proponents of means-tested vouchers believe that vouchers should be large enough that entrepreneurs would be tempted to set up new schools to attract these children. This means that the voucher would probably be worth about $8,000 per student.

   Outline a plan for determining eligibility for these vouchers. Should families with more children have lower income limits or higher income limits? How would you handle the labor-supply disincentive effects if the program were to supply a full voucher (or a full set of vouchers for a family with several children) at one income and nothing at all if the income were a dollar more than the cut-off level? If you phase down the value of vouchers as income rises, how do you maintain work incentives in a family with several children? At what point does the phase-down start? How far does it go? (That is, what is the maximum income a family can have and still get some payments from vouchers?)

References

Angrist, Joshua, and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics* 114(2):533–576.

———. 2004. "The Effect of High Stakes Achievement Awards: Evidence from a School Centered Randomized Trial." Forschunsinstitut zur Zukunft der Arbeit (IZA) Discussion Paper 1146.

Argys, L., D. Rees, and D. Brewer. 1996. "Detracking America's Schools: Equity and Zero Cost." *Journal of Policy Analysis and Management* 15:623–645.

Basu, Kaushik. 1999. "Child Labor: Cause, Consequence and Cure, with Remarks on International Labor Standards." *Journal of Economic Literature* 37:1083–1119.

Bettinger, Eric P., and Robert L. Slonim. 2004. "The Effect of Educational Vouchers on Academic and Nonacademic Outcomes." *Upjohn Institute Employment Research* 11(2):6.

Black, Sandra. 1999. "Do Better Schools Matter? Parental Valuation of Elementary Education." *Quarterly Journal of Economics* 114:577–599.

Black, Sandra, Paul E. Devereux, and Kjell G. Salvanes. 2003. "Why the Apple Doesn't

Fall Far: Understanding Intergenerational Transmission of Human Capital." National Bureau of Economic Research Working Paper 10066.

Bolin-Hort, Per. 1989. *Work, Family, and the State: Child Labor and the Organization of Production in the British Cotton Industry, 1780–1920.* Lund, Sweden: Lund University Press.

Brown, Martin, Jens Christiansen, and Peter Philips. 1992. "The Decline of Child Labor in the US Fruit and Vegetable Canning Industry: Law or Economics?" *Business History Review* 66:723–770.

Card, David, and Alan B. Krueger. 1992. "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *Journal of Political Economy* 100:1–40.

Currie, Janet. 2001. "Early Childhood Education Programs." *Journal of Economic Perspectives* 15:213–238.

Dee, Thomas. 2003. "Are There Civic Returns to Education?" National Bureau of Economic Research Working Paper 9588.

Dehejia, Rajeev, and Roberta Gatti. 2002. "Child Labor: The Role of Income Variability and Access to Credit across Countries." National Bureau of Economic Research Working Paper 9018.

Downes, T. A., and J. E. Zabel. 2002. "The Impact of School Characteristics on House Prices: Chicago, 1987–1991." *Journal of Urban Economics* 52(1):1–25.

Epple, Dennis, and Richard Romano. 1998. "Competition between Public and Private Schools, Vouchers, and Peer-Group Effects." *American Economic Review* 88:33–62.

Epple, Dennis, David N. Figlio, and Richard Romano. 2000. "Competition between Private and Public Schools: Testing Stratification and Pricing Predictions." National Bureau of Economic Research Working Paper 7956.

Feinberg, Joel. 1992. *Freedom and Fulfillment: Philosophical Essays.* Princeton, NJ: Princeton University Press.

Figlio, David N., and Maurice E. Lucas. 2000. "What's in a Grade? School Report Cards and House Prices." National Bureau of Economic Research Working Paper 8019.

Figlio, David N., and Joshua Winnicki. 2002. "Food for Thought: The Effects of School Accountability Plans on School Nutrition." National Bureau of Economic Research Working Paper 9319.

Gaviria, Alejandro, and Steven Raphael. 2001. "School-Based Peer Effects and Juvenile Behavior." *The Review of Economics and Statistics* 83:257–268.

Glewwe, Paul, Nauman Ilais, and Michael Kremer. 2002. "Teacher Incentives." Mimeo, Economics Department, University of Minnesota.

Hanushek, Eric A. 1997. "Assessing the Effects of School Resources on Student Performance: An Update." *Educational Evaluation and Policy Review* 19(2):141–164.

Hanushek, Eric, and Margaret E. Raymond. 2004. "Does School Accountability Lead to Improved School Performance?" National Bureau of Economic Research Working Paper 10591.

Hanushek, Eric A., et al. "Does Peer Ability Affect Student Achievement?" National Bureau of Economic Research Working Paper 8502.

Houtenville, Andrew. 1996. "Parental Effort, Student Achievement, and the Quality of Schooling." Working paper, Economics Department, University of New Hampshire.

Hoxby, Carolyn M. 1994. "Do Private Schools Provide Competition for Public Schools?" National Bureau of Economic Research Working Paper 4978.

———. 1998. "What Do America's 'Traditional' Forms of School Choice Teach Us

about School Choice Reform?" *Federal Reserve Bank of New York Economic Policy Review* 4:47–59.

———. 2000a. "The Effects of Class Size on Student Achievement: Evidence from Population Variation." *Quarterly Journal of Economics* 115(4):1239–1285.

———. 2000b. "Peer Effects in the Classroom: Learning from Gender and Race Variation." National Bureau of Economic Research Working Paper 7867.

Hsieh, Chang-tai, and Miguel Urquiola. 2002. "When Schools Compete, How Do They Compete? An Assessment of Chile's Nationwide School Voucher Program." Mimeo, Department of Economics, Princeton University.

Jacob, Brian. 2002. "Accountability, Incentives, and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." National Bureau of Economic Research Working Paper 8968.

Jacob, Brian, and Lars John Lefgren. 2003. "Are Idle Hands the Devil's Workshop? Incapacitation, Concentration, and Juvenile Crime." National Bureau of Economic Research Working Paper 9653.

Jacob, Brian, and Steven Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics* 118(3):843–879.

Kisker, Ellen, and Rebecca Maynard. 1995. "Quality, Cost, and Parental Choice of Child Care." In David M. Blau, ed., *The Economics of Child Care,* paperback edition. New York: Russell Sage Foundation.

Krueger, Alan B. 1997. "Experimental Estimates of Education Production Functions." National Bureau of Economic Research Working Paper 6051.

———. 1998. "Reassessing the View That American Schools Are Broken." *Federal Reserve Bank of New York Economic Policy Review* 4:29–46.

Ladd, Helen F. 2002. "School Vouchers: A Critical View." *Journal of Economic Perspectives* 16(4):3–25.

Lavy, Victor. 2002a. "Evaluating the Effect of Teacher Group Performance Incentives on Student Achievement." *Journal of Political Economy* 110(6):1286–1317.

———. 2002b. "Paying for Performance: The Effect of Teachers' Financial Outcomes on Students' Scholastic Outcomes." Mimeo, Economics Department, Hebrew University.

Lazear, Edward P. 2001. "Educational Production." *Quarterly Journal of Economics* 116(3):777–804.

Levy, Frank, and Richard Murnane. 1998. "Standards, Information, and the Demand for Student Achievement." *Federal Reserve Bank of New York Economic Policy Review* 4(1):117–124.

Lleras-Muney, Adriana. 2002a. "The Relationship between Education and Adult Mortality in the United States." Unpublished paper, *www.princeton.edu/~alleras.*

Lleras-Muney, Adriana. 2002b. "Were Compulsory Attendance and Child Labor Laws Effective? An Analysis from 1915 to 1939." *Journal of Law and Economics* 45(2):401–435.

Lochner, Lance, and Enrico Moretti. 2004. "The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports." *American Economic Review* 94(1):155–189.

Manski, Charles. 1992. "Educational Choice (Vouchers) and Social Mobility." *Economics of Education Review* 11:351–369.

McEwan, Patrick J., and Martin Carnoy. 2000. "The Effectiveness and Efficiency of Private Schools in Chile's Voucher System." *Educational Evaluation and Policy Analysis* 22(3):213–239.

Milligan, Kevin, Enrico Moretti, and Philip Oreopoulos. 2003. "Does Education Improve Citizenship? Evidence from the United States and the United Kingdom." National Bureau of Economic Research Working Paper 9584.

Mocan, H. Naci, Benjamin Scafidi, and Erdal Tekin. 2002. "Catholic Schools and Bad Behavior." National Bureau of Economic Research Working Paper 9172.

Moehling, Carolyn M. 1998. "State Child Labor Laws and the Decline of Child Labor." Mimeo, Economics Department, Ohio State University.

Nechyba, Thomas J. 1999. "A Model of Multiple Districts and Private Schools: The Role of Mobility, Targeting, and Private School Vouchers." National Bureau of Economic Research Working Paper 7239.

Oreopoulos, Philip, Marianne E. Page, and Ann Huff Stevens. 2003. "Does Human Capital Transfer from Parent to Child? The Intergenerational Effects of Compulsory Schooling." National Bureau of Economic Research Working Paper 10164.

Psacharopoulos, George. 1994. "Returns to Investment in Education: A Global Update." *World Development* 22(9):1325–1343.

Richards, Craig, and Tian M. Sheu. 1992. "The South Carolina Incentive Reward Program: A Policy Analysis." *Economics of Education Review* 11(1):71–86.

Scholliers, Peter. 1995. "Grown-ups, Boys, and Girls in the Ghent Cotton Industry: The Voortman Mills, 1835–1914." *Social History* 20:201–218.

Sonnenstein, Freya L. 1991. "The Child Care Preferences of Parents with Young Children." In Janet Hyde and Marilyn Essex, eds., *Parental Leave and Child Care: Setting a Research and Policy Agenda.* Philadelphia: Temple University Press.

Summers, Andrew, and Barbara Wolfe. 1977, "Do Schools Make a Difference?" *American Economic Review* 67:639–652.

Teske, Paul, and Mark Schneider. 2001. "What Research Can Tell Policymakers about School Choice." *Journal of Policy Analysis and Management* 20(4):609–632.

U.S. Bureau of the Census. 2000. *Statistical Abstract 2000.* Washington, DC: Government Printing Office.

Walker, James R. 1995. "Public Policy and the Supply of Child Care Services." In David M. Blau, ed., *The Economics of Child Care,* paperback edition. New York: Russell Sage.

———. 1992. "New Evidence on the Supply of Child Care: A Statistical Portrait of Family Providers and an Analysis of Their Fees." *Journal of Human Resources* 27(1):40–69.

Zimmer, Ron W., and Eugenia F. Toma. 2000. "Peer Effects in Private and Public Schools across Countries." *Journal of Policy Analysis and Management* 18(1):75–92.

Chapter 11

Race and Space

————

W. E. B. Du Bois thought that in the long run "the greatest human development is going to take place under experiences of widest individual contact" (Lewis 1995, p. 558). Cities should be the places for widest individual contact. Wonderful things can happen when large numbers of people gather together, and they can coordinate their actions through a market or some other system. Diversity should be a positive force in a market society: the gains from trade are greatest when traders are most different, and innovation is usually spurred when differing traditions and ways of looking at the world encounter each other. A chapter about race should be a chapter about the good things that city living brings.

Yet almost all systems of coordination, markets especially, work poorly if their participants significantly fear, distrust, or dislike each other. The cost of racial and ethnic discord is that it can keep people from taking advantage of the increasing returns to scale that city life brings, and from enjoying the cultural and technological dynamism that diversity fuels.

To the extent that minority presence makes nonminorities eschew cities in general, or certain parts of cities in particular, or public transportation, or public pools or schools, or hiring certain types of job applicants, people lose out on advantages of city life that they could have had if race or ethnicity mattered less. The same result occurs when minorities find themselves losing out on jobs, mortgages, marriages, health care, or houses in some neighborhoods, because they can't complete what would otherwise be Pareto-improving deals.

The question I ask in this chapter, therefore, is whether the people in American cities are gaining as much as they should be from the diversity of their populations.

We begin with a short discussion of the definitions and biology of race and ethnicity. The next section asks how race and ethnicity affect the way cit-

————

To get the most out of this chapter, you should be familiar with these concepts: location theory, the monocentric city model, potential Pareto improvement, and sprawl. You can find these terms in the Glossary.

ies are arranged—where people live, whether they own or rent, how good that housing is, how much they pay for it, how their children are educated. Race matters an awful lot. The third section asks what the consequences of this arrangement are—the first part looks at whether minorities gain or lose from the current situation, and the second part does the same for whites. In Chapter 12 we will examine policy options.

I. What Race and Ethnicity Mean These Days

The federal government's official division of people in this country is two-dimensional: people are classified based on what their "race" is, and on whether they are Hispanic or not. Thus everyone is supposed to fit into one (and before 1997, only one) of the following twelve boxes:

	Hispanic	Not Hispanic
White	☐	☐
Black or African American	☐	☐
Native American or Aleut	☐	☐
Asian	☐	☐
Native Hawaiian or Pacific Islander	☐	☐
Other	☐	☐

You get to choose which box to put yourself in, and in fact many people change boxes over the course of their lifetime.

It's easy to see that these categories have nothing to do with biology (why, for instance, are people from India grouped with people from China, while people from Guyana are different from people from Venezuela?). They aren't about geography either, because everyone in question lives in the United States.

If not biology or geography, what are these categories, then, in the Office of Management and Budget's boxes? Not culture: you don't have to like salsa to be counted as Hispanic or Bach to be non-Hispanic white. Not income. Not education.

Probably the most accurate description is that roughly between 1840 and 1930, many of the people who ran the United States thought certain easily recognizable groups were genetically inferior—descendants of slaves, Chinese, Mexicans, Indians. The OMB's categories are essentially an attempt, starting in the early 1970s, to sanitize and universalize these old stereotypes, with everyone else being labeled "white." (A good deal of politics has also been added. For instance, the old category "Asian or Pacific Islander" was broken up because Hawaiians didn't want to be included in the same college admissions pool as Chinese and Korean students.)

Still, silly though they may be, these distinctions are important, because

what people think of you and what they thought of your ancestors are important. This is an empirical proposition, and it will be amply demonstrated in this chapter.

II. How Race Matters

This section is devoted to summarizing the evidence that these seemingly nonsensical divisions matter a lot. We'll postpone most analysis to the next chapter. When I say "minorities," I'll mean African Americans, Hispanics of any race, and sometimes Asians and Native Americans. Most of the information presented will be about African Americans. Social scientists have gathered much more information about African Americans than about other minority groups, and African Americans have been a much larger presence in American cities much longer.

A. Segregation

In the United States today, if you know what neighborhood somebody lives in, you can almost always make a very good guess about what their race is and what their neighbor's race is. This phenomenon is called **segregation.** Segregation means that different races live mainly separately, in different neighborhoods.

By itself, there's nothing wrong with segregation. Common interests and common tastes may draw similar people together; healthy segregation would be the result. But less benign forces could be at work, too. Segregation might reflect coercive efforts to keep some people from living where they want to live and where it makes sense for them to live; reducing segregation in this case would be a potential Pareto improvement. Or segregation could arise because people fight too much in mixed-race communities; in that case reducing the proclivity to fight could help a lot of people.

Segregation, then, is a symptom, like a pulse of 200 beats per minute or a weight of 300 pounds. If you've just climbed thirty flights of stairs, it's perfectly fine to have a pulse of 200, and if you're six feet five and all muscle, 300 pounds is a good weight. But under ordinary circumstances, either symptom calls for further investigation.

The first step in any investigation is to measure the phenomenon. Segregation is measured in two different ways.

One is called the **index of dissimilarity.** This index starts with the idea of perfect integration, or no segregation. If a city or a metropolitan area is perfectly integrated, then by definition every neighborhood has the same proportion of minorities as every other neighborhood. The index of dissimilarity measures the distance a city is from perfect integration by figuring the size of the shake-up that would be needed to achieve perfect integration; the bigger

the shake-up needed, the greater the degree of segregation. Thus the index equals the minimum proportion of minorities who would have to change neighborhoods to make the city perfectly integrated (if no whites move). If a city is perfectly segregated—that is, if every neighborhood is entirely one race or the other—the index of dissimilarity equals one (if a city is 20 percent minority and perfectly segregated so that all minorities live in all-minority neighborhoods, then each of those neighborhoods would have to become empty, since no whites live there, to achieve perfect integration, and so 100 percent of the population of each of those tracts would have to relocate; the index of dissimilarity would be one). So the index of dissimilarity ranges between zero and one, with larger values indicating greater segregation. The index of dissimilarity compares two groups.

The other measure is called the **index of isolation.** This index equals the proportion of minorities who live in the neighborhood that the average minority group member lives in; that is, the probability that a neighbor of a minority group member is also a member of that minority group. With perfect integration, this index is the proportion of that minority in the whole city (since every neighborhood has the same composition), and with perfect segregation the index equals one (every minority group member lives in an all-minority neighborhood). Once again, higher values indicate greater segregation.

How much segregation is there? A lot. First, let's look at African Americans. For blacks versus non-Hispanic whites, the average index of dissimilarity for major U.S. metropolitan areas was .65 in 2000, down from .74 in 1980 (this weights cities by black population). Two-thirds of blacks would have to move to achieve perfect integration. Some cities had very high dissimilarity indexes in 2000: .85 in Detroit, .82 in New York and Milwaukee-Waukesha, .81 in Chicago, and .80 in Newark. Even the least segregated cities with substantial African American populations had dissimilarity indexes far from zero: .46 in Norfolk–Virginia Beach–Newport News, Raleigh-Durham–Chapel Hill, and in Augusta-Aiken (Lewis Mumford Center, 2001a). Children were even more segregated, with an average dissimilarity index of .68 (Lewis Mumford Center, 2001b).

Isolation indexes were large, too. In 2000, the average African American lived in a neighborhood where 51 percent of the neighbors were African American (down from 61 percent in 1980), and the average African American child was growing up in a neighborhood where 56 percent of the children were African American. The index of isolation for Chicago was .79; for Detroit and Memphis it was .73.

Hispanics were also segregated, but not quite so highly. The average index of dissimilarity for Hispanics versus non-Hispanic whites was .52, up a bit from .51 in 1980 (weighted by Hispanic population). New York had the highest segregation index at .65, Laredo the lowest at .29 (the population of Laredo

is close to 95 percent Hispanic). Hispanic children were also considerably more segregated than Hispanic adults, with an average index of dissimilarity of .55. Among Hispanic children, the average index of isolation was almost as high (.53) as the average index of isolation among black children (.56)—but the index of isolation gives heavy weight to metropolitan areas, like Laredo, that are almost entirely Hispanic.

Even among Hispanics, color matters a lot. Among Caribbean Hispanics in 1980, black and white Hispanics were highly segregated from each other, with an average dissimilarity index of .61 in ten major metropolitan areas (Denton and Massey, 1989). Moreover, among Caribbean Hispanics in those ten cities, the index of dissimilarity between black Hispanics and non-Hispanic whites was much higher than the index between white Hispanics and non-Hispanic whites (.80 as opposed to .52).

Ethnicity also matters, with Cubans and Mexicans being less segregated from non-Hispanic whites than Puerto Ricans and Dominicans. In 2000, the average index of dissimilarity between Cubans and non-Hispanic whites was .49; for Dominicans, it was .81.

Asians are less segregated than African Americans and Hispanics, with an average index of dissimilarity in 2000 of .42 in the forty metropolitan areas with the most Asians. In some of the metropolitan areas with substantial Asian populations, however, segregation was considerably greater—.50 in New York and Stockton-Lodi (California), for instance.

The different minority groups also tend to be segregated from each other, but not so segregated as they are from non-Hispanic whites.

Of course, these high degrees of segregation could be explained benignly. But the benign explanations do not hold up well, especially for African American segregation. Maybe different races live apart because they have different incomes; perhaps what we really are observing is segregation by class, not by race. In fact, however, income explains only a small portion of the racial segregation we observe.

One way of seeing this is to look at the index of dissimilarity for households in a particular income class. If income were responsible for segregation between blacks and whites, say, then if you just looked at the indexes of dissimilarity for particular income classes, they would be very small. Poor blacks would be integrated with poor whites and rich blacks would be integrated with rich whites, but blacks and whites would appear to live in different neighborhoods because they belonged to different income classes.

This simple check shows that income differences are most definitely not responsible for most of the segregation we observe. For instance, if you take the top ten metropolitan areas in 2000 and compare the index of dissimilarity for the whole population with the index of dissimilarity for middle-income households (income between 175 percent and 350 percent of the poverty level), you find essentially no difference. The index of dissimilarity between

middle-income blacks and middle-income whites is in fact on average some-what higher than the overall index. Middle-income Hispanics are slightly less segregated, and middle-income Asians are almost exactly as segregated as Asians overall.

Taking a more sophisticated approach to this question, Sethi and Somanathan (2001) calculated what the index of dissimilarity would be if in-come were all that mattered in determining where people lived. Every rich neighborhood would have the same racial composition as every other rich neighborhood; every poor neighborhood, the same racial composition as ev-ery other poor neighborhood. If income were all that mattered, most U.S. metropolitan areas in 1990 would have had indexes of dissimilarity ranging from .08 to .18, with an average of .12. This is about one-fifth of the actual dissimilarity indexes for the cities they looked at.

The other benign explanation for segregation is clannishness: people of the same background like to live together so they can talk their own language, swap their own stories, play their own games, patronize their own restaurants, and attend their own churches. The segregated neighborhoods we see today, in this view, are just like the famous Little Italys and Germantowns at the be-ginning of the twentieth century.

This explanation also fails. We've already seen that Hispanics and Asians, immigrant groups that include many non-English speakers, are less segregated from non-Hispanic whites than blacks are, even though most blacks speak English and fewer are non-English-speaking immigrants. It is difficult to see, moreover, why culture would split black Hispanics from white, or make white Hispanics live more closely with white non-Hispanics than with black His-panics.

In another test, Cutler, Glaeser, and Vigdor (1999) argued that if clan-nishness were responsible for segregation, then recent arrivals from out of state in large cities would be the most clannish of African Americans, and so they would be willing to pay more to live in heavily black neighborhoods. Looking at data from several censuses, they find no evidence of this.

Although no benign explanation for current levels of segregation seems to work, and although segregation for most minorities today is much greater than it was for immigrant groups in the early 1900s, the situation is not get-ting any worse. The dominant trend since 1970 has been toward lessening seg-regation of African Americans and stable or only modestly increasing segrega-tion of Hispanics and Asians. Segregation of African Americans increased in every census between 1900 and 1970 and has fallen in every census since then. This trend has three chief components.

The first is the shift of population to the Sun Belt. Sun Belt cities, like Sacramento, Augusta-Aiken, Norfolk–Virginia Beach–Newport News, Tucson, and Anaheim, are among the least segregated in the country (retirement com-munities are an exception); Frost Belt manufacturing centers, like Detroit,

Chicago, Cleveland, and Newark, are among the most. As the proportion of Americans who live in the Frost Belt falls, average segregation decreases.

The second trend has been a drastic decrease in the number of neighborhoods that are essentially all white. In 1970, 62.6 percent of whites in metropolitan areas lived in neighborhoods (census tracts) that were less than 1 percent black; by 1990 only 35.6 percent of whites did. This trend has affected Frost Belt cities as well as Sun Belt (Ellen 2000b).

Third, integration is much more stable in many neighborhoods than it used to be. Integration used to be called "the time between when the first black family moves in and when the last white family moves out," but since 1980 especially, quite a few neighborhoods with sizable proportions of both blacks and whites have stayed that way for a long time. Between 1980 and 1990, 76 percent of the tracts that were integrated at the beginning of the decade were still integrated at the end; in 53 percent of them, white population grew (Ellen 2000b).

Indeed, the classic story of "white flight"—all the whites quickly leave a neighborhood once a few blacks move in—no longer seems to be the mechanism through which segregation is maintained. People move all the time, for all sorts of reasons, but Ellen finds that the proportion of blacks in a neighborhood has little effect on the propensity of whites to move out of it. On the other hand, whites are very reluctant to move into a neighborhood with a significant proportion of blacks. So it is the small inflows—of whites to neighborhoods with a significant number of blacks, and blacks to neighborhoods with very few other blacks—that maintains segregation, not large outflows of panicking whites.

B. Centrality

Not only do minorities have an unusually high propensity to live in neighborhoods with other minorities, they also have an unusually high propensity to live in neighborhoods that are close to historic central business districts of cities, the CBDs. Minority neighborhoods, particularly African American neighborhoods, are much closer to the CBD, on average, than white neighborhoods. That's why the adjective *inner-city* is a synonym for *minority*.

The measurement of centrality is similar to measurement of segregation. The question you ask is what proportion of minorities would have to move, at a minimum, if minorities were to end up having the same distribution of residences by distance from the CBD as whites have. The difference is that moves out (away from the CBD) are counted as positive, moves in are counted as negative. So the index of centrality equals one if all minorities live closer to the CBD than all whites; it equals zero if minorities have the same distribution of distance to the CBD as whites do; and it equals minus one if all minorities live farther from the CBD than all whites.

The index of centrality for African Americans is quite high in most metropolitan areas. In 1980, for the eighteen northern cities with the largest black populations, it averaged .88; for the twelve biggest southern cities it averaged .75 (Massey and Denton 1993). Little is known about centrality for Hispanics and Asians.

Income doesn't explain much about why black housing is so centrally located. If people lined up perfectly in order of income, with the poorest households closest to the CBD, and in every city the distribution of income by race were the same as it was in the United States as a whole in 1997, then the index of centrality would be around .23 or .24 (U.S. Bureau of the Census 1999, table 752), not .88 or .75. In fact, income is not very well correlated with distance from the CBD—in many old cities, like New York and Chicago, some of the wealthiest neighborhoods are immediately adjacent to the CBD. Income doesn't seem to be able to explain much more than a quarter to a third of why blacks live so much closer to CBDs than whites do.

Land near the CBD is one of the most valuable resources in any metropolitan area, and it is a resource in fixed supply. Apparently race is an important determinant of who uses that land. This is strange, since there is no obvious connection between skin color, hair texture, and the benefits or costs of living next to a CBD.

The optimal allocation of people to neighborhoods—and the allocation a well-functioning market would accomplish—assigns to each neighborhood the people who benefit the most from being there. To argue that concentrating minorities in inner cities is a good idea, then, you have to argue both that they derive benefits from being there and that no one else would derive more benefits. Since cities are great places, the former proposition is probably true, even though there is some evidence to doubt it (I'll discuss this in section III.A.3, on the spatial mismatch hypothesis). The latter proposition is more dubious.

C. Commuting Time

You would expect that because they live so close to CBDs, African Americans have shorter commuting times than whites do. Shorter commuting time is the main benefit that people get from living near a CBD, and so almost any argument in favor of the current arrangement has to start with the benefits that minorities get from being close to their jobs. Giving minorities great access to the CBD makes sense only if access to the CBD is differentially valuable to minorities.

It turns out, though, that minorities have longer commuting times, not shorter. On average, the commutes of African Americans who live in central cities are longer than those of any group that lives in the suburbs. Table 11.1 provides 1990 data.

Nor is there any easy, benign explanation for this difference to be had by

Table 11.1   Mean travel time to work, 1990 (minutes)

	Residents of metropolitan areas		
	All residents	Central city	Not from central city
White	22.6	20.8	23.5
Black	26.1	26.5	25.5
Hispanic	24.7	24.9	24.4

Source: 1990 Census of Population, Social and Economic Characteristics: United States, 1990 CP-2-1 (Washington, DC: U.S. Bureau of the Census, November 1993), tables 63, 64, 67.

appeal to the other traits of minority workers. For data from 1985, Gabriel and Rosenthal (1996) find that blacks have 23-percent longer commutes and Asians have 25-percent longer commutes than whites do, even when you hold earnings, age, gender, education, and family composition constant. If minorities gain more than whites from living near the CBD, it's not clear how they do so.

D. Homeownership and Housing Quality

Minorities are less likely than whites to be homeowners, and more likely to live in low-quality housing than whites with comparable incomes, families, and background. This is especially true in cities, where most housing in the older, central part is apartments.

In 2000, 73.7 percent of white non-Hispanic households lived in houses they owned, as opposed to 46.7 percent of black households and 45.4 percent of Hispanic households. Part of this difference, of course, stems from minorities' lower incomes, and slightly younger ages, but not all of it. Wachter and Megbolugbe (1992), for instance, found that 19 percent of the difference in homeownership rates could not be explained by standard, nonracial variables. Freeman and Hamilton (2002), however, find that the unexplained disparity in homeownership rates fell in the 1990s, but did not disappear, possibly as a result of federal mortgage initiatives.

Once they buy houses, moreover, African Americans are also less likely to "trade up" to better houses—their first home is often their last. Again, this result is true holding most observable characteristics, like income, constant (Boehm and Schlottman 2004).

Race and ethnicity also make a big difference in the quality of housing that people purchase. Among only poor households, for instance, in 1989, 4 percent of white households were overcrowded (more than one person per room), as compared with 9 percent of black households and 26 percent of Hispanic (Yinger 1995, p. 109).

DiPasquale and Kahn (1999) constructed an index of structure quality

and an index of neighborhood quality for Los Angeles county in 1990. They found that black and Hispanic recent movers on average live in neighborhoods considerably worse than whites with the same income, age, household composition, and nativity status. Minority renters move to structures with about the same quality as comparable white renters move to, but minority owners buy houses of somewhat lower quality than comparable white owners. Even with owners, the neighborhood difference is much larger than the structure difference.

E. Housing Prices

Do minorities pay more than whites do for comparable housing? This question is very difficult to answer, primarily because it is difficult to phrase. Housing is a complex bundle, encompassing structure, amenities, neighborhood, and location. Minorities and whites buy bundles that differ in many ways, some obvious and some subtle. Figuring out what "comparable housing" might mean is the crux of the problem, and it may be impossible to answer.

For instance, if you decide that "comparable housing" means "comparable structure" and you ignore neighborhood quality completely, then most of the literature indicates that since 1970 or so, minorities have paid less for housing than whites have (Mieszkowski 1979, Chambers 1992) or that housing in minority neighborhoods is cheaper than housing of the same (structural) quality in white neighborhoods (DiPasquale and Kahn 1999, table 3; Gablinger 2003). (Apparently this was not the case before 1970, when the pressure of the Great Migration and obviously coerced segregation made blacks pay more for housing of lower structural quality [Mieszkowski 1979; Cutler, Glaeser, and Vigdor 1999; Kain and Quigley 1975].)

We have seen, however, that minorities generally live in lower quality neighborhoods—neighborhoods with less successful schools, more crime, higher car insurance rates, higher density, more of some kinds of pollution. Gablinger (2003) controlled for some of these factors, but not all, and found that mainly African American neighborhoods were still cheaper than mainly white neighborhoods, but integrated neighborhoods appear to be cheaper than both. If minorities pay lower prices for lower neighborhood quality, are the prices low enough to make up for the lower neighborhood quality?

From this phrasing, you can see that our discussion has degenerated into asking a very difficult if not impossible question. Apples are better than oranges on some dimensions, but oranges are squishier and easier to share. Who would want to debate whether apple buyers or orange buyers are paying a higher price for fruit?

We can say a few things about prices, however, without venturing into nonsense. Race does not seem to matter much for the price an individual pays,

given the composition of a neighborhood. Blacks in white neighborhoods pay about what whites do; whites in black neighborhoods pay about what blacks do (Chambers 1992; Gablinger 2003). Whites don't seem to face significant **discrimination** in moving to minority neighborhoods, but they rarely do so (in the Los Angeles County data, the average white recent mover was in a neighborhood that was 5 or 6 percent black); thus their actions tell us that for most whites the lower prices in minority and especially black neighborhoods are still too high to offset the lower neighborhood quality as they perceive it. You could make a similar statement about minorities finding prices in white neighborhoods too high, but you would have to be very careful to include among the qualities of a neighborhood the barriers that realtors, community members, and racist maniacs might erect to keep minorities out. I discuss these barriers in the next chapter.

F. Car Ownership

Minority households are less likely than other households at all education levels to own cars. Overall, in the years 1992 to 1994, 25 percent of black households and 20 percent of Hispanic households lacked cars, as compared with 5 percent of white households. The average black household had 0.671 cars per adult; the average Hispanic household, 0.725; and the average white household had 1.135. Among high school graduates, the average black household had 0.648 cars per adult; the average Hispanic household, 0.692; and the average white household, 1.146 (Raphael and Stoll 2001).

G. Education

African Americans and Hispanics don't do as well in school as whites and Asians. They're less likely to finish high school or college, and less likely to score well on standardized tests. Partly the problem arises from poverty, but poverty is not the whole story. Holding socioeconomic background constant, African Americans and Hispanics do worse on standardized tests than whites and Asians, although the gap is slowly decreasing (Fletcher 2001). This gap exists even in integrated, affluent school districts like Shaker Heights in Ohio and South Orange–Maplewood in New Jersey.

Minority students work at least as hard as white students, and in some important dimensions they work harder. In Shaker Heights, black students spent as much time on homework as whites, and ranked at least as high on measures of academic motivation (Fletcher 2001). Nationally, minority students are likely to persist at least as long in school as white students with the same family background, and in many instances to persist longer (Cameron and Heckman 2001). There is no evidence that genetic differences are responsible for the test score gap.

H. Summary

Race matters. It matters because people think it matters (or people think that people think that it matters, or something at an even greater remove). Minorities disproportionately live in segregated neighborhoods close to the CBD. Income explains little of this centralization and segregation. Blacks commute longer distances. Everything else being equal, minorities are considerably less likely to own their own homes or cars or to live in neighborhoods with successful schools and low crime rates. They score less well on standardized tests of educational achievement, even with the same work effort and family background.

III. Consequences

Sorting where people live by a largely irrelevant characteristic like skin color or hair texture, and locating large numbers of people who don't work there right next to the central business district seems, on the face of it, a stupid way to organize a city. Before dismissing this pattern, however, we need to examine its consequences more closely. Perhaps the pattern has some hidden advantages that explain its prevalence or justify its continuation. In the first part of this section we will look at how this pattern affects minorities; then we'll do the same for whites. Data on these questions have been almost entirely concentrated on African Americans.

A. Consequences for Minorities

1. Psychological Benefits of Togetherness

One possible explanation for the pattern of segregation is that minority groups might feel better living together. In segregated neighborhoods they can relax more easily, find friends and people with whom they share experiences more quickly, be themselves more openly, cultivate their cultural inheritance more deeply, understand themselves better. If this is the case, and something in the way markets work normally prevents these aspirations from being realized as fully as they should be, then perhaps maintaining the current pattern is worthwhile.

Notice, however, that a preference for togetherness is an argument for segregation, not for centralization. Harlem could perform these functions just as effectively somewhere in southern New Jersey as it could fifty blocks from Midtown.

Even this version of segregation without centralization finds little support among minorities in survey data. In the early 1990s, 98 percent of blacks in the Detroit area and 99 percent in the Los Angeles area said they were willing to move into a neighborhood that was 53 percent black. A neighborhood

that was around half black was the first choice of most respondents. In the early 1970s in Detroit, an all-black neighborhood was the first choice of only 12 percent of those surveyed and the last choice of 27 percent—and 31 percent of blacks said they would refuse to move into such a neighborhood (Massey and Denton 1993, pp. 91, 89). In the 1990s, a quarter of blacks said they would go so far as to refuse to move into an all-black neighborhood (Ellen 2000b, p. 135).

The way that minorities have chosen the metropolitan areas they live in also shows that whatever appeal separatism has is quite weak. Minorities are rather well dispersed among metropolitan areas in the United States; they are not concentrated in only a handful. For blacks (versus nonblacks) the index of dissimilarity for the twenty-five largest metropolitan areas in 1996 (treating each metropolitan area as a "neighborhood") was .28. Only a little over a quarter of African Americans would have to move between metropolitan areas to achieve a uniform distribution, as opposed to two-thirds who would have to move within metropolitan areas (U.S. Bureau of the Census 1998, table 44). The index of dissimilarity for Hispanics between metropolitan areas—.37 in 1997—is also lower than the index within metropolitan areas. Both blacks and Hispanics are more segregated *within* metropolitan areas than between metropolitan areas, and for blacks the gap is huge. (The index of dissimilarity for Jews across states is .40, according to data in Singer 1999. Thus if clannishness is measured by segregation across the United States, blacks and Hispanics are probably less clannish than Jews.)

This high degree of integration between metropolitan areas indicates that increasing returns to scale in the benefits to ethnic or racial togetherness seem to be very limited. The arguments for the psychic advantages of segregation don't say how big a minority neighborhood has to be to realize them, but there is no hint in these arguments that a minority neighborhood of 500,000 does any better at realizing them than twenty-five neighborhoods of 20,000 each.

Thus, while minorities gain some benefits from living in neighborhoods with more than a handful of other minorities like themselves, it does not appear that they derive any substantial benefits from living in highly centralized aggregations of hundreds of thousands of people of the same race. Many patterns considerably different from the ones we observe now could satisfy almost all the desires minority people actually have for segregation.

2. Business

Another argument for the current pattern is that it encourages minority business. Large minority populations permit minority businesses to get started, and shelter them from white competition until they have grown strong enough to stand on their own in the wider marketplace. Notice that, once again, this is an alleged advantage of segregation, not centralization.

If segregation promotes minority business, then eighty years of intense segregation should have turned African Americans into the most successful group of entrepreneurs the world has ever seen. But it has not.

On the contrary, even when you control for education, family background, income, gender, and other standard characteristics, African Americans are considerably less likely than whites in general and members of other ethnic groups in particular to be self-employed. Meyer (1990), for instance, finds that only 3.4 percent of black workers are self-employed, as opposed to 10.0 percent of white. Only about a quarter of this self-employment gap is attributable to differences in age, education, gender, marital status, the presence of children, and region.

Of course, segregation probably is not directly responsible for this large disparity; attitudes of consumers and lenders probably matter, too, and are probably more important. But even if segregation is a positive force, it is clearly not a particularly powerful one for promoting minority entrepreneurship. And to the extent that other factors make minority entrepreneurship a shaky enterprise, investing in segregation to promote it is not wise. If your car lacks a motor, for any reason, you don't buy it expensive new tires.

In an extensive study of minority business, Bates (1997) finds direct evidence on the role of segregation. Minority businesses that are located in minority neighborhoods and serve mainly minority customers are less successful than those businesses that operate in a wider world. Holding the entrepreneur's background and several other characteristics of the business constant, Bates found that those businesses that primarily served minority customers were significantly less profitable and no more likely to survive (pp. 161, 158). Since 1970, the growth in black-owned businesses has been in fields where few customers are minorities; the traditional small businesses are declining in number. Segregation appears to create an environment in which minority businesses fail, not one in which they thrive.

3. Jobs

While the first two consequences arise from segregation, a third consequence results from centralization. In almost all metropolitan areas, the largest single concentration of jobs is in the historical CBD. Centralization should make minority neighborhoods great places for commuting to work and for finding jobs. Location theory implies that land prices will be bid up high enough that residents of these great places will be no better off than those otherwise identical people who live in less favored locations, but if for some reason the market fails to accomplish this equalization, centralization will be an unalloyed labor market benefit for minorities.

But just as segregation has not produced minority entrepreneurs, centralization has not produced minority workers with short commutes. We have already seen that minorities have longer commutes than whites, not shorter

ones. Blacks and Hispanics also are less likely to be employed than whites are, at every age and education level.

Would the employment situation be even worse if minority neighborhoods were not so centralized? This question has been studied intensively. The overwhelming consensus answer is no. In fact, the debate on this question is about how much residential centralization *hurts* minorities' employment situation, not about whether it helps.

The idea that residential centralization hurts minorities' employment situation is called the **spatial mismatch hypothesis.** In this view, CBD jobs mainly require high levels of education, while the low-skill jobs that minorities need have mostly moved to distant reaches of the suburbs.

The spatial mismatch hypothesis has been challenged in several different ways. Ellwood (1986), for instance, compared the South Side and the West Side of Chicago in 1970. The two neighborhoods were almost entirely African American and about equal distances from Chicago's CBD, the Loop (although the further reaches of the South Side were farther than anything in the West Side). The West Side, though, happened to be chock-full of factories and warehouses offering plenty of low-skill jobs, while the South Side was almost completely devoid of major employment centers. Ellwood reasoned that adults who have jobs may move closer to them, and so adult employment may not offer a good test of the spatial mismatch hypothesis. However, families usually don't relocate to accommodate the jobs that teenagers hold, and so if spatial mismatch meant anything, teenagers on the West Side should have higher employment rates than teenagers on the South Side.

They didn't. Most jobs in West Side factories and warehouses were held by people from outside the neighborhood, many of them white. So far as getting jobs was concerned, the Chicago metropolitan area of 1970 could be looked at as a single unified labor market, one where people of all races could fairly easily find out about substantial numbers of jobs anywhere. Blacks did worse in that market, and the reason, in Ellwood's phrase, was "race, not space." Leonard (1985) reached a similar conclusion for Los Angeles in 1980; so did Raphael (1998) when he used Ellwood's methods for analyzing 1990 data for the San Francisco Bay Area.

At the same time, a number of other studies have found that minority teenagers in neighborhoods with better job access are more likely to be employed. The difficulty here is in measuring "better job access": essentially you want to know how easy it would be for people who don't have jobs to get to the jobs they would have if they had jobs. Clearly no measure of this is going to be perfect, and many substantially different answers to this hypothetical question are equally viable.

Ihlanfeldt (1992), for instance, measured a neighborhood's quality of job access for teenagers of a particular race by the average commuting time of employed, low-skilled adults of the same race in that neighborhood. For out-of-

school teenagers living at home, he found that this measure of job access had a substantial effect in the 1980 census: the longer the average commuting time for low-skilled blacks in a black youth's neighborhood, the less likely he or she was to have a job, holding the usual characteristics constant. The poorer job access that black neighborhoods suffer from, in Ihlanfeldt's calculation, accounts for about a third of the rather substantial difference in employment rates between black and white teenagers living at home.

Of course spatial mismatch is not the only possible interpretation of Ihlanfeldt's result. Better job access should lead to higher rents, and higher rents should lead to more employed teenagers living with their families rather on their own. But when more employed teenagers live with their families rather than on their own, that raises the employment rate among teenagers living at home. So Ihlanfeldt may be reporting on the housing market and household formation, not the labor market.

Raphael (1998) comes to a conclusion very similar to Ihlanfeldt's but with a different measure of job accessibility. Like Ellwood and Ihlanfeldt, he looks at out-of-school youths, sixteen to nineteen years old, living at home; his data are from the San Francisco Bay Area in 1990. His measure of job accessibility is the increase in employment in the area near a youth's home, not the level of employment. The argument is that new jobs are more important than existing jobs for people entering the labor force; turnover in existing jobs matches job vacancies with newly unemployed workers. He finds that living close to new jobs increases teenage employment, and that about a third of the otherwise unexplained advantage that white teenagers have in being employed disappears when proximity to new jobs is considered. Raphael does not find that proximity to new jobs explains much about why employment rates differ among black neighborhoods, but his sample may be too small and have too little variability to answer this question.

One theoretical weakness is that Raphael's formulation ignores workers who retire, die, or withdraw from the labor force for other reasons. This seems to be an important source of new job opportunities: new Social Security benefits were awarded to 1.7 million retired workers in 1999, for instance, although total employment in that boom year grew by about 1.9 million (U.S. Social Security Administration 2001, p. 12; U.S. Bureau of the Census 1999, table 662).

The differences among the participants in the debate about the spatial mismatch hypothesis tend to obscure the points of agreement, however, and for many purposes these are far more important than the areas of disagreement. There are two main points on which they agree.

The first is that, so far as getting jobs or keeping jobs is concerned, blacks gain nothing from where they live. Maybe they lose, but no one contends that they gain.

The second point is that, for commuting to the jobs they actually have and are most likely to get, blacks live in the wrong places, or at least in places

that are substantially worse than the places where comparable whites live. Were blacks to live in the same neighborhoods that whites with the same jobs and income live in, they would commute less.

Together these points suggest that black neighborhoods should have lower rents to make up for their poorer job access. If race didn't matter in the housing market and minorities could live anywhere they wanted to and could afford to, this would clearly be the case. But we have seen that race matters.

4. Child Development and Education

Another possible benefit of the current residential pattern is that minority children might be able to grow and develop their capabilities more easily and more fully if they are in a more hospitable environment, where they don't have to confront so often the stereotypes and prejudices of whites, and they can encounter positive role models frequently and see that success is possible for people like themselves if they work hard enough. This is a benefit of segregation, not centralization.

Some evidence supports the idea that segregation enhances minority achievement. Constantine (1995) finds that black students who attend historically black colleges and universities (HBCUs) earn substantially more after graduation than comparable students who attend other colleges. HBCUs have high proportions of black students, faculty, and administrators, and an atmosphere in which black success and achievement are clearly visible.

No study has yet shown a sizable effect for role models. Almost all studies that have been done, however, have asked how the presence or absence of female faculty influences the decisions or achievements of female college or graduate students; these are reviewed in Holzer and Neumark (2000). Little is known about how minority mentors and role models affect minority students, although many theories imply that the situation of minorities is similar to that of women. Thus it is not clear whether it is role models or some other constituent of what HBCUs offer that is responsible for the greater success that their students enjoy.

For residential segregation to contribute to better socialization experiences for minority children, moreover, it would have to create neighborhoods where poor children are more likely to come into contact with successful and educated minority adults than they otherwise would be. In other words, racial segregation would have to promote class integration within the minority community. There is no evidence that it does this. In fact, greater racial segregation in a metropolitan area is usually associated with greater economic segregation within the African American community; the correlation between the two measures is fairly high (Cutler and Glaeser 1997). Residential racial segregation doesn't produce neighborhoods that resemble HBCUs.

Despite the positive evidence for at least one kind of segregation, most studies indicate that residential racial segregation in U.S. cities harms African

American children who grow up in those neighborhoods. Cutler and Glaeser (1997) used 1990 census data on young adults to study how segregation affected the probability of finishing high school, finishing college, being idle (not being employed, in school, or in the military), and becoming an unmarried mother. Blacks in more segregated cities were considerably less likely to finish high school and more likely to be idle or to be an unwed mother than blacks in less segregated cities; segregation made no difference for whites. (Segregation also hurt college graduation, but the effect was smaller and sometimes not statistically significant.) The effect was big: if segregation were eliminated entirely, the black-white difference in high school graduation and idleness would disappear, as would two-thirds of the difference in rates of unwed motherhood. Cutler and Glaeser conducted a number of tests to see whether the bad outcomes were causing segregation, rather than the reverse, and they concluded that causality runs in the direction they thought it did—segregation causes bad outcomes.

Since segregation is correlated with centralization, and centralization hurts job access, one possibility is that poor job access is causing the bad outcomes, not any difficulties with child development in segregated neighborhoods. Cutler and Glaeser checked for this possibility as well, and found that job access is one reason why segregated cities produce worse outcomes, but only a minor reason.

O'Regan and Quigley (1996) also find that segregation hurts young people. According to their 1990 data, minority youths between the ages of sixteen and nineteen who were living at home were less likely to have a job if they lived in a metropolitan area with a high index of isolation. Living in neighborhoods with white people helps; living in neighborhoods with minorities or poor people hurts. (O'Regan and Quigley held constant a large number of personal and labor market characteristics.) If the average metropolitan area were completely integrated, about half of the employment gap between white and minority youths would be eliminated.

Hanushek, Kain, and Rivkin (2002) look more specifically at the relationship between school racial composition and test scores in Texas. Everything else being equal, they find that black children do more poorly in schools that have a higher proportion of black students. This effect is concentrated among students who come from the most advantaged backgrounds—black children from better homes do more poorly in school if they are exposed to more black classmates.

Finally, Ellen (2000a) finds that segregation hurts young black children even from the minute they are born. Using essentially the same methods as Cutler and Glaeser, she shows that African American mothers in more segregated cities are more likely to give birth to low-birth-weight babies than African American mothers in less segregated cities. Segregation has little or no effect on white mothers. Eliminating segregation would remove about 40 per-

cent of the difference between black and white mothers in their proportions of low-birth-weight children. According to Ellen (p. 203), "Low birth weight is a major cause of infant mortality and is associated with greater childhood illness and such developmental disorders as cerebral palsy, deafness, blindness, epilepsy, chronic lung disease, learning disabilities, and attention deficit disorder."

Thus, something about segregation seems to make bearing and raising children successfully more difficult for African American families. Since opportunities for child rearing are an important part of the value of a house, this is another reason why trying to figure out whether minorities pay more for housing is extremely difficult.

All four sets of authors try to specify the mechanism by which segregation causes this harm, but they are more successful at ruling out possibilities than ruling them in. Socialization, negative role models, poor information flows from and about the larger society are frequently cited mechanisms, but there is no strong evidence to support these possibilities—just a lack of support for other stories that link segregation to outcomes. Another possible mechanism—similarly lacking in empirical support—is that segregation operates on white people, not black: white people who grow up with little contact with blacks may be more likely to harbor negative stereotypes and be more inclined to engage in discriminatory actions. The harsher environment that these attitudes create for black people may lead to worse outcomes. Not all interracial contact produces greater harmony, but social psychologists have found that sustained, personal, informal, noncompetitive contact on an equal-status basis reduces prejudice (Jackman and Crane 1986). More directly, living in integrated neighborhoods appears to lead to greater interracial contacts and friendships (Sigelman et al. 1996).

5. Politics

Politics is the final domain in which the current residential pattern may help minorities. Through both segregation and centralization, this pattern may give minorities more control than they would otherwise have over governments that can help them.

The U.S. electoral system is a winner-take-all system, and so in theory those who lose by a little fare no worse than those who lose by a lot. If minorities were spread evenly throughout a metropolitan area, they would control no district or town, and governments would treat them poorly everywhere. Segregation should help minorities by assuring that they are winners in those districts and jurisdictions that they dominate, even if they are losers elsewhere.

Some evidence supports this view. Black mayors are much more likely to be elected in jurisdictions where the majority of voters are black than in other jurisdictions, and black-owned businesses do much better in metropolitan ar-

eas where the mayor of the largest city is black. Bates (1997) shows that, during the 1980s, black-owned businesses were more likely to start up in those metropolitan areas, everything else being equal, and total sales of black-owned businesses were greater.

Still, the evidence supporting political benefits from segregation is meager. The two most crucial local government services are the police force and education. Minority political leadership does influence the racial composition of police departments, as well as police officers' attitudes toward civilians, but Lovrich and Steel (1983) found no evidence that racial composition had any impact on any measure of crime or safety. Lott (2000) argued that minority political control and minority police representation increased crime, but his findings have been seriously questioned (Holzer and Neumark 2000). Donohue and Levitt (1998) found that having police officers of the same race as the citizens cut property crime but had no measurable effect on violent crime.

On education, recall that Cutler and Glaeser (1997) found that segregation reduced high school graduation rates (and maybe college graduation rates, too). So whatever the positive effects segregation might have on the politics of educational attainment, they are not big enough to offset the negative effects of other aspects of segregation.

There is no evidence I am aware of, pro or con, on the political benefits of centralization.

Arguments about political gains and losses, moreover, presuppose the set of political rules that are in place now in U.S. metropolitan areas—the way jurisdiction boundaries are drawn, the limitations on the local franchise, strict majority rule, and so on. But these are not immutable conditions. For instance, jurisdictions could be bigger (metropolitan government) or smaller (neighborhood government); people could be allowed to vote on local matters in more than one jurisdiction; representatives could be elected through proportional representation or approval voting. It seems more reasonable to adapt political rules so that they work well when people live where it makes sense for them to live, rather than to force people to live in places they would not otherwise live just so an arbitrary historical set of rules does not go seriously haywire.

6. Summary of Consequences for Minorities

We have not found large benefits for minorities, on average, from the current residential pattern in U.S. cities. Centralization hurts job access, resulting in both higher commuting costs and worse chances of getting jobs. Segregation seems to make the successful raising of children more difficult. No benefits from psychic togetherness, business development, or political power have yet been documented that would offset these losses. Most of the benefits of segregation, moreover, seem to be realizable from decentralized neighborhoods of

10,000 or 20,000 people. Centralized agglomerations with hundreds of thousands of minority residents seem to have no rationale whatsoever.

That the average minority person probably loses from centralization and segregation does not imply that every minority person does. It does not imply that everyone should immediately give up the advantages of a neighborhood they have come to think of as their home. Nor does it even imply that changing the current pattern would be a potential Pareto improvement; we'll have to see how much whites benefit from the current pattern before reaching any such conclusion. But the weight of the evidence indicates that conscious efforts at maintaining the current pattern seem to hurt minorities on average.

B. Consequences for Whites

1. Child Development and Education

Perhaps white kids develop better when they have less contact with minorities. Segregation might curtail fighting and bickering. In addition, one possible interpretation of Cutler and Glaeser's (1997) results (not one that they endorse) for blacks is that something about African American culture inhibits healthy child development. If this were so, then whites less exposed to this culture would be more likely to graduate from high school and less likely to be idle or to become unwed mothers. The distinct unwillingness of many white parents to send their children to overwhelmingly minority schools might arise from a belief in such dangers.

Cutler and Glaeser, however, test for this effect, and they don't find it. Young white adults in more segregated metropolitan areas have slightly better outcomes than otherwise similar young white adults in less segregated places, but the effect of segregation is tiny and often in the opposite direction when other characteristics are held constant. Similarly, Hanushek, Kain, and Rivkin (2002) in examining test scores in Texas schools, don't find any significant effect of the proportion of black classmates on white achievement. Angrist and Lang (2004) find that busing in black students from Boston did not reduce the test scores of white students in suburban schools outside of that city. Contact with blacks doesn't appear to harm white kids.

Although they don't gain anything in test scores from avoiding contact with minorities, white kids could quite possibly lose out morally and culturally from their isolation. Hispanic and African American cultures are among the most vibrant in the world, and African Americans are the most religious group of people in the developed world. Avoiding contact with such people may not be a good way to raise children.

2. Commuting

If centralization makes minorities commute more, does it let whites commute less? The answer is probably no; on average, whites probably commute more than they would if black residences were more integrated and less centralized.

The reasoning is fairly simple. Many whites live in the suburbs and work in the CBD. Their commute requires them to pass through minority neighborhoods. If minority neighborhoods were concentrated on the outskirts of metropolitan areas rather than in their centers, many whites would have shorter commutes. When South African whites designed apartheid to suit their purposes, they put the black townships like Soweto far from the centers of their cities.

On the other hand, whites who work in the suburbs might be forced to commute longer if minority neighborhoods were less centralized. Overall, it's impossible to say whether whites would have longer or shorter commutes without specifying precisely where minorities would live in some alternative to the current system.

We can make a stronger statement, however, about total commuting time and segregation. Hold the location of jobs fixed. Suppose families were segregated by location of work, not by race: all families that work in the CBD live near the CBD, all families that work in the suburbs live near their jobs in the suburbs. Then no one commuting to work in one direction ever passes anyone else commuting to work in the opposite direction. This pattern clearly minimizes aggregate commuting cost and, since minorities work everywhere, reduces both segregation and centralization. Thus, relative to at least one plausible alternative with less centralization and less segregation, if the current pattern provides any reduction in commuting cost for whites (which it probably doesn't), this reduction is more than offset by an increase in commuting time for minorities.

3. Cheaper Labor

Spatial mismatch implies that the CBD is surrounded by large numbers of minority workers who don't have good job alternatives elsewhere. This should lower wages in the CBD, drive up minority employment, and increase land values. CBD landowners should be the big winners from spatial mismatch.

But the centralization of minority residences cuts both ways for CBD landowners. The more minority people there are living close to the CBD, the fewer white people there are; that's a principle from physics, holding density constant. The value of land that white people occupy is constrained from falling too far by the wages offered by suburban employers, and so the wages that CBD employers must offer white workers are greater the longer they must commute, and the more they think of the CBD as an alien, minority environment.

Centralization thus raises the wage CBD employers have to pay whites, even while it reduces the wage they have to pay minorities. Which effect is stronger?

Mills and Price (1984) and Mills (1985) looked at this question and concluded that centralization of minority residences hurts the CBD. Moving whites away was worse than moving minorities near. They measured central-

ization in a metropolitan area by the ratio of the proportion of blacks in the central city population to the proportion of blacks in the suburban population. Metropolitan areas where the black population was more centralized had CBDs that lost employment share more rapidly than metropolitan areas where black population was less centralized. A cheap minority labor force failed to attract jobs to the CBD—apparently the greater difficulty of attracting white workers offsets any gain this could have created.

4. Open Space and Sprawl

Whites might gain if segregation and centralization force minorities to occupy less land than they would otherwise occupy. Reducing segregation might induce minorities to buy more land, increase the size of metropolitan areas, and cut whites off from open space they would otherwise enjoy.

Segregation and centralization are almost surely connected with minorities' low homeownership rates, and homeowners use more land per household than renters do. It may be no coincidence that some of the metropolitan areas where sprawl is greatest (for example, Tucson and Sacramento) are among the areas with the least segregation; and vice versa (New York, Chicago).

The relationship between sprawl and segregation, however, is by no means so simple. First suppose that job locations are held constant as minority residences decentralize. The places minorities now live in would most likely be taken over by whites who work in the CBD. If these white households decreased their land consumption by more than minorities increased theirs, the result would be a net decrease in the land used in the metropolitan area.

Whites moving closer to the CBD might decrease their land consumption for several reasons. They would be moving to more valuable land than they formerly occupied, and location theory tells us that this greater value will give them an incentive to economize on their use of land. Moreover, the neighborhoods near the CBD would have a higher concentration of CBD workers than their old neighborhoods, and this concentration might permit more frequent and cheaper mass transit.

Finally, if we permit job locations to vary, too, we see another reason why alternatives to the current pattern might increase overall density. Decentralizing the minority population would increase the proportion of jobs in the CBD, if Mills and Price are right. This would increase the movement of whites to neighborhoods close to the CBD, and discourage the residential expansion of the metropolitan area at great distances from the CBD. Since land is more expensive in the CBD, moreover, businesses located there are likely to use land more intensively than they would if they were in the suburbs. Thus residential decentralization might also reduce total land use by businesses.

Do all of these secondary reductions add up to more than the first-order increase in land use that would ensue if minorities lived like other Americans of the same age and income? It's hard to say; no serious study has been done

on this question. It doesn't seem, though, that whites derive huge benefits from any open space that the centralization and segregation of minorities might produce.

5. Togetherness and Culture

Another possible benefit that whites could derive from their segregation is the warm comfort of being among their own kind, the ability to be themselves without pressure from outsiders, the opportunity to experience and develop their own culture. White Americans, like any other group, might need a place to call their own. Notice that this is an argument for segregation, not for the centralization of minorities.

This argument, however, has some serious weaknesses. First, isolation from minorities is not something most white Americans say they want. Even as long ago as 1978, 88 percent of white people said blacks have a right to live anywhere they want to, and 86 percent said they would not move if "a black person came to live next door." In 1976, only 7 percent of whites in the Detroit area (one of the most segregated metropolitan areas) said they would try to move out of a neighborhood that was 8 percent black (Massey and Denton 1993, pp. 91–93).

Second, white Americans are quite numerous and are the richest people in the world. They hold the top positions in many of the world's most powerful organizations, and control many of the most important media outlets. Does their culture really need the meager protection that residential segregation provides?

6. Neighborhood Preservation

Many people are attached to the neighborhoods in which they live. The streets and parks carry many sweet memories (and some bitter ones, too); their oldest friends live there, and their relatives; and even people they don't consider friends they have grown used to, come to understand, and worked out pleasant ways of interacting with. They know the people they pass on the street every day, and those people understand what they have accomplished. Not everybody values neighborhoods this way, but some people do. I'll call them "neighborhood preservationists."

For these people, a large, abrupt change in the physical characteristics of a neighborhood or in its population is a major loss. Even if you continue to live in the same house, your friends and acquaintances are gone; the place looks different. If racial transition causes rapid changes in a neighborhood, then neighborhood preservationists become worse off. To the extent that the current residential patterns protect neighborhoods from abrupt changes, neighborhood preservationists are better off.

Thus the preservationists in a particular white neighborhood clearly realize benefits if various defenses succeed in keeping that neighborhood from

changing quickly from almost all white to almost all minority. (Since preservationists have strong incentives to be neighborhood activists as well, this may be partly why such defenses get erected.) But the actual benefits to the preservationists of the current situation in an entire metropolitan area are considerably less than the sum of the benefits preservationists expect in particular neighborhoods. In fact, preservationists would probably be better off with integration.

There are two reasons why the actual whole is less than the sum of the expected parts. The first is that, since about 1970, racial transition usually hasn't happened quickly. This is apparent in Ellen's data. Racial transitions may have taken place much more quickly while the Great Migration was still occurring, and many people still think "white flight" is a plausible expectation, but the fears of neighborhood preservationists that they will lose all their friends quickly are not well founded. The presence of minorities causes only a very modest increase in the rate at which white people move out of neighborhoods (although it causes a big decrease in the rate at which they move in). Since in any normal neighborhood people move out all the time and those who remain are always being forced to make new friends and acquire new acquaintances, the burden of racial transition is no longer large. A slightly accelerated rate of turnover and, possibly, a greater difficulty in making friends with people of a different race are real costs, but they're not huge.

The second reason is that racial transition is a problem only because metropolitan areas are segregated. The more neighborhoods that erect barriers to minority residence, the greater the pressure on the neighborhood with the least effective barriers. Segregation, by concentrating minority demand on only a few neighborhoods at a time, makes racial transition in those neighborhoods both swifter and more complete than it would be if most neighborhoods were open to minorities. By creating a situation in which few whites actually live in neighborhoods with substantial numbers of minorities, segregation causes whites to avoid integrated neighborhoods, first by stigmatizing those who live in integrated settings ("They must not care much about their kids if they send them to *those* schools"), and second by affording whites who want to avoid minorities many attractive options for doing so. In an integrated metropolitan area, racial transition would not be a major problem. All neighborhood preservationists would be more secure if none tried to be secure.

7. Summary of Consequences for Whites

The review of the consequences for whites of segregation and the centralization of minorities has necessarily been much sketchier than our review of the same question for minorities. Social scientists have devoted amazingly little effort to trying to find out how segregation and centralization affect the largest and most powerful group in U.S. society.

The current pattern doesn't seem to help white kids in the aggregate, reduce sprawl, or cut commuting times; in fact, it may hurt. Plus, CBD land values are probably hurt.

On the other hand, some whites place great value on avoiding minority neighborhoods and protecting the neighborhoods they know. Given current patterns, segregation and centralization are the best way they know of, working independently, to achieve those goals. That's where the large benefits of segregation for whites come from (centralization doesn't seem to serve any purpose for these goals). But there seem to be alternative arrangements of metropolitan residences that, combined with some education, would allow whites to realize most of their goals. No small group of white people, acting alone, however, could make the changes necessary to effectuate these alternative arrangements. Like drivers stuck in a traffic jam, anyone who acts differently alone will be worse off, but everyone would be better off if everyone acted differently.

IV. Conclusion

Diversity should make living in cities more productive and more interesting. But in the United States today, the way people live doesn't allow enough of the advantages of diversity to be realized. The costs of current racial patterns are significant. In the next chapter we will look at what should be done.

Questions

1. Several years ago a child custody case revolved around the merits, or lack thereof, of integration. The custodial parent and the child were living in South Orange, New Jersey, and the child was in a school system in which approximately 40 percent of the students were African American. The noncustodial parent was living in Chatham, about eight miles away, where the proportion of African Americans in the school system was well below 10 percent. Chatham and South Orange were on average about equally wealthy, and the adult populations were similar in educational attainment. Average and median scores on standardized achievement tests, however, were lower in South Orange. The parenting qualities of the two contending individuals were not significantly different. The noncustodial parent argued in part that the Chatham school system would be better for the child.

    How would you assess the relevance of the argument? How would you

decide which parent to award custody to? How, if at all, does your answer depend on the race or gender of the child (which I have not specified)? Does it depend on the races of the contending parents? What are the public policy implications of your decision?

2. Left-handed and right-handed people live in Directional City. Directional City has three neighborhoods, and in the 2000 census they had the following populations:

Neighborhood	No. of lefties	No. of righties	Total pop.
1	1,990	10	2,000
2	100	900	1,000
3	10	1,490	1,500
Total pop.	2,100	2,400	4,500

a. Find the index of dissimilarity between lefties and righties.
b. Find the index of isolation for lefties.
c. Find the index of isolation for righties.
   In 2010, the census reports the following populations:

Neighborhood	No. of lefties	No. of righties	Total pop.
1	1,990	10	2,000
2	60	940	1,000
3	50	1,450	1,500
Total pop.	2,100	2,400	4,500

d. How does the index of dissimilarity change?
e. How do the indexes of isolation change?
f. Which measure better reflects the segregation over the decade? Why?

3. Most literature on housing-market segregation and discrimination concentrates on African Americans. Is this concentration appropriate? Why or why not?

4. Suppose a city consists of blacks and whites. White household income is distributed uniformly between 0 and 100; black income, between 0 and 60. There are three schools: two that whites attend and one that blacks attend. Racial segregation is perfect, and among whites income segregation is also perfect. The two white schools are of equal size.
a. What is the average income in the black school?
b. What is the average income in the poor white school? The rich white school?
c. Consider only kids from households below the white median income. What is their average income in the school that white kids attend? What is their average income in the school that black kids attend?
d. Consider only kids from households above the white median income.

What is their average income in the school that white kids attend? What is their average income in the school that black kids attend?

    e. Since peer effects matter, would you expect a bigger racial test-score gap among richer children or poorer?

References

Angrist, Joshua, and Kevin Lang. 2004. "Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program." Forschunsinstitut zur Zukunft der Arbeit (IZA) Discussion Paper 976.

Bates, Timothy. 1997. *Race, Self-Employment and Upward Mobility: An Illusive American Dream.* Washington, DC, and Baltimore: The Woodrow Wilson Center and Johns Hopkins University Press.

Boehm, Thomas P., and Alan M. Schlottman. 2004. "The Dynamics of Race, Income, and Homeownership." *Journal of Urban Economics* 55(1):113–130.

Cameron, Steven, and James Heckman. 2001. "The Dynamics of Educational Attainment for Black, Hispanic, and White Males." *Journal of Political Economy* 109(3):455–499.

Chambers, Daniel. 1992. "The Racial Housing Price Differential and Racially Transitional Neighborhoods." *Journal of Urban Economics* 32:214–232.

Constantine, Jill M. 1995. "The Effect of Attending Historically Black Colleges and Universities on Future Wages of Black Students." *Industrial and Labor Relations Review* 48:531–546.

Cutler, David, and Edward Glaeser. 1997. "Are Ghettoes Good or Bad?" *Quarterly Journal of Economics* 112(3):827–871.

Cutler, David, Edward Glaeser, and Jacob Vigdor. 1999. "The Rise and Decline of the American Ghetto." *Journal of Political Economy* 107(3):455–506.

Denton, Nancy A., and Douglas Massey. 1989. "Racial Identity among Caribbean Hispanics: The Effect of Double Minority Status on Residential Segregation." *American Sociological Review* 54:790–808.

DiPasquale, Denise, and Matthew Kahn. 1999. "Measuring Neighborhood Investments: An Examination of Community Choice," *Real Estate Economics* 27:389–424.

Donohue, John J. III, and Steven D. Levitt. 1998. "The Impact of Race on Policing, Arrest Patterns, and Crime." National Bureau of Economic Research Working Paper 6784.

Ellen, Ingrid Gould. 2000a. "Is Segregation Bad for Your Health? The Case of Low Birth Weight." *Brookings-Wharton Papers on Urban Affairs* 1:203–238.

———. 2000b. *Sharing America's Neighborhoods: The Prospects for Stable Racial Integration.* Cambridge, MA: Harvard University Press.

Ellwood, David. 1986. "The Spatial Mismatch Hypothesis: Are Teenage Jobs Missing in the Ghetto?" In R. B. Freeman and H. J. Holzer, eds., *The Black Youth Employment Crisis,* pp. 147–187. Chicago: University of Chicago Press.

Fletcher, Michael. 2001. "Long Division." *The New Crisis* (September–October):26–31.

Freeman, Lance, and Darrick Hamilton. 2002. "A Dream Deferred or Realized: The Impact of Public Policy on Fostering Black Homeownership in New York City Throughout the 1990s." *American Economic Review* 92(2):320–324.

Gablinger, Ynon. 2003. "The Extent and Nature of Discrimination in Housing." PhD diss., Columbia University.

Gabriel, Stuart A., and Stuart S. Rosenthal. 1996. "Commutes, Neighborhood Effects, and Earnings: An Analysis of Racial Discrimination and Compensating Differentials." *Journal of Urban Economics* 40:61–83.

Hanushek, Eric, John Kain, and Steven Rivkin. 2002. "New Evidence about *Brown v. Board of Education:* The Complex Effects of School Racial Composition on Achievement." National Bureau of Economic Research Working Paper 8741.

Holzer, Harry, and David Neumark. 2000. "Assessing Affirmative Action." *Journal of Economic Literature* 38(3):483–568.

Ihlanfeldt, Keith R. 1992. *Job Accessibility and the Employment and School Enrollment of Teenagers.* Kalamazoo, MI: Upjohn Institute.

Jackman, Mary R., and Marie Crane. 1986. "'Some of My Best Friends Are Black . . .': Interracial Friendship and Whites' Racial Attitudes," *Public Opinion Quarterly* 50:459–486.

Kain, John, and John Quigley. 1975. *Housing Markets and Racial Discrimination.* New York: National Bureau of Economic Research.

Leonard, Jonathan. 1985. "Space, Time, and Unemployment." Unpublished manuscript. University of California, Berkeley.

Lewis, David Leavering, ed. 1995. *W. E. B. Du Bois: A Reader.* New York: Henry Holt.

Lewis Mumford Center for Urban and Regional Research. 2001a. "Ethnic Diversity Grows, Neighborhood Integration Lags Behind." State University of New York at Albany. Accessed on September 23, 2002, at *http://mumford1.dydns.org/cen2000/report.html.*

———. 2001b. "Separating the Children." State University of New York at Albany. Accessed on September 23, 2002, at *http://mumford1.dydns.org/cen2000/report.html.*

Lott, John R. Jr. 2000. "Does a Helping Hand Put Others at Risk? Affirmative Action, Police Departments and Crime." *Economic Inquiry* 38(2):239–277.

Lovrich, Nicholas, and Brent Steel. 1983. "Affirmative Action and Productivity in Law Enforcement Agencies." *Review of Public Personnel Administration* 4:55–66.

Massey, Douglas S., and Nancy Denton. 1993. *American Apartheid: Segregation and the Making of the Underclass.* Cambridge, MA: Harvard University Press.

Meyer, Bruce. 1990. "Why Are There So Few Black Entrepreneurs?" National Bureau of Economic Research Working Paper 3537.

Mieszkowski, Peter. 1979. "Studies of Prejudice and Discrimination in Urban Housing Markets." Special study, Federal Reserve Bank of Boston.

Mills, Edwin. 1985. "Open Housing Laws as Stimulus to Central City Employment." *Journal of Urban Economics* 17:184–188.

Mills, Edwin, and R. Price. 1984. "Metropolitan Suburbanization and Central City Problems." *Journal of Urban Economics* 15:1–17.

O'Regan, Katherine, and John Quigley. 1996. "Teenage Employment and the Spatial Isolation of Minority and Poverty Households." *Journal of Human Resources* 31:692–702.

Raphael, Steven. 1998. "The Spatial Mismatch Hypothesis and Black Youth Joblessness: Evidence from the San Francisco Bay Area." *Journal of Urban Economics* 43:79–111.

Raphael, Steven, and Michael Stoll. 2001. "Can Boosting Minority Car-Ownership Rates Narrow Inter-Racial Employment Gaps?" *Brookings-Wharton Papers on Urban Economic Affairs* 2.

Sethi, Rajiv, and Rohini Somanathan. 2001. "Racial Income Disparities and the Measurement of Segregation." Barnard College Working Paper 01-07.

Sigelman, Lee, et al. 1996. "Making Contact? Black-White Social Interaction in an Urban Setting." *American Journal of Sociology* 101(5):1306–1332.

Singer, David, ed. 1999. *American Jewish Yearbook.* New York: American Jewish Committee.

U.S. Bureau of the Census. 1998. *Statistical Abstract 1998.* Washington, DC: Government Printing Office.

———. 1999. *Statistical Abstract 1999.* Washington, DC: Government Printing Office.

U.S. Social Security Administration. 2001. "Fast Facts and Figures about Social Security." Accessed at *http://www.ssa.gov/policy/docs/chartbooks/fast_facts/2001/index.html.*

Wachter, Susan M., and Isaac F. Megbolugbe. 1992. "Racial and Ethnic Disparities in Homeownership." *Housing Policy Debate* 3:333–370.

Yinger, John. 1995. *Closed Doors, Opportunities Lost: The Continuing Costs of Housing Discrimination.* New York: Russell Sage Foundation.

Race and Policy

American cities are not as pleasant and productive as they could be, because of how people treat race and ethnicity. That's what we saw in Chapter 11. Different policies could probably produce a potential Pareto improvement. Before we look for those policies, however, we need to ask why the current situation prevails. Remedies proceed best from an accurate diagnosis of the underlying malady.

The easy answer to the question is "racism," but it's not a very informative answer. Since these problems wouldn't exist unless people recognized different races, this answer is correct, but it doesn't point in an obvious way to a potential Pareto improvement. Americans are likely to believe in races for a long time to come.

Nor is laying the blame on slavery particularly helpful. True, if there had been no slavery in this country, it's unlikely that the United States would now be experiencing these racial problems. But deplore it as we might, there is nothing we can do now to undo the historical fact that slavery existed (reparations may or may not assuage resentment or guilt, but they are unlikely, say, to make white home buyers flock to a black neighborhood like Brooklyn's Bedford-Stuyvesant).

Simply answering "discrimination" or "whites dislike minorities" is not adequate either. This sort of answer is just the middle of a story without a beginning or an ending.

The beginning that's missing from this story is an explanation for why whites should have such preferences. When we deal with ordinary goods, economists don't usually ask about the source of preferences, but races and ethnic groups aren't ordinary goods. They are sets of people with little in

---

To get the most out of this chapter, you should be familiar with these concepts: centralization, the closed-city case, consumer surplus, equilibrium, fixed cost, index of dissimilarity, the open-city case, Pareto improvement, Pigouvian subsidies, potential Pareto improvement, public goods, sprawl, and zoning. You can find these terms in the Glossary.

common except various visible features that have little or no productive significance. Why anyone would have strong feelings about such abstractions is a question that needs answering in a coherent story about **discrimination.** I'm not denying that some people have such feelings; I'm just maintaining that any story that assumes such feelings without explaining them is incomplete.

The ending that needs to be included in a story about discrimination is an account of how ill feelings harm the people they are directed toward. Being hated doesn't always make a person worse off. No matter how qualified I am, no one is going to hire me, an American, for any kind of responsible job in Iran, Myanmar, or North Korea. This prejudice has not hurt my career in any way. Similarly, large numbers of my friends despise the British royal family and consider it an example of hereditary subnormality, but those beliefs have not made the royal family worse off. Thus a story about discrimination is incomplete unless it describes how animosity translates into harm.

In the first section of this chapter, I will ask how we can explain the pattern of segregation, centralization, and poor education. Why don't more whites move to minority neighborhoods or more minorities move to white neighborhoods? Why don't minorities do better in school? No solid, well-accepted, comprehensive theory exists, but we will attempt some answers.

I. Causes

A. Whites in Minority Neighborhoods

Several good reasons can be cited for why more whites don't move to minority neighborhoods. Notice, first, that we don't have to explain why no whites would ever, under any conditions, move to these neighborhoods; we have to explain only why enough minorities would be willing to bid more to live there than almost all whites. If a neighborhood has 100 houses, and the hundredth most eager minority household is willing to pay $500 a month to live there while the most eager white household is willing to pay $499, no whites at all will live in the neighborhood. But since there are many more white households than minority households and since minority neighborhoods have features—propinquity to the CBD, especially—that should be appealing to at least some whites, we still need to explain why so few whites are willing to pay more than minorities to live in minority neighborhoods.

One reason might be hostility: some whites dislike minorities and would be willing to pay a premium not to live near them. It is no doubt true that many whites hold such antagonistic attitudes toward minorities. Almost a quarter of white respondents to a Detroit-area survey in 1978, for instance, said they would feel uncomfortable in a neighborhood that was as little as 8 percent black, and 27 percent said they would be unwilling to move into such a neighborhood (Massey and Denton 1993, p. 93). In 1996, 13 percent of

whites surveyed agreed with the statement that whites had a right to keep blacks out of their neighborhoods, and 33 percent disapproved of racial intermarriage, with 13 percent favoring laws to ban it (Schuman et al. 1997, p. 107). But it is hard to see how simple animosity could be so pervasive among whites that no appreciable number would move to minority neighborhoods if prices there for the same quality of housing were lower than in white neighborhoods.

Think about a typical U.S. metropolitan area in which 15 percent of the population is African American. This is the average in the data used by Ellen (2000). In this metropolitan area, 46.1 percent of African American households are in majority white neighborhoods. This leaves 53.9 percent of the African American population, which represents 8.1 percent of the total population, living in majority African American neighborhoods. If whites representing just 2.7 percent of the total population, or 3.2 percent of the white population, moved there, those neighborhoods would have a population that was one quarter white, and they would be integrated. For pure hostility to work, if minority housing prices are lower than white, more than 96 percent of whites would have to dislike African Americans so much that they would pass up a bargain.

Survey data do not support this idea, however. In the early 1990s almost a third of whites said they were willing to live in a neighborhood whose population was half black. In 1990, 66 percent of whites said they would have no objection to sending their children to a school that had a majority of black students, and in 1996, 75 percent said they would not move if black people in great numbers lived in their neighborhood (Schuman et al. 1997, p. 143). In 1992, 11 percent of whites said they strongly favored the use of quotas to help blacks in college admissions, 6 percent strongly favored preferential hiring, and 23 percent favored preferences in hiring and promotion in cases where there had been job discrimination in the past (ibid., pp. 175–176). All of these percentages are considerably greater than the 4 percent of whites that would be required to integrate all minority neighborhoods.

Another possible explanation for why so few whites are willing to outbid minorities in minority neighborhoods is that whites hold negative stereotypes about minority neighborhoods (or about minorities). Ellen (2000) provides some evidence for this explanation: those whites who are least likely to use neighborhood public goods—renters, households without children—are least sensitive to racial composition. Many whites—more whites than harbor antipathy per se—may believe that minority neighborhoods (possibly through no fault of their residents) inevitably have more crime, worse schools, dirtier streets, more noise, less prestige. Whites who don't hold these stereotypes but think other people they care about do—employers, relatives, friends, and potential friends—will also feel reluctant to spend a lot of money for houses in minority neighborhoods. Thus we can delineate another large set of whites not willing to bid much to live in minority neighborhoods.

This explanation, however, also has a problem, though maybe not a fatal one. Minorities and whites live in the same country, watch the same television shows, and read the same newspapers. If whites are prejudiced against minority neighborhoods, or worry about other people who are prejudiced, then minorities should be prejudiced too, and worried about other prejudiced people too. If whites are no more prejudiced (at first or second remove) than minorities are, then prejudice gives no reason for segregation.

Is there reason to think that minorities are less likely to hold negative stereotypes, or to worry about others who might? Possibly, although there is no evidence on this question. Minorities may rely on direct personal contact and experience for forming their beliefs about minorities, and less on the media, which may have a bias toward carrying more bad news than good. The question, though, is whether different sources of information can support a racial gap in stereotypes large enough to support the segregation we observe. Many minorities during their lives must have received worse information about minorities and minority neighborhoods than some whites have.

A third explanation relies on trust. Trust is an important component of good neighborhoods. Leaving the key to your house with your neighbors so you don't get locked out or so they can feed the cat, reporting to them truthfully whether their children have been reasonably well behaved, simply refraining from harassing them—all of these actions help make a neighborhood a good place to live, and all of them depend on trust. The kind of trust they depend on is not just trust of particular individuals you know, but expected trust of the potential new neighbors you don't know, of the people who might move in after them, of the people down the street who walk their dog in front of your house. If whites trust minorities less than minorities trust other minorities, then they will be more reluctant to move into minority neighborhoods.

Is there reason to believe that mistrust of minorities is so pervasive among whites as to explain the high degree of segregation we observe? Possibly, especially since some of the relevant mistrust in a neighborhood deals with the *worst* person on the block, not the average. Even if everyone else is very nice, a neighborhood can be miserable if one or two people hate and harass others. It is not implausible that many whites believe that a neighborhood of 100 people is more likely to contain one anti-white person if that neighborhood is all minorities than if it is all whites.

B. Minorities in White Neighborhoods

It's unlikely that hostility toward whites is a major reason why more minorities don't live in mainly white neighborhoods. Survey data do not reveal any considerable amount of animosity on the part of minorities toward whites. In 1996, for instance, 83 percent of African Americans said they would have no objection to sending their children to a majority white school (recall that the

comparable figure for whites was 66 percent in 1990). In 1976, only 22 percent said their most preferred neighborhood mix was all black or mostly black. The comparable figure for whites then was 73 percent (Schuman et al. 1997, pp. 254–255, 140). Minorities don't hold negative stereotypes about white neighborhoods: many white neighborhoods are beautiful and safe, have great schools, carry immense prestige—and everyone knows it.

Possibly more important than minority hostility is a long history of efforts by certain whites to keep blacks and other minorities out of "their" neighborhoods. Let's review some of that history. It's not immediately relevant to current problems, because the practices in question have stopped or diminished in prevalence and ferocity, but since history has ways of affecting the present and the future—the slate doesn't get wiped clean every morning—we need to consider whether we have escaped its reach.

1. Coercion and Covenants

The role of violence in the history of segregation and centralization is very clear. In the late nineteenth century, when blacks lived mainly in the rural South, segregation in urban areas was not especially significant. In southern cities, especially, blacks and whites lived in close proximity.

During World War I, blacks began to migrate to cities, especially northern cities. That's when the urban violence started. A black boy swimming near a "white" beach, for instance, touched off a major race riot in Chicago in 1919; in six days of rioting, 38 people died, most of them black, more than 500 were injured, and a thousand were left homeless (Drake and Cayton, 1945, p. 65). Other major riots occurred in New York in 1900; Evansville, Indiana, in 1903; Springfield, Illinois, in 1908; and East St. Louis, Illinois, in 1917.

Even in 1955, race riots were possible in northern cities like Detroit. Sugrue (1996, p. 232) describes what happened when the Wilsons, an African American family, purchased a house in a predominantly white neighborhood.

> Indignant white neighbors launched a five-month siege on 18199 Riopelle
> . . . [J]ust before the Wilsons moved in, someone broke into the house,
> turned on all the faucets, blocked the kitchen sink, flooded the basement,
> and spattered black paint on the walls and floors. Later that day, after the
> Wilsons cleaned up the mess and left, vandals broke all the front windows
> in the house. Despite the noise, no one reported the attack to the police. On
> Tuesday, April 26, the Wilsons moved in. The onslaught escalated. White
> members of the Cadillac Improvement Association approached the Wilsons
> and demanded that they sell the house. That evening, someone threw a
> stone through the bathroom window. For two straight nights, the phone
> rang with angry, anonymous calls.
>
>   On Friday night, after dinner, a small crowd gathered on Riopelle Street
> in front of the Wilsons' house. They were soon joined by more than four

hundred picketing and chanting whites, summoned by young boys who rode their bikes up and down the street, blowing whistles. The crowd drew together a cross-section of neighborhood residents: as Mrs. Wilson reported, "[I]t was children; it was old people; it was teenagers; in fact, all ages were there." The demonstrators screamed epithets . . . A rock shattered the dining room window . . . The following evening, more protesters filled the street in front of the Wilson house; despite police surveillance, someone threw a large rock toward the house so hard that it stuck in the asbestos siding.

Aside from such public, communal violence, there has also been a tradition of more stealthy, private violence against minority families that have moved into predominantly white neighborhoods. Blacks moving into white neighborhoods have frequently received threatening letters and warnings of dire consequences, and experienced personal harassment as well as rock throwing, gunshots, cross burnings, and physical attacks. Between 1917 and 1921, fifty-eight black homes were bombed in Chicago alone (Massey and Denton 1993, pp. 34–35).

Whites have tried to keep blacks and other minorities out of predominantly white neighborhoods through legal action, too, as well as violence. In Chapter 8, I mentioned the laws that California communities adopted in the 1870s and 1880s banning Chinese people from taking up residence, and the laws against laundries that replaced them after the residency bans were declared unconstitutional. Similarly, around the time of World War I several cities enacted laws banning blacks from certain neighborhoods, and fifteen state courts supported those laws, but the Supreme Court held that they, too, were unconstitutional (*Buchanan v. Warley,* 1917), and this form of law never became popular.

Instead, **racial covenants** developed as the major legal impediment to black movement to white neighborhoods. A racial covenant is a provision in the deed to a property that forbids the owner from selling or renting the property to minorities (in many cases, selling or renting to Jews was also prohibited). A developer would insert racial covenants in the deeds to all the houses when he first built a neighborhood, and those covenants would be passed down from buyer to buyer. Alternatively, realtors or neighborhood associations sometimes canvassed already built neighborhoods and got most homeowners to add covenants to their deeds.

At one time, racial covenants were quite widely used. Clarke and Perlman (1948) estimated that 80 percent of deeds in some areas included restrictive covenants. Cutler, Glaeser, and Vigdor (1999) examined a sample of deeds entered into between 1908 and 1928. Two-thirds of the deeds in highly segregated cities contained resale restrictions; half of the deeds in less segregated cities contained them. The Supreme Court, however, declared racial covenants

unenforceable in *Shelley v. Kraemer* (1948). (In *Gandolfo v. Hartman* in 1891 a federal court had struck down covenants against Chinese buyers, but that decision was based on a treaty with the Chinese government and didn't create a precedent.)

What does this history of riots, firebombings, and covenants tell us about where people live in cities? For 1950 or so, it tells us pretty clearly that people were living in the wrong neighborhoods; Pareto improvements could have been achieved by reassigning them. The market did not produce the segregation of 1950.

Thus, it appears that in 1950 a lot of people were living in the wrong places. This is a conclusion about 1950, however, not about today. Racial covenants have been unenforceable for more than half a century, and while hate crimes still occur, the volume nowhere approaches that of the early part of the twentieth century. The official response is also much different today.

But there is much reason to think the history remains relevant. Segregation is not much lower now on average than it was in 1950, and a city's relative ranking in segregation in 1950 is a pretty good predictor for its relative ranking today.

Moreover, although violence in recent decades has in no way approached the levels of the 1920s or the 1950s, hate crimes, like cross burnings and graffiti attacks, as well as less blatant forms of harassment, still occur. The possibility of becoming the victim of such an attack reduces the willingness of minorities to pay to live in white neighborhoods.

2. Realtors and Landlords

The evidence that landlords and realtors treat minorities differently is substantially more current than the history of firebombs and covenants.

Numerous court cases in the past decade have found particular landlords guilty of rejecting qualified minority apartment seekers when they were accepting white apartment seekers. Even if only some landlords act this way, this practice will keep some minorities out of white neighborhoods (because their next best alternative will be in a minority neighborhood), and will dissuade others from searching in white neighborhoods, since they are less likely to be successful.

Since this behavior is illegal, the obvious question is why some landlords do it. Many of the reasons for white behavior we have already considered probably apply: personal animosity, stereotypes of minorities as bad tenants (loud, dirty, unreliable), mistrust, and fears about how current and prospective tenants would react. Because turnover is expensive and risky, landlords are neighborhood preservationists.

The most influential body of evidence on discrimination comes from a series of carefully designed "audit" experiments in the late 1980s and 1990s. In these experiments, pairs of auditors, one minority and one white, were

matched with each other and with a newspaper advertisement for a house or an apartment. The auditors were the same gender and roughly the same age, and were instructed the same way on what to tell the realtor or landlord about their income, job, education, and family status. Then they were sent to answer the ads—sometimes the minority auditor first, sometimes the white—and to record how they were treated in a number of dimensions.

We should not expect the auditors to be treated exactly the same way. In some cases the apartment may have been rented or the house sold in between the two visits (or a deal that was in place at the first visit may have fallen through before the second). The realtor may have been feeling lousy when she met the first auditor, or she may have received some very good news just before the second auditor arrived. She may have been in a hurry to take care of some personal business when she met one auditor, or especially eager to make money when she met the second. Personalities could have "clicked" with one auditor but not with the other.

But if race doesn't matter, these random events should occur just as often with minority auditors as with white; the minority auditors should have been treated more favorably as often as the white auditors were. This is not what the studies found. On most dimensions, white auditors were treated more favorably more often than minority auditors.

For instance, in a very large audit study funded by the federal government, black auditors were told 11 percent of the time that a house advertised for sale was no longer available, but their white partners were told this only 6 percent of the time. With Hispanic auditors, 9.5 percent were told the house was no longer available, as compared with 5 percent of their white counterparts (Yinger 1995, pp. 34, 44). Seventeen percent of black auditors and 15 percent of Hispanic auditors were told that an advertised apartment was no longer available, as opposed 12 percent and 7 percent of their white partners, respectively.

Race and ethnicity also matter in how much assistance realtors provide to their customers. Realtors typically try to interest people who inquire about one particular unit in as many other units as possible; this improves their chances of making a sale and collecting a commission. It is therefore interesting to note how often a white auditor was told that more units were available but the minority auditor was not. This happened in 44 percent of the black-white audits and 43.5 percent of the Hispanic-white audits. The opposite—the minority auditor being told that more units were available—occurred in 24.5 percent of the black-white audits, 27 percent of the Hispanic-white audits (the Hispanic difference is not statistically significant in this case, but all other contrasts are).

Realtors are also more likely to show minorities, especially blacks, houses in minority neighborhoods. Excluding the advertised units, the houses that blacks were shown were in more heavily minority neighborhoods than those

that whites were recommended or shown in 12 percent of the audits (Yinger 1995, p. 54).

Because realtors on average treat minorities less well, searching for housing is more costly for minorities, and so they move less often. Yinger (1995, p. 101) estimates that moving costs minority home buyers between $3,000 and $4,000 more than it costs whites and that as a result minorities move approximately 10 percent less often. (These estimates include the cost of the mortgage denials that I will discuss below.)

These costs, however, may be overestimates because of the way the studies developed the list of houses and apartments that auditors were sent to ask for. These housing units were a random sample of the units advertised in major newspapers. The audits showed that a large number of the advertising realtors did not discriminate against minorities. If minority searchers have some idea of who these nondiscriminating realtors are, they can do business only with them and avoid the costs of dealing with realtors who will not treat them well. Real searchers don't respond randomly to ads in the newspaper.

Newspaper ads have another flaw, too, that probably affects more seriously our understanding of how real estate markets work. Firms that operate primarily in minority neighborhoods are considerably less likely than other firms to advertise their properties in major newspapers. Firms that operate in both minority and white neighborhoods are less likely to advertise their minority-neighborhood properties. Thus, the average available house in the black-white audits was in a neighborhood that was 9 percent black, even though the average neighborhood in the cities where these audits were conducted was 20 percent black. Similarly, the average house in the Hispanic-white audits was in a neighborhood that was 16 percent Hispanic, although the cities where the audits were conducted were 24 percent Hispanic (Yinger 1995, p. 57).

On the one hand, this bifurcation of the real estate market confirms the notion that many minority searchers are smart enough to avoid realtors who treat them poorly, since firms that operate mainly in minority neighborhoods probably don't discriminate against minorities. The actual volume of poor treatment of minorities is probably less than that which would occur if all searchers used newspaper ads.

On the other hand, this pattern of realtor specialization gives us some insight into one mechanism by which segregation has been maintained. Firms in the audit studies were on average less likely to show blacks houses in white neighborhoods, but the differential (12 percent) is far too small to explain the high degree of segregation that actually exists. In major metropolitan areas in 1990, the difference between the proportion of whites who lived in majority white neighborhoods (98.3 percent, according to Ellen 2000, p. 30) and the proportion of blacks who lived in majority white neighborhoods (46.1 percent) was far greater than 12 percent. If all "steering" occurred at the level of

the real estate firm, metropolitan areas would be far less segregated than they are.

Segregation is maintained because firms are segregated. Because many firms that serve minority neighborhoods don't participate in multiple listing services, and few firms serve both minority and white neighborhoods, it's difficult to search in both kinds of neighborhoods at the same time. Whites don't go to firms that mainly serve minority neighborhoods. Blacks may encounter poor treatment when they go to firms that serve mainly white neighborhoods, and they are unlikely to find out much about minority neighborhoods from these firms; they may therefore be less likely to go to these firms. The real estate market thus makes it expensive for minorities to learn about houses in white neighborhoods.

3. Mortgages

Realtors are not the only participants in the contemporary real estate market who treat minorities and whites differently. Mortgage lenders, too, seem to take race into account when they approve or disapprove mortgage applications. Since houses are more expensive in white neighborhoods than minority ones and because a higher proportion of residences in white neighborhoods are owner-occupied, difficulty in getting mortgage credit translates into difficulty in moving to a predominantly white neighborhood.

Minorities who apply for mortgages are more likely to be turned down than whites are. By itself, this fact does not establish that lenders treat minorities differently: minorities on average have different incomes, assets, and credit histories, and are seeking loans on different kinds of properties. We can't say that race matters for mortgage approvals unless we can find that minority applicants are more likely to be turned down than otherwise identical whites.

A team of researchers at the Boston Federal Reserve Bank (Munnell et al., 1996) conducted the most ambitious and complete study of this question. They used a large data set that included all the mortgage applications filed by minorities in the Boston area, with all types of lenders, and a large sample of white applicants. They gathered information on every variable that lenders had told them might bear on their approval decisions. The conclusion was that race mattered considerably: a white applicant with the average characteristics of whites in the study had an 11 percent chance of being turned down; a minority applicant with the exact same set of characteristics had a 17 percent chance of being turned down.

This work has stood up well under a torrent of challenges. Notice that this study's conclusion is different from the conclusions that were drawn from the audit studies. The Boston Fed study deals with actual behavior—what happened when minorities applied to the lenders they actually applied to—rather than a hypothetical question—what would happen if minorities answered newspaper ads randomly. So the Boston study, in contrast to the

realtor audits, shows that minorities are actually harmed by the treatment they receive.

On the other hand, the Boston Fed study cannot pinpoint which lenders treat minorities worse, or even by itself show that any individual lender does so (perhaps for some reason minorities end up disproportionately applying to lenders who are tougher). The actual mechanism that results in the worse treatment that minorities receive in fact remains something of a mystery, since most lenders use reasonably objective nonracial criteria to make mortgage decisions, and most approvals or denials are pretty clear-cut.

The most plausible interpretation is that loan officers, perhaps unconsciously, provide extra, friendly help to white applicants more than they do to minority applicants. For instance, loan officers can help applicants by soliciting additional information about their credit histories, by finding their explanations more credible and putting them in better order, by hinting at how to put a slightly better face on their failings. Most applicants, of course, either clearly qualify or clearly don't, but in the borderline cases that matter for determining rejection rates, a little help can go a long way. In a number of legal suits involving racially disproportionate rejection rates at individual institutions, informal assistance like this was found to be the chief reason why whites were approved more frequently, holding objective characteristics constant.

One criticism of the Boston Fed study has been that lenders aren't *really* basing their decisions on race. Instead they're using race as a proxy for other characteristics that make borrowers more likely to default but are impossible to observe—things like rich relatives who will bail you out and a neighborhood where property values are very likely to rise. At least one study (Berkovec, Quercia, and Stegman 1992) has found that minorities are more likely than whites to have their properties foreclosed, even holding many characteristics constant. Thus lenders would seem to have some incentive to use race as a proxy.

Exactly how relevant this evidence is, however, has been disputed (see, for instance, Yinger 1995, pp. 75–78). Lenders may or may not lose money on a mortgage they foreclose, and the amount of money they lose varies widely. Higher foreclosure rates do not necessarily indicate that loans to minorities are less profitable on average. Foreclosures, moreover, are not events solely under the control of the borrower. When a borrower falls behind in paying off a mortgage, the lender has many options aside from immediate foreclosure. Lower white foreclosure rates may merely indicate greater willingness on the part of loan officers to try to arrange "work-outs" (rescheduling the mortgage payments) or "wraparounds" (another party buys the property and assumes the mortgage or pays it off).

In addition, minorities are less likely to prepay their mortgages when interest rates fall (Deng and Gabriel 2004). Partly this is because minorities are less likely to trade up to more expensive homes; partly it is because minorities are less ruthless about refinancing. If you lend somebody money at a high in-

terest rate and market interest rates fall, you don't want her to prepay—you want to keep collecting the high interest that the market can no longer give you. Everything else being equal, because whites are ruthless about prepaying (possibly because they can refinance more easily), minorities are more attractive mortgage customers.

More seriously, though, why lenders treat minorities differently is irrelevant to the main point that race matters in mortgage markets in a way that makes it hard for minorities to move to white neighborhoods. Lenders' motivations may be relevant to whether they will be admitted to heaven (although helping one's friends probably doesn't bar one from heaven either), but this question is not my business here.

Loan approval is not the only place where lenders can treat minorities differently, though it is the area that has been most intensively studied. Lenders also seem to engage in fewer and less-effective outreach activities with minorities and toward minority communities.

Holding almost everything else equal, banks in New York City are less likely to open branches in minority neighborhoods (Chang, Chaudhuri, and Jayartne 1997). Thus minorities may be less likely to deal with a bank when they want a mortgage, their time costs of getting a mortgage are higher, and they may find it more difficult to establish a good reputation and a long-term relationship with a neighborhood loan officer.

Audit studies in Pontiac, Michigan; Louisville; and Chicago have shown that before minorities file mortgage applications, lenders show them less enthusiasm than they show whites (Smith and Cloud 1993, Lawton 1993, Galster 1993). They seem less interested in giving information to black customers than white; they urge blacks more frequently to go to other lenders; they warn black customers more often that the application process is arduous; they give white customers more information about options; and they help white customers more.

There is also some evidence, most of it fairly old, that minorities on average pay higher interest rates than comparable white borrowers (Yinger 1995, pp. 79–80). This disparity may stem from the different types of lenders who specialize in different neighborhoods, or from the greater assistance that whites may receive in looking for rates or in timing their actions. Documented cases in which the same loan officer on the same day discriminates on the basis of race are rare (though not unheard of).

Thus we have several reasons why most minorities should have more difficulty gaining access to credit, and hence greater difficulty buying houses in white neighborhoods.

4. Redlining—A Digression

Redlining is the practice of denying mortgages for properties in minority neighborhoods. The name comes from the red lines on maps that at one time demarcated neighborhoods in which banks were not supposed to lend. Red-

lining is geographic discrimination, as opposed to the personal discrimination discussed in the last section. Redlining claims that even white loan applicants are denied mortgages when they want to buy in minority neighborhoods; personal discrimination claims that minority applicants are denied mortgages even when they are buying in white neighborhoods.

While the evidence is pretty clear that lenders treat minorities differently (personal discrimination), most modern studies find no evidence of redlining. For instance, Tootell (1996) uses the Boston Fed data to investigate whether lenders are more likely to deny mortgage applications for properties in minority neighborhoods, everything else being equal, including the race of the applicant. He finds that the applicant's race matters, but that the composition of the neighborhood the property is located in does not. Bovino (2000), however, challenges this conclusion by recoding the Boston Fed data and showing that mortgages in heavily African American neighborhoods are more likely to be denied; Ross and Tootell (2004) reach a similar conclusion. Mortgage applicants from minority neighborhoods are more likely to be required to have private mortgage insurance; how you treat this requirement determines what you think about redlining today.

Although redlining may not occur today, between the 1920s and 1960s most lenders operated with clear, written instructions about not lending in minority neighborhoods or transitional neighborhoods. When the Federal Housing Administration (FHA) was established in 1937, it incorporated these general practices as its own and codified them as part of its regulations. These regulations were also incorporated in the Veterans Administration (VA) program. Jackson (1985) details how strong and blatant these regulations were.

Redlining, however, even if it is still occurring, is not a way of keeping blacks out of white neighborhoods. Instead, it's a way of making property owners in black neighborhoods worse off. If anything, it encourages blacks to move to white neighborhoods by making black neighborhoods less attractive.

5. Local Governments

Local governments may use their police powers to keep minorities out of white towns and neighborhoods. Neighborhood preservationists have more of an incentive than other people to become involved in local politics, since change in neighborhood amenities affects them more than people who can move easily, and so local governments and civic associations are likely to place a high premium on neighborhood preservation. Local governments can make life miserable for new minority residents in any number of (perfectly legal) ways: holding up building permits, inspecting their houses frequently for building and zoning code violations, stopping them and their guests on the street to ask what they're doing or where they're going. Cutler and Glaeser (1997) find that metropolitan areas with more local governments are more segregated than those with fewer. Anticipating such difficulties with local gov-

ernments makes minorities less willing to bid on houses in white neighborhoods.

6. Trust

The final factor holding down the number of minorities in white neighborhoods is minority reluctance to move to those neighborhoods. Animosity and stereotypes probably do not contribute to this reluctance as much as they do to whites' corresponding reluctance to move to minority neighborhoods, but the role of mistrust is probably important. The day-to-day neighborly interactions that I described in the preceding section are important, and so is avoiding maniacs who will harass you. Of the blacks in Detroit in 1992 who said that they would not move into an all-white neighborhood, 90 percent gave as their reason a belief that whites would not welcome them (Ellen 2000, p. 58). Mistrust probably explains in part why minorities may not bid as high as whites for housing in white neighborhoods.

Mistrust may also matter for education. Actual learning depends on a complex of conjectures that students make about what their teachers and classmates are thinking about them, and that teachers make about what students are learning or not learning. Histories of stereotypes and animosity can easily get in the way in this process. (For instance, in a psychological experiment, black and white students did equally well on a test that they were told it was of no significance, but a matching group of black students did much worse on the same test when they were told it measured their academic ability [Fletcher 2001].) Schools are not like a gas stations; teachers don't simply pour learning into you. School is a place where people try to make you a different person. What you think of those people, what they think of you, and what you think they think of you all have to matter. So classroom situations that work perfectly well for white students may not work as well for minority students, and vice versa.

This educational relationship may matter for the housing market, too. Minority parents may not be as willing to pay as high a price as whites for educational inputs that might not be as effective with their children. Why pay a fortune to buy a house that will allow you to put your kids in a great school system, if the people there may treat them like idiots?

C. Interaction and History

We've seen why people of one race might be willing to pay less to live in a neighborhood dominated by people of another race, other things being equal. How does this play itself out in the housing market?

The open-city model (possibly modified for the presence of neighborhood preservationists) is the proper setting for examining this question. Segregation and the centralization of minorities have been features of American

cities for at least eight decades, and massive migrations of minorities have oc-
curred during this time. People of the same race, earning power, and tastes in
weather should not be markedly better off in one city than in any other.

But there is no mechanism for ensuring that otherwise identical people
of different races will be equally well off, even within the same city. You can
move from city to city, but (with very limited exceptions) you can't move from
race to race. Immobility between races would not prevent outcome equaliza-
tion if everyone else were color-blind: for instance, if you were born on Satur-
day you can never become a person born on Sunday, but if two people differ
only by the day of their birth, you would not expect that in equilibrium one
would envy the other. The envier could always outbid the envied for anything
he or she desired. But people treat different races differently, and so there is no
reason to think that in equilibrium no race will envy another.

In the 1940s and 1950s, the era of institutionalized Jim Crow laws, blacks
in the South were probably worse off than other Americans with the same
earning ability, if only because they were forced to endure the insults of legal
segregation and disenfranchisement. The open-city model implies that north-
ern blacks were no better off than southern blacks, and so northern blacks
were probably worse off than otherwise identical northern whites before the
civil rights laws were passed in the 1960s. The strong findings on housing
prices from this era support this conclusion.

Because blacks could expect violent opposition in any white neighbor-
hoods they tried to live in, and because many whites found living near blacks
so distasteful that housing prices in white neighborhoods near black neigh-
borhoods were lower than those in other white neighborhoods, when rising
city wages (and falling agricultural wages) in the first half of the twentieth
century drew more blacks from the rural South, black neighborhoods ex-
panded in population but remained compact and centralized.

Since this time, two relevant trends have affected cities: the relative ad-
vantages of central locations have declined (because of cars, television, and
frozen food, as explained in Chapter 7), and the relative well-being of African
Americans in the rural South has increased (because of air conditioning, civil
rights laws, judicial decisions on employment, education, public accommoda-
tions, voting, and so on).

In an open-city model, both of these trends act in the same direction:
they reduce the African American population of central-city minority neigh-
borhoods. That indeed was the dominant trend between 1970 and 1990. In
New York, for instance, between 1970 and 1990 the population of the South
Bronx fell 38 percent, the population of Harlem fell 26 percent, and the popu-
lation of North Brooklyn fell 14 percent.

Both trends also depress housing prices in these neighborhoods, pro-
vided some African Americans continue to live there. If race didn't matter to
whites, or if the next best alternative blacks had to living in a black neighbor-

hood was pretty similar to the next best alternative whites had, these lower prices would induce whites to move into black neighborhoods. But we have seen little such movement, even though substantial numbers of whites now have attitudes quite favorable to blacks. Even in the 1990s, despite great media attention to phenomena like the gentrification of Harlem, the proportion of non-Hispanic whites fell in Harlem and in New York City's other large minority neighborhoods.

On the other hand, the two trends have worked in opposite directions for the movement of blacks into white neighborhoods. The rising attractiveness of locations far from central business districts have made blacks willing to pay more to live in white neighborhoods, but the rising attractiveness of the rural South (and possibly government programs and greater political power in inner cities as well) has improved prospects for blacks living outside white neighborhoods and so reduced blacks' willingness to pay for living in a white neighborhood. The proportion of blacks living in white neighborhoods has increased rather substantially since 1970, and so the improvement in opportunities outside white neighborhoods probably was not so large. But if those opportunities had not improved, even more blacks would be living in white neighborhoods.

Thus relative opportunities now seem to be in just the right range to assure segregation and stagnation. If blacks had worse opportunities outside white neighborhoods, they would be willing to pay enough to bust out of ghettoes and spread throughout metropolitan areas. If black opportunities outside black urban neighborhoods were better, prices in those neighborhoods would be low enough to entice whites to move there.

And so, lines drawn in the 1950s and before have a greater effect on where many people live than current considerations. No wonder the losses from the segregation and centralization detailed in Chapter 11 appear to be so large.

D. Hispanics and Asians

We know a lot less about why Hispanics and Asians live where they do than we know about African Americans. The real estate audits described earlier found discrimination against Hispanics looking for a place to live, and so part of their story may be similar to the story about blacks. But we don't know whether Hispanic and Asian residence patterns are costly, and we don't know whether they're permanent.

II. Policies

Not surprisingly, many policies have been suggested to improve this situation. In this section we'll look at some of the more popular and promising of those

policies. Many analyses concentrate entirely on the policies' effects on minorities, and that's quite understandable. In this book, however, I have analyzed policies using the potential Pareto improvement criterion, and I don't plan to make an exception in this chapter.

Aside from consistency, there are two reasons why I think potential Pareto improvement is the right criterion to use, even when your only concern is the well-being of minorities. Politically, racial and ethnic minorities are not the majority, and so policies that cause great harm to whites are unlikely to endure for long. Even if political considerations don't concern you, if you're willing to think about reparations you should also think about potential Pareto improvements. To see why, think about some policy that benefits minorities but is not a potential Pareto improvement—the gains to minorities are less than the losses to whites. Then replacing that policy with a reparations scheme in which whites pay minorities some amount greater than the minority gain from the policy and less than the white loss makes everyone better off than implementing the policy. The reparations scheme is both better for minorities and more politically palatable. Thus there doesn't seem to be any argument for any policy that would have smaller gains than losses.

Are there other criteria we might want to use to judge policies intended to improve the racial situation? Possibly, but it's hard to formulate them in a consistent and usable fashion. For instance, some might argue that any form of discrimination is abhorrent and everyone should always act in a color-blind fashion. Aside from the logical difficulties of this principle (how can you find out whether the principle is being adhered to unless someone can recognize people's races?), the policy consequences of this position seem somewhat at variance with most people's inclinations. Should we, for instance, prohibit racial references in personal advertising the same way they're prohibited in employment advertising? (And if people are free to discriminate in probably the most important decisions ordinary people make—whom to marry or raise children with—can we consistently maintain that we value the principle of color-blindness?)

So while looking for potential Pareto improvements may not be the only way to approach the question of race, it surely is one of the better ways.

A. Making Moving to White Neighborhoods Easier for Minorities

Probably the longest standing American policies on neighborhood racial composition have been those that make it more difficult for whites to exclude minorities from neighborhoods in which they predominate. Several Supreme Court decisions that I have already mentioned laid the groundwork for these policies: the laundry cases of the 1880s, which struck down the California ordinances barring Chinese residents from various towns (and the 1891 federal court decision in *Gandolfo v. Hartman* prohibiting racial covenants against

Chinese home buyers); *Buchanan v. Warley* (1917), which struck down ordinances that forbade blacks or members of other minority groups from living in white-zoned neighborhoods; and *Shelly v. Kraemer* (1948), which held that it was unconstitutional for any government to enforce a racial covenant.

During the 1960s, the federal government took several steps to make it harder for whites to exclude minorities. In 1962 President Kennedy required federal agencies to act against discrimination in all housing programs that received federal support. The 1964 Civil Rights Act outlawed discrimination in public accommodations. In April 1968 Congress passed the Civil Rights Act of 1968, which includes as Title VIII the Fair Housing Act. Two months later the Supreme Court ruled (in the case of *Jones v. Maeger*) that the 1866 Civil Rights Act, which had been largely ignored for a long time, banned racial and ethnic discrimination in housing market transactions. In 1988 the Fair Housing Act was substantially amended to make it easier to enforce and to extend its coverage to cases based on family status and disability.

The Fair Housing Act and the 1866 Civil Rights Act are the chief tools the federal government uses to combat discrimination in housing. The majority of states have laws against housing discrimination, too, most of them similar to the Fair Housing Act. Under these acts, realtors are prohibited from racial "steering," "blockbusting" (using comments about racial or ethnic change to encourage panic selling), and discrimination in advertising or in providing information about available housing units. Sellers and landlords cannot discriminate in their decisions to sell or rent, or in the terms and conditions they offer. Mortgage lenders and insurers are similarly banned from considering the race of an applicant or the racial composition of the neighborhood in which a property is located when they approve mortgages or set terms.

The key questions on civil rights laws are how they are enforced and what standards of proof are required. The 1866 Civil Rights Act is easy for private parties to enforce, but it has a hard standard of proof: anyone who believes he or she has been discriminated against can file a civil suit, and there are no limits on damages or the attorney fees awarded, but plaintiffs have to prove discriminatory *intent,* not just discriminatory effect. In other words, they have to show both that the defendant's practices harmed them and that the practices were adopted specifically to target minorities.

The Fair Housing Act, by contrast, has a lower standard of proof: if the plaintiff can show discriminatory effect, then the burden is on the defendant to prove that the practices in question were based on business necessity, and that alternative, nondiscriminatory practices were not available.

Who is covered by the Fair Housing Act is severely limited. Most realtors and mortgage lenders are covered, but all buyers are exempt, as are the sellers of roughly 80 percent of U.S. housing stock—almost all single-family homes and owner-occupied buildings that have two to four units. (In all cases, however, sellers are prohibited from making statements that indicate discrimina-

tory preferences or intent.) The 1866 Civil Rights Act does not contain similar formal restrictions, but the requirement of proving intent is a very high hurdle for the small sellers and buyers who are excluded under the Fair Housing Act. Thus, while the refusal of almost all whites to consider living in predominantly minority neighborhoods is discrimination, it can't be prosecuted under either statute.

Do these laws, limited as they are, create potential Pareto improvements? Do they make minorities better off? Very little empirical evidence is available on either question, and theoretical arguments can be made in both directions.

Antidiscrimination laws are neither universal nor perfectly enforced. Imperfect enforcement creates two types of problems, both of which reduce the net benefits of the laws: some instances of discrimination go unpunished, and some agents are forced to do things they wouldn't otherwise do just to prove that they are not discriminating (landlords, for instance, may rent to minority tenants about whom they have serious misgivings out of fear of being sued). Altering the standards and burdens of proof can reduce the frequency of one type of problem, but only at the expense of increasing the frequency of the other. The easier it is for defendants in antidiscrimination cases to prevail, the more instances of discrimination will go unpunished; the easier it is for plaintiffs to prevail, the more often nondiscriminatory agents will be forced to do things that make no sense.

Laws that are not universal will sometimes force individual transactions in which losers lose more than winners gain, and they may even end up hurting the people they are supposed to help. For example, a law that required employers to provide mirror-image versions of all common office equipment for any left-handed employees would help those left-handers who'd already been hired, but would probably make most lefties worse off because employers would no longer hire them or would pay them much lower wages. Could it happen? Yes. Acemoglu and Angrist (2001) in fact show that the Americans with Disabilities Act almost certainly caused a large and steep decline in the employment of disabled people. Thus antidiscrimination laws have to be assessed on a case-by-case basis.

The public accommodations provisions of the 1964 Civil Rights Act have probably been successful; nobody pays any attention to them any more. They made it much easier for minorities to live and work in white neighborhoods. They were easy to enforce, and they mostly forced transactions where many people had obvious gains.

Antidiscrimination laws covering large landlords have two parts—a prohibition on differential access, and a prohibition on differential rents. Minorities may gain because the laws let them move to apartments they would not otherwise be able to move to, and since we have seen that traditional locations for minorities are bad places for many minorities to live, these opportunities should be valuable.

We have seen many reasons why some landlords and white tenants might be unhappy if more minority tenants lived in their building, but some of these reasons reflect ignorance, and others arise out of comparisons. A landlord in a white neighborhood who accepts minority tenants provides a positive externality to other tenants and landlords in white neighborhoods—she makes it less likely that they will be overwhelmed by minority demand if they are known to be nondiscriminatory. Thus, although an individual landlord might lose if she alone were compelled to accept minority tenants, there is no reason to believe that *all* landlords will lose if *all* are compelled to accept minority tenants. The universal nature of these laws, at least within the class of large landlords, also reduces search costs for minorities without any corresponding losses for whites.

Several arguments, however, can be made against these provisions. The first concerns the differential rent prohibition. Consider a potential landlord who, for whatever reason, would prefer not having minority tenants. Suppose she has a choice of building an apartment building in a location attractive to minorities or in a location unattractive to most minorities. If she can't charge differential rent, she will build in the location unattractive to minorities, even though potential minority tenants would be willing to pay her enough in extra rent to make it more than worth her while to build in the location more attractive to them. In this case the differential rent provision forces both landlord and minority tenants into a Pareto-inferior outcome.

This example can be extended. The alternative to building in a location attractive to minorities might be not building at all, or, more likely, building stores or single-family homes not covered by fair housing laws. By placing restrictions on one particular type of housing—large apartment buildings—fair housing laws discourage its construction. Fair housing laws might also encourage suburban towns to use their zoning powers to prevent construction of large apartment complexes, since these complexes cannot prohibit minorities from moving in or charge them higher rent. Because zoning codes, the income tax code, and possibly tenancy laws also discourage multifamily housing relative to owner-occupied dwellings, this effect of fair housing laws could make an existing distortion worse—and contribute to sprawl.

You can make similar arguments for and against the prohibitions on discrimination by realtors, mortgage brokers, and other middlemen in the housing market. These prohibitions help minorities by lowering search costs and opening up opportunities that might not otherwise be available to them. As with landlords, compliance by other realtors and brokers makes compliance by each individual realtor or broker less costly, since it prevents the lone tolerant agent from being swamped by all sorts of bad risks, and so the laws do not necessarily make the agents worse off, even though none might have complied with them voluntarily.

On the other hand, the laws may force realtors to waste their own time

and that of their minority customers by showing those customers houses in white neighborhoods in which they have no interest. They may also force mortgage brokers to use more objective and bureaucratic procedures in deciding when to lend; this might make obtaining a mortgage more uncomfortable for everyone and could introduce all sorts of errors that subjective human intuition could avert. The laws may also induce agents to use more subtle and more costly ways to avoid minority customers—by greatly restricting or sharply targeting their advertising, for instance; by locating their offices more strategically; or by splitting one company into several with different selling strategies and different prices.

In summary, then, the theoretical case for imperfectly enforced laws that cover only a small portion of the housing market is not overwhelming. The efficacy and efficiency of the current laws are ultimately empirical questions, and few answers are known.

Only a handful of empirical studies have been done on the effects of fair housing laws. Gablinger and Virabhak (2003) found that segregation decreased faster in the 1990s in states with more stringent fair housing laws, while Collins (2004) found no effects from variations in state laws in the 1960s. We can say, tentatively, that it seems reasonable to conclude that fair housing laws have a modest effect on segregation, and if they affect anything else, they don't make a big difference.

B. Making Minority Neighborhoods More Attractive

The other major response to the problem of race in cities has been to try to make minority neighborhoods more attractive. The public discourse around such programs as Model Cities in the 1960s, Community Development Block Grants in the 1970s, urban enterprise zones in the 1980s, and urban empowerment zones in the 1990s justified these initiatives on the basis of what they would do in poor minority neighborhoods, even though nothing in the legislation itself specifically invoked race. A large number of other smaller programs, both state and federal, have been earmarked for "distressed neighborhoods" or "impacted communities," and have been established with the same ostensible goal in mind, although once again race is never explicit in the legislation. Recent state programs—like those in Texas and Florida—to replace affirmative action in admissions to elite universities with percent rules—the universities will accept a certain percentage of top students from every high school in the state—are also attempts to make minority neighborhoods more attractive.

These programs use a multitude of different means: they reduce corporate or sales taxes on businesses in the targeted neighborhoods, they provide low-interest or no-interest loans to either commercial or residential borrowers, they build infrastructure, they improve schools, they pay for demolishing

abandoned buildings or adding to the police force, they establish employment training or social services programs in specific locations, they subsidize site preparation for locations that have sustained significant environmental damage, they construct subsidized housing. Some of these activities are designed to be potential Pareto improvements—better schools and job training programs, for instance—although they don't always succeed.

However, the most expensive and most frequently discussed programs, such as urban enterprise zones, would not be potential Pareto improvements in a market economy that functioned reasonably well. In such an economy, firms locate employment where it is most productive to do so (relative to other producers and consumers). Forcing them to move jobs into minority neighborhoods would obviously be inefficient—the gainers could not compensate the losers—because it would substitute a less productive location for a more productive one. Enticing firms into making the same move by offering tax reductions or direct subsidies is no more efficient: compared with coercion, it merely rearranges the names of winners and losers without changing the fundamental fact that winners can't compensate losers. Brownfield programs (subsidizing the cleaning of environmentally damaged sites so that they can be built upon) are similar: they try to get firms to locate on more expensive sites rather than cheaper ones.

The arguments in favor of these programs, then, must start from some sort of market failure. Clearly, many things are wrong with urban land markets in the United States today. When we analyze these programs, we will ask whether they are the best way of addressing the problems we have found. We'll work with the open-city model.

What happens when minority neighborhoods become more attractive in a single city? Since other alternatives don't change, nothing happens outside minority neighborhoods—there is neither more integration nor less. Within minority neighborhoods, population and density increase (at the expense of other cities), but consumers are no better off, because the price of land rises. Landowners—owners of land in minority neighborhoods—are the only people who gain.

Can these programs promote integration by inducing whites to move into minority neighborhoods? Because white outside options don't change, more-attractive minority neighborhoods make whites willing to bid more to live there, but the amount minorities are willing to pay to live in these neighborhoods also increases. Unless the improvements are designed deliberately to be more attractive to whites than to minorities—discounts on sun block, for instance, or ice hockey rinks—whites will be no more likely to move to minority neighborhoods.

In the open-city model, then, since owners of land in minority neighborhoods are not a particularly deserving class, programs that would not otherwise be potential Pareto improvements are not a good idea.

How appropriate is the open-city model for analyzing these programs? For state programs it is obvious that the open-city model is the right one to use. On the federal level, many of the programs were or are aimed at a few chosen cities—model cities and urban empowerment zones, for instance—while others rely on application processes that funnel money to those cities and organizations that ask for it best. Thus for a substantial proportion of the programs that seek to improve minority neighborhoods, the open-city assumption approximates their effect on the market. This holds true even for national programs. Minorities in U.S. cities have emigrated from places all over the world—the Caribbean, Latin America, China, the rural South—and they can also move back to those places.

The extreme mobility assumption of the open-city model, however, may miss a few of the beneficiaries. In particular, neighborhood preservationists and tenants with long-term leases or rent-controlled apartments are not part of the open-city model, but they would clearly also gain from measures that improved their neighborhoods. Their gains, though, essentially come at the expense of landowners and so do not change the fundamental negative conclusions that these programs help only landowners and do not spur integration.

Notice that these somewhat negative arguments apply only to programs that wouldn't otherwise be potential Pareto improvements—programs whose only appeal might be promoting integration. Programs that can be justified in their own right without appeal to land markets are not subject to this criticism.

C. Helping Minorities Move to Mainly White Neighborhoods

The most direct and economically interesting way to promote minority movement to mainly white neighborhoods would be to subsidize it: to pay landlords more or require a smaller contribution from tenants when a minority tenant moves to a mainly white neighborhood. The subsidies would no longer be available when the minority percentage in a neighborhood reached a certain level—say, twice the minority percentage in the metropolitan area. This would be an explicitly race-based program.

In an open-city model, the chief result of such a program is that white neighborhoods are more attractive to minorities, and so the population of minorities increases in those neighborhoods and in the metropolitan area as a whole. This raises prices, which drives some whites out of these neighborhoods and out of the metropolitan area. Alternatively, prices could fall because whites find the neighborhoods less attractive. Minority neighborhoods are unaffected, since they become neither more nor less attractive.

Under some circumstances, such a program could be a potential Pareto improvement. The rise in land prices would have to exceed the amount of the

subsidies, and for this to happen, unsubsidized minorities would have to move to the neighborhoods. Early movers would be providing a positive externality to later movers by opening up the neighborhood to them and making them feel comfortable there.

The subsidies, in other words, could work in Pigouvian fashion. This would happen if, race aside, neighborhoods occupied almost entirely by whites were really good places for minorities to live—because they were close to jobs, say, or the blocks were well laid out—but minorities did not live there because they were made to feel uncomfortable or fearful. We've seen that this is probably the case for many neighborhoods—that minorities live in the wrong places.

On the other hand, such a program would fail to be a potential Pareto improvement if few nonsubsidized minorities entered the neighborhood, or if nonsubsidized minorities entered because whites left and prices fell.

Designing a subsidy program that creates a potential Pareto improvement is a difficult task, although maybe not impossible in principle, if given excellent information. Aside from the inescapable problems of defining "minority" and "neighborhood" in a usable fashion, the program must halt subsidies when there are enough new minority residents in a neighborhood to encourage more to enter, but few enough to keep up white demand, too. No one knows how to do this now.

Not surprisingly, at present no programs like this have been tried. But some efforts have been made that bear a strong resemblance to subsidies for minorities who move to white neighborhoods. The chief difference is that these programs have been race-blind, in theory anyway: they have encouraged residents of public housing projects in high-poverty neighborhoods to move to low-poverty neighborhoods (which is different from encouraging minorities to move to white neighborhoods). In practice, almost all of the participants have been minorities, but promoting economic integration is a different goal from promoting racial integration. Also, in many cases the incentives have been rather weak—mainly just counseling.

The most famous program of this type grew out of a desegregation case in Chicago that was decided in 1976. A group of public housing residents sued the Chicago Housing Authority (CHA) and the U.S. Department of Housing and Urban Development (HUD), claiming that the CHA had discriminated in selecting housing sites—concentrating projects in minority neighborhoods—and in placing tenants—concentrating minority tenants in minority projects. The court sided with the tenants, and the ensuing settlement established the Gautreaux Assisted Housing Program (Gautreaux was the name of the lead plaintiff in the case).

This program offered certificates (called Section 8 certificates) to CHA tenants and families on the CHA waiting list. A family with a Section 8 certificate can live in any apartment that will accept the certificate, provided it

meets certain quality standards. The family pays 30 percent of its income in rent, and the government makes up the difference between this amount and an administratively determined fair market rent. In addition to providing the certificates, the program counseled participants and advertised to landlords in an attempt to encourage participants—most of whom were African American—to move to low-poverty neighborhoods, most of which were mainly white. The Section 8 program made white neighborhoods no more expensive than minority neighborhoods for participants; the counseling and outreach encouraged them to move there.

About half of Gautreaux participants did in fact move to mainly white neighborhoods. After a few years, those who moved out of minority neighborhoods were found to be more likely to have jobs than those who didn't. Their children were more likely to attend college, more likely to be employed if they didn't attend college, and more likely to have good jobs than the children of participants who decided not to move.

These differences were due in part to selection (people who already have jobs near white neighborhoods or who have good prospects for such jobs are going to find the prospect of living in that white neighborhood more attractive), but the Chicago results, which involved several thousand participants, inspired a number of other metropolitan areas to try similar programs, largely also in conjunction with discrimination litigation.

The federal government established a Gautreaux-type program too, called Moving to Opportunity (MTO). Originally established in the early 1990s, it ran into congressional trouble early in its life when members became afraid of being portrayed as financing "invasions of welfare recipients." (The preponderance of evidence from Chicago indicates that Gautreaux participants don't increase crime rates or decrease property values in the neighborhoods they move to; Yinger 1995, pp. 152–153, 235–236.) The MTO program was resurrected as a ten-year demonstration program in five cities. Participants are few, but scientific rigor is high. Residents of public housing projects in high-poverty neighborhoods are randomly divided into three groups. One group gets nothing; one group gets Section 8 certificates good anywhere in the metropolitan area; and one group gets Section 8 certificates that can be used only in low-poverty neighborhoods. Both groups that get certificates also receive extensive counseling, and anyone who wants to can continue living in the public housing project. Almost all of the participants are minorities. More than 90 percent of participating households are headed by women.

Moving to low-poverty neighborhoods seems to help MTO participants, but the effect is not nearly so powerful with the Gautreaux participants. After five years, adult participants' mental health improved—they were less harried and worried—and they were less likely to be obese. Teenage girls also experienced better mental health, and they reduced delinquency and problem behavior. But the adults who moved to low-poverty neighborhoods earned no

more money and suffered no less unemployment than those who stayed be-
hind in public housing projects, and children's cognitive achievement showed
no statistically significant improvement. Teenage boys were more likely to ex-
hibit problem behavior and more likely to be arrested for property crimes
(Kling, Ludwig, and Katz 2004; Kling et al. 2004; Orr et al. 2003).

What accounts for the difference between Gautreaux and MTO? One
possibility is that because Gautreaux participants were self-selected, only the
more ambitious tenants joined the program. Perhaps those who moved would
have been more successful anyway.

An alternative explanation is that MTO is an experiment in *economic* in-
tegration while Gautreaux is an experiment in *racial* integration. The low-
poverty neighborhoods that MTO participants moved to were heavily minor-
ity. Families who had the opportunity to move from public housing projects
to low-poverty neighborhoods lived in neighborhoods that were on average
38 percent black (and not all of them changed neighborhoods), while families
who did not receive a voucher lived in neighborhoods that were on average 48
percent black (Goering 2003). Race may be the key to the big benefits that
Gautreaux found, not income.

D. Integration Maintenance

The phrase "integration maintenance" refers to a wide array of activities that
have as their goal promoting integrated communities. Such activities include
mortgage subsidies for whites moving into heavily minority neighborhoods,
mortgage subsidies for minorities moving into heavily white neighborhoods,
housing counseling designed to encourage moves by minorities into white
neighborhoods and whites into minority neighborhoods, bans on visible for-
sale signs and active encouragement by realtors of "panic" selling, advertising
of the advantages of integrated living, active enforcement of building and
housing codes to prevent deterioration of areas undergoing racial transition,
enhanced public services for these areas, programs to promote intergroup un-
derstanding, maintaining separate waiting lists by race for subsidized housing
projects and quotas in each building or on each floor, and equity assurance
programs—programs that try to assuage white homeowners' fears that their
investments will lose value if they stay in an integrated neighborhood by reim-
bursing residents for 80 percent of losses in home value after five years (Yinger
1995, 125–130).

Integration maintenance activities are mainly seen now in upper-middle-
and middle-income suburbs near heavily minority cities: Oak Park and Park
Forest near Chicago, for instance; Shaker Heights and Cleveland Heights near
Cleveland; Park Hill in Denver; West Mount Airy in Philadelphia; and Maple-
wood and South Orange near Newark. There are generally two motivations
behind these efforts: the value of integration per se, and the wish (of neigh-

borhood preservationists) not to lose a community to a rapid, nasty, complete racial turnover. Virtually no federal or state activities consciously support housing integration now.

The variety of integration maintenance activities calls for a variety of analyses. We have already seen that, under some circumstances, carefully designed subsidies for minorities moving into white neighborhoods could be potential Pareto improvements. Obviously the same reasoning applies for subsidies for whites moving into mainly minority neighborhoods.

However, existing subsidies for whites moving into minority neighborhoods are not for heavily or historically minority neighborhoods like Harlem, but for neighborhoods in leafy suburbs where the minority proportion is low but rising quickly. These subsidies are an attempt to slow that rise. The external benefits that whites who move into these neighborhoods provide are enjoyed not primarily by subsequent white families who move in, but by the white neighborhood preservationists who live there already.

While some such external benefits are no doubt real and should be included in any cost-benefit analysis, caution should be used in accepting arguments about the size of these benefits. These subsidies, because they seek to slow minority entry into white neighborhoods, might be characterized by their detractors as a modern and genteel version of the firebombings and cross burnings that frequently greeted minority entry in earlier times. Whites may exaggerate the benefits of these programs because some of the benefits are based on racist ignorance, because they may overestimate the extent to which minorities gain from their presence, and because some of the losses that neighborhoods suffer when minorities enter in large numbers are offset by gains in other neighborhoods—for instance, if highly skilled and experienced teachers transfer out of schools serving a minority neighborhood, they make the neighborhoods they transfer to more attractive.

The same sort of analysis applies to equity assurance programs, and even to publicity programs designed to attract whites to neighborhoods with rising minority populations. (South Orange and Maplewood, New Jersey, for instance, run ads in community newspapers serving the Upper West Side in Manhattan and Park Slope in Brooklyn, two areas with large numbers of wealthy, mobile, liberal whites.) These activities may be beneficial or not to the current populations of those neighborhoods, but they could have detrimental effects on society as a whole. No evidence supports the idea that they should be banned, but it is probably wrong to think that higher levels of government should encourage more towns and neighborhoods to undertake them.

The exception to this general precept about integration maintenance activities may be housing counseling and the provision of accurate information about the advantages of integrated settings. Helping people find out where they would be best off living in the absence of racial considerations is likely to

be a potential Pareto improvement. As Ellen (2000) points out, these efforts are probably best undertaken on a metropolitan-wide or even a national basis, rather than the local basis that's now predominant. The more different integrated environments a household is exposed to, the more probable it is that it will find one to its liking. Thus metropolitan efforts are more likely both to promote integration and to produce potential Pareto improvements.

On the other hand, some activities that have operated under the rubric of integration maintenance are pretty clearly inefficient. These include bans on solicitation by realtors, prohibitions on display of for-sale signs, and quotas. All of these policies keep people from making Pareto-improving deals, and all of them are directed at keeping minorities out of integrated neighborhoods. Of course these deals should be discouraged if they have significant negative externalities, but the size of the negative externalities from having minority neighbors (if they exist at all) is probably not great enough to merit actions of this type.

These measures are also unlikely to be successful in promoting stable integration. Integrated areas become segregated not because whites flee but because they are not attractive enough to outsider whites who would replace those whites who leave in the normal course of events. These policies do little to make a neighborhood more attractive to whites who don't already live there. On the contrary, they prevent outsiders from learning about possible vacancies in the neighborhood and present the neighborhood to the world as troubled, insular, and defensive.

The final problem with solicitation bans, sign prohibitions, and quotas is that under most circumstances they are unconstitutional.

E. Encouraging Cars

Raphael and Stoll (2001) have argued for encouraging minority car ownership, because it will lead to better job opportunities. The discussion in Chapter 3 gives some idea of how this might be done. The United States fights pollution by requiring clean and expensive cars, and it deals with accidents by requiring insurance premiums. Neither payment is tied to actual driving; both represent fixed costs of owning a car. Relative to charging for pollution and insurance by the mile, these procedures discourage car ownership by people who wouldn't drive much—poor people in particular. So moving to a more efficient way of handling pollution and accidents might improve job prospects for minorities.

Car ownership would also improve educational opportunities. Not having access to cars means that minority parents have limited choices on where to send their kids to school. This is especially important for child care, because the car is the primary mode of delivery to formal organized facilities, especially to those where cognitive development is stressed. Thus whites (who own

cars) rely much more heavily on day care centers to provide child care, and minorities (who don't own cars) rely more on relatives (U.S. Bureau of the Census 1995). Car ownership may matter for other levels of schooling as well, especially if magnet schools, charter schools, and voucher programs proliferate. Current evidence on magnet and charter schools, however, indicates no consistent racial pattern in who takes advantage of these opportunities (Teske and Schneider, 2001).

F. Doing Nothing

The argument for doing nothing is that all the possible policy options would do more harm than good. You can believe that the current situation is very bad and still think doing nothing is the best course of action. You can also believe that housing markets and education systems are not the most opportune places for policy interventions; perhaps labor markets or marriage patterns or reparations should be where the effort is concentrated. Such considerations are beyond the scope of this book.

III. Conclusion

In Chapter 11 we saw that U.S. cities had too much segregation, and too much centralization of minorities. Minority education wasn't good enough. Cities could be better, for everybody.

In this chapter we looked at policies that might achieve those potential Pareto improvements. We did not find any strong answers—no magic programs, no demonic idiocies. Instead we found lots of open questions. The opportunities to make cities better are large enough that answering those questions is a good idea.

Questions

1. Suppose an organization to promote integration is being set up in the neighborhood in which you live. Would you join it? Why or why not? If you joined, what activities would you ask it to engage in? What activities would you oppose? Why? How would your answer differ if you were primarily motivated to improve the well-being of minorities currently living outside your neighborhood? How would it differ for improving the well-being of whites outside your neighborhood?

2. Employment discrimination laws focus on the actions of buyers (employ-

ers) in the labor market. Why don't housing discrimination laws focus on the actions of buyers, too? What would happen if they did? Why don't they?

3. How do you expect the progress of information technology to change the way houses are bought and sold? What effects should this have on realtor behavior and the search costs of minorities? What effects do you expect this technological progress to have on segregation?

4. Triple City is a closed city. The population consists of 300 non-Hispanic white (NHW) households, 100 Asian households, 100 African American households, and 100 Hispanic households. We will refer to all groups but NHWs as "minorities." There are three neighborhoods, each with 200 houses.

   NHW households prefer to live with other NHW households. Let w denote the NHW proportion in a neighborhood. So w is between 0 and 100. NHWs are willing to pay up to

   $$400 + w$$

   to live in that neighborhood. No one else cares about race. Minorities are willing to pay up to 500 for any house anywhere.

   a. Show that 100 NHWs and 100 minorities in each neighborhood is an equilibrium. (That is, show that no household wants to move to another neighborhood.) Call this the integrated equilibrium.
   b. Calculate the index of dissimilarity D for NHWs versus minorities.
   c. Find price in each neighborhood. How much consumer surplus do minority households realize?
   d. Show that 200 NHWs in one neighborhood and 50 in each of the other two is also an equilibrium, when house prices in each neighborhood equal the NHW willingness to pay. (That is, show that no household wants to move to a different neighborhood at the prevailing prices, and that if prices changed in one neighborhood, holding other prices constant, supply would not equal demand in that neighborhood.) Call this the segregated equilibrium.
   e. Calculate the index of dissimilarity D for NHWs versus minorities for this equilibrium.
   f. Find consumer surplus for minorities.
   g. At which equilibrium are minorities better off?

References

Acemoglu, Daron, and Joshua D. Angrist. 2001. "Consequences of Employment Protection? The Case of the Americans with Disabilities Act." *Journal of Political Economy* 109(5):915–957.

Berkovec, James, Roberto G. Quercia, and Michael A. Stegman. 1992. "Residential Mortgage Default: A Review of the Literature." *Journal of Housing Research* 3:341–379.

Bovino, Beth Ann. 2001. "Essays on Race and Discrimination in Financial Markets." PhD diss., Columbia University.

Chang, Angela, Shubham Chaudhuri, and Jithendra Jayartne. 1997. "Rational Herding and the Spatial Clustering of Bank Branches: An Empirical Analysis." Columbia University Department of Economics Discussion Paper 9697-24.

Clarke, T., and P. Perlman. 1948. *Prejudice and Property: An Historic Brief against Racial Covenants.* Washington, DC: Public Affairs Press.

Collins, William. 2004. "The Housing Market Impact of State-Level Anti-Discrimination Laws, 1960–1970." *Journal of Urban Economics* 55(3):534–564.

Cutler, David, and Edward Glaeser. 1997. "Are Ghettoes Good or Bad?" *Quarterly Journal of Economics* 112(3):827–871.

Cutler, David, Edward Glaeser, and Jacob Vigdor. 1999. "The Rise and Decline of the American Ghetto." *Journal of Political Economy* 107(3):455–506.

Deng, Yongheng, and Stuart Gabriel. 2004. "Are Underserved Borrowers Lower Risk? New Evidence on the Performance and Pricing of FHA-Insured Mortgages." Working paper, Lusk Center for Real Estate, University of Southern California.

Drake, St. Clair, and Horace R. Cayton. 1945. *Black Metropolis: A Study of Negro Life in a Northern City.* New York: Harper and Row.

Ellen, Ingrid Gould. 2000. *Sharing America's Neighborhoods: The Prospects for Stable Racial Integration.* Cambridge, MA: Harvard University Press.

Fletcher, Michael. 2001. "Long Division." *The New Crisis* (September–October):26–31.

Gablinger, Ynon, and Suchin Virabhak. 2003. "Discrimination in the Housing Market: What Can We Learn from Inter-State Differences in Fair Housing Laws?" Mimeo, Department of Economics, Columbia University.

Galster, George C. 1993. "Use of Testers in Investigating Discrimination in Mortgage Lending and Insurance." In M. Fix and R. J. Struyk, eds., *Clear and Convincing Evidence,* pp. 287–334. Washington, DC: Urban Institute.

Goering, John. 2003. "The Impacts of New Neighborhoods on Poor Families: Evaluating the Policy Implications of the Moving to Opportunity Demonstration." *Federal Reserve Bank of New York Economic Policy Review* 9(3):113–140.

Jackson, Kenneth T. 1985. *Crabgrass Frontier: The Suburbanization of the United States.* New York: Oxford University Press.

Kling, Jeffrey R., Jens Ludwig, and Lawrence Katz. 2004. "Youth Criminal Behavior in Moving to Opportunity." Accessed at *www.wws.princeton.edu/~kling/mto/recent.html.*

Kling, Jeffrey R., et al. 2004. "Moving to Opportunity and Tranquility: Neighborhood Effects on Adult Economic Self-Sufficiency and Health from a Randomized Housing Voucher Experiment." Accessed at *www.wws.princeton.edu/~kling/mto/recent.html.*

Lawton, Rachel. 1993. "Pre-Application Mortgage Lending Testing Program: Lender Testing by a Local Agency." Paper presented at Discrimination and Mortgage Lending: Research and Enforcement Conference, sponsored by U.S. Department of Housing and Urban Development, Washington, DC, May 18.

Massey, Douglas S., and Nancy Denton. 1993. *American Apartheid: Segregation and the Making of the Underclass.* Cambridge, MA: Harvard University Press.

Munnell, A., et al. 1996. "Mortgage Lending in Boston: Interpreting the HMDA Data." *American Economic Review* 86:25–53.

Orr, Larry, et al. 2003. "Moving to Opportunity: Interim Impacts Evaluation." Accessed at *www.wws.princeton.edu/~kling/mto/recent.html.*

Raphael, Steven, and Michael Stoll. 2001. "Can Boosting Minority Car-Ownership

Rates Narrow Inter-Racial Employment Gaps?" *Brookings-Wharton Papers on Urban Affairs* 2.

Ross, Stephen L., and Geoffrey B. Tootell. 2004. "Redlining, the Community Reinvestment Act, and Private Mortgage Insurance." *Journal of Urban Economics* 55(2):278–297.

Schuman, Howard, et al. 1997. *Racial Attitudes in America: Trends and Interpretations,* rev. ed., Cambridge, MA: Harvard University Press.

Smith, Shanna, and Cathy Cloud. 1993. "The Role of Private, Nonprofit Fair Housing Enforcement Organizations in Lending Testing." Paper presented at Discrimination and Mortgage Lending: Research and Enforcement Conference, sponsored by U.S. Department of Housing and Urban Development, Washington, DC, May 18.

Sugrue, Thomas J. 1996. *The Origins of the Urban Crisis: Race and Inequality in Postwar Detroit.* Princeton, NJ: Princeton University Press.

Teske, Paul, and Mark Schneider. 2001. "What Research Can Tell Policymakers about School Choice." *Journal of Policy Analysis and Management* 20(4):609–632.

Tootell, Geoffrey. 1996. "Mortgage Lending in Boston: Do Mortgage Lenders Discriminate against Neighborhoods?" *Quarterly Journal of Economics* 111(4):1049–1079.

U.S. Bureau of the Census. 1995. *Population Reports,* PTO-53. Washington, DC: U.S. Bureau of the Census.

Yinger, John. 1995. *Closed Doors, Opportunities Lost: The Continuing Costs of Housing Discrimination.* New York: Russell Sage Foundation.

Chapter 13

Housing: The Big Picture

Housing serves many purposes. It shelters people and their animal companions from cold, rain, and excessive heat. It protects their TVs, books, furniture, games, clothing, clocks, pictures, food supplies, computers, and financial records from the weather, marauding animals, and unfriendly humans. It hides them from nosy neighbors. It blocks out noises and foul odors. It removes urine and feces and temporarily stores other wastes. It washes their bodies, their dishes, and their clothing, and captures the resulting sullage. It gives children and dogs a place to play, and adults a place to garden or have parties or use their NordicTrack. It gives sick people a place to recover and store their medicines. It is where mail is delivered and phone calls come. It provides electricity, gas, water, cable and Internet connections, and parking spaces for cars. Cats get vantage points from which to survey the world, and cooks get ovens and stoves to use. Houses carry rights, too: their inhabitants get to vote in local elections, send their kids to local public schools, play on municipal tennis courts, dump paper and plastic at the town recycling center, and object to proposed zoning variances within 200 feet of their home. Where your house is located also affects who your neighbors are and what type of kids your children play with.

Housing, then, is a bundle: brick and mortar; wood and nails; pipes and sinks; furnaces and air conditioners; windows, doors, and locks; toilets and showers; wires and fiber optic cables; lawns and driveways; garages and gardens; swimming pools and swing sets; ovens and Jacuzzis. Not every house has all of these pieces, and all of them can be purchased separately.

Why are these particular goods and these particular purposes so often bundled together? The answer is the usual one: increasing returns to scale and transportation costs. Think first about storing your possessions. Storage requires some sort of a box, and building boxes is governed by the two-thirds

---

To get the most out of this chapter, you should be familiar with these concepts: adverse selection, deadweight loss, externalities, moral hazard, potential Pareto improvement, and second-best. You will also need to understand the two-thirds rule and economies of scale. You can find all these terms in the Glossary.

rule: doubling the storage volume requires less than twice the cost of constructing the box. So it is cheaper to build one box for all your possessions—and for your body, too—than a series of separate boxes. One big box also saves on transportation costs—you don't have to go to one place for your toothbrush, another for your toothpaste, a third for your sink, and so on. Foot transportation—especially when you're barefoot in the middle of the night—is expensive, and so the more of your destinations there are in the same big box, the better. Because they save on transportation costs and take advantage of increasing returns to scale, houses are like miniature cities.

What makes good sense for one person, moreover, makes doubly good sense for two. It's cheaper to build one big box to store two peoples' possessions than two smaller boxes, one for each. Two people can share, also: bathrooms, comfortable chairs, newspapers, and stoves aren't usually in use at all times, and so the excess capacity that one person would normally have can be used costlessly by the other. The same applies to water pipes, telephone lines, and cable connections. Two may not be able to live quite so cheaply as one, but it's far less than twice as expensive.

Of course, there's no reason to stop at two. Three, four, five people, even a dozen—the more people, the greater the savings on construction cost and excess capacity. But just as there are limits on city sizes, so there are reasons why we don't spend all of our time in one big indoor commune. The reasons are just like the reasons why cities don't grow ever so big: congestion can kick in around the shower in the morning and the television at night; indoor trips can become very long. Security and privacy are probably much more important for limiting house population than they are for limiting city population. A hundred people living in one big room have reason to fear each other, and to worry about how their possessions will stay their own. Even among saints, privacy is a problem: most people have at least a few things they don't want other people to know about. So is coordination: negotiating with fifteen other people before you watch a television program can take a lot of time, even if they're all obliging.

So in developed Western societies, generally fewer than a half dozen people share a single "housing unit"—an indoor space with some common facilities and no interior locked doors. Almost always, moreover, the people who share a housing unit are bound by some ties of affection, real or imagined—they are lovers, spouses, siblings, members of a religious order—and that makes negotiating the inevitable frictions and loss of privacy easier; affection internalizes externalities. Not everyone lives this way: there are military barracks and psychiatric hospitals, college dormitories and monasteries, assisted-living residences and unrelated roommates who divide everything according to strict rules, long-stay hotels and homeless people. But most people in developed Western countries have some place they call home and some small set of people with whom they share it.

Not all activities are done at home, of course, nor do people keep all of

their possessions in the same big box. That's because some other activities have more strongly increasing returns to scale and don't require the same level of trust and affection in order to realize those returns. People venture forth from home to work, shop, attend school, and participate in sporting events, and they do these things with a lot of people they wouldn't feel good about sharing a kitchen or bedroom with.

Because housing bundles together the means to pursue so many different ends, it ties up a large part of the world's wealth. The value of U.S. residential real estate in 2003 was $13.4 trillion, out of a total net stock of wealth of $34.8 trillion (U.S. Bureau of Economic Analysis 2005). American houses in 2003 were worth about the same amount as the total combined value of the thirteen largest stock markets outside the United States (those of Japan, the United Kingdom, Germany, France, Switzerland, Canada, Hong Kong, China, Italy, Australia, the Netherlands, Taiwan, and South Korea; U.S. Bureau of the Census 2004, table 1377). For most Americans, wealth is housing.

It's not surprising, then, that public policy debates about housing abound. Houses represent mini-cities in the midst of larger cities. The larger cities become chaotic if the mini-cities do too much. The mini-cities wither if the larger cities do too much. Most public policy debates are really about how to draw the line—whether to encourage more activity inside houses and more of the kinds of activity done in houses, or to allow more activity subject to the formal coordinating mechanisms of the market and the government.

In this chapter, we'll consider three major sets of policies. All of them try to make the mini-cities that houses represent more independent, self-reliant, and self-sufficient. The three sets of policies are the homeowners' tax preferences, which promote ownership in particular and housing consumption in general; tenancy laws, which promote stability and independence among renters; and mortgage programs that make it easier for people to own houses but change the terms on which they own.

To understand these policies, we will first have to answer at two more fundamental questions: whether housing markets, left on their own, would underproduce housing, and whether they would leave too many people renting rather than owning. It doesn't make sense to have programs to address problems unless those problems really exist. The first two sections of this chapter are devoted to these questions. Then we'll look at each of the three sets of policies separately.

Home is not a morally neutral word. How we express love, whom we share privacy and intimacy with, whether we have a roof over our heads—these are questions that are basic to our lives. Since housing policies affect the answers to these questions, debates about housing are almost always at least partially informed by ethical concerns, including concerns about the types of lives other people live and the types of values other people hold. In thinking about policies, we will have to think about these concerns, even if they might

seem meddlesome. But we won't abandon our usual concerns about Pareto improvements, potential and actual, and about thinking through all the implications of a policy. Indeed, it's because people care so much about housing that we need to analyze it carefully.

I. Not Enough Housing?

There are several arguments for why policies to increase housing consumption might be a potential Pareto improvement. First, activities within the home are hard to organize—explicit contracts are usually out of the question, hold-up problems can be large because the physical and emotional costs of leaving a household are great, affection is always in short supply. Left on their own, then, people may fail to realize all the economies of scale that come from sharing housing. As a result, people may eat alone in restaurants too often, and together at home too infrequently; they may not do enough of the activities that can be done best at home. Everyone would be better off if more people reached agreements that would allow them to eat at home and watch TV together more, but such agreements could not be enforced after they were made. In the absence of such agreements, government policies that encourage people to spend more on housing than they would otherwise spend might be potential Pareto improvements.

The problems of raising children create another argument for policies that encourage housing. Housing is an important part of raising children: it gives kids a place to eat with their family, a place to play, a bedroom to feel secure in, locations for books and toys and computers and doing homework. It's hard for kids to eat in restaurants, study in libraries, and sleep in hotels. Better and bigger houses may improve children's health, too, by protecting them from cold, drafts, and infections; poor health in childhood often leads to poor learning as well as to illness in adulthood.

If children were solely the property of their parents, the usual arguments about consumer sovereignty would imply that parents would spend the right amount on their upbringing. If children and stereos both existed only for the enjoyment of parents, then parents would be uniquely and best qualified to allocate their expenditures between the two goods. But children are independent actors in ways that stereos are not; they are, or will become, individuals capable of striving for goals, feeling pain, making their own decisions. Like stereos, however, they cannot engage in Coasian bargaining with parents, and so it's quite possible that parents will raise them less well than they would if kids could credibly promise to pay parents for the quality of their child rearing. True, most parents love their children dearly, and some children do later take care of parents in their old age, but so long as there are some parents who don't think their children's consumption is equivalent to their own (or don't

understand their kids perfectly), and so long as kids' promises to their parents are not credible or enforceable, children won't be treated as well as they should be. Forcing parents to treat kids better would be a potential Pareto improvement.

Other people, too—especially other children—have an interest in how children are raised. You and your parents are not the only ones to realize the benefits if you grow up to be an informed and wise voter or a generous contributor to many good causes or a good parent or a valued friend; nor do you and your parents realize all the costs if you turn out to be a criminal, a ward of the state, or someone whose car stalls in the Lincoln Tunnel during rush hour. Child rearing has externalities.

Regular externalities are a final reason why markets may not provide enough housing. The neighbors may not like it if your house is decrepit or your yard is unkempt; or they may enjoy looking at your garden or your artfully designed parapet. If only one or two immediate neighbors care about these things, of course, Coasian bargaining can handle any apparent problems. But if large numbers of passersby are affected, there may be some basis for government action.

II. Not Enough Responsibility?

Even if people are consuming the right amount of housing, a number of arguments suggest that they may not be consuming it the right way: the terms under which households occupy houses may not give them incentives to treat that housing, their neighbors, and their neighborhoods the right way. In particular, many people think that the robust independence of ownership gives better incentives than the less responsible lifestyle that tenancy fosters.

A. Maintenance

This argument takes a number of different forms. First, tenants may have less incentive to maintain their houses well. An owner loses money if the house looks like it's ready to collapse when he moves out, but a tenant doesn't. As I will point out later, this argument has some weaknesses. But if it is true, it has some consequences, both direct and indirect.

To the extent that the maintenance tenants neglect matters to outsiders, tenancy is a problem. This is the direct consequence. The indirect consequence is that, because tenants' incentive problems make being a landlord a more difficult job, fewer people become landlords, they invest less, and as a result less housing is available than should be. There is some empirical evidence to support at least one step in this rather long chain of reasoning: Wang et al. (1991) found that the values of rental properties in the San Antonio area were 3.7 percent less than otherwise comparable ownership properties.

B. Stability

The next argument for ownership arises from the value of stability. In every age group, the average number of years that a renting household has lived at its current address is about half of the comparable number for an owning household (Caplin et al. 1997, p. 10). Why? Because moving costs a lot more for owners. Buying a house includes the costs of closing on a mortgage, title insurance, origination fees, and lawyers' fees. Selling a house entails paying commissions to real estate agents and brokers. DiPasquale and Wheaton (1996) estimate total out-of-pocket costs for the purchase and sale of a house to be around 6 to 9 of the house's value. On top of these are the other costs of moving, such as searching for a house and having strangers tramp through your bedroom every weekend when you are trying to sell. Tenants' moving costs are much smaller.

When a family moves, it upsets other peoples' lives as well. The family forces their former neighbors to get to know new people and adjust to new routines. The former neighbors may in fact be losing valued friends. Movers themselves experience none of these costs. (Movers lose old friends, but this loss is accounted for in their decision to move.) Similarly, children in a family may gain more from stability than adults do, and so to the extent that family decision making doesn't give enough weight to kids' well-being, households may move too much. Kids usually spend more time in a neighborhood and in neighborhood schools than adults do. Adults don't usually have to change jobs when they move, but kids usually have to change schools; adults can often maintain old friendships, but kids often don't. Thus, discouraging moving might be a potential Pareto improvement. Again, empirical evidence supports this position. Wang et al. (1991) found that being surrounded by at least two rental properties lowered the value of an owned property by about 2 percent.

Despite the empirical evidence, however, there are three reasons to be skeptical about the value of promoting stability. First, the neighbors may be confusing correlation with causality, at least on a societywide level. Some people are just more footloose than others—they don't have or particularly want children, they crave excitement and adventure, they make new friends easily or don't care how many they have. Because of moving costs, these footloose people would naturally prefer to rent rather than own. Thus the difference in stability between renters and owners may reflect differences in preferences more than differences in opportunities; increasing home ownership would not increase stability as much as the cross-sectional differences would imply.

Moreover, if other personality traits that make people bad neighbors are correlated with footlooseness—throwing loud, long parties, for instance—then increasing homeownership among the footloose is not going to create fewer bad neighbors in the aggregate. It's just going to mean that bad neighbors will own their homes rather than rent them.

A second reason to be skeptical about the importance of stability as a reason for promoting ownership is the behavior of landlords. If stability really did provide sizable external benefits, landlords, especially large landlords, would promote it much more aggressively than they seem to. They have both the incentives and the tools; so do large condominium associations. If stability is good for neighbors, and those neighbors are mainly residents of the same apartment or condominium complex, then the owner or owners of that complex gain from stability and should be willing to pay for it. The owner has the ability to pay for it by charging lower rents to tenants with longer tenures. Empirical investigation (for example, Guasch and Marshall 1987) has shown that landlords do charge long-term tenants less, but the discount is pretty small. If landlords who stand to gain from stability are not willing to pay much for it, it's appropriate to ask whether governments should be willing to pay more than that for stability.

A final cause for skepticism about arguments for stability is the link between residents and voting. The public officials who represent any area were selected by the area's current residents. The public officials know that those residents voted for them at least once; they don't know how any set of future residents might vote. If they want to stay in office, they are probably better off keeping in place the voters who elected them, rather than allowing strangers to come in. So the interest that public officials have in promoting stability is greater than the interest society as a whole has in promoting it (because the gain to current public officials from keeping their jobs is offset by the loss suffered by those who want to replace them). Skepticism is always in order when somebody is arguing in favor of something that helps him or her personally.

C. A Stake in the Community

The idea of this argument is that homeownership gives you a stake in your community and a stake in society. This makes you more willing to join in the collective efforts to enhance or protect that community or society. You have something to lose, something to protect, when you own a home. For instance, in an open-city model, renters don't gain from anything that makes a city more attractive or lose from anything that makes it less attractive. So they have no reason to join the Kiwanis Club or to found a community arts group or to vote for candidates who will promote the general good.

D. Kids

Children are the focus of much community activism; good schools and playgrounds and Little Leagues require a lot of volunteer work, formal and informal, and such programs naturally work better with a stable pool of volunteers. Similarly, kids may gain more than adults from residential stability. Both reasons together suggest that ownership is better for kids than renting. But be-

cause kids' interests may not be well represented in family decision making, too few families may own houses.

Empirical evidence supports this idea. Green and White (1997) find that, compared with the children of renters, the children of owners are less likely to drop out of school or become teenage mothers. Aaronson (2000); Boehm and Schottman (1999); Haurin, Parcel, and Haurin (2000); and Harkness and Newman (2002) come to similar conclusions. This emerging line of research, however, has not yet identified what it is about homeownership that helps children. Stability surely plays a role, but the role of the larger houses and backyards that owners have has not been clearly analyzed. Ownership per se may not be the cause.

E. Capital Market Problems

Most people need to borrow money in order to buy a house; almost nobody needs to borrow money to rent an apartment. Because of the legitimate fears lenders have about whether they will be repaid, capital markets don't work perfectly. Many times people can't get loans, even though those loans would be Pareto improvements if they could bind themselves to repay (at least at those times when luck worked in their favor). For this reason, homeownership may be too rare.

But one should be careful about accepting this argument uncritically. It applies only if homeownership is the only activity that capital markets inhibit. In fact, people may seek loans for many other reasons—for starting or expanding a business, for getting an education, for going on vacation or buying a boat—and all of these other loans are inhibited by capital market imperfections too. Because you can use a house as collateral, borrowing for houses may be easier than borrowing for these other purposes. Making homeownership easier, then, while it helps people make better decisions between renting and owning, also forces them to make worse decisions between buying a house and buying a boat. (This is a second-best problem, like the question of reducing subway fares because car travel was underpriced. Recall from Chapter 5 that if subway fares were set below marginal cost by the same proportion that car travel was underpriced, people would allocate trips the right way between cars and subways, but make too many trips in total. Subsidizing homeownership to offset capital market imperfections completely might allocate people properly between owning and renting, but it would cause overinvestment in housing relative to large consumer expenditures.) It's not clear that promoting homeownership under these circumstances will always be a potential Pareto improvement.

Thus we have several reasons to think that perhaps governments should encourage people to spend more on housing than they otherwise would, or at least to alter their approach to it. Since many programs are designed to accom-

plish these ends, this is reassuring news: there is indeed an illness that might be treated by the medicine the doctor is prescribing. But we have not yet seen either that that illness is afflicting the patient we're concerned about, or that that medicine would work on that illness if the patient had it, or that no other medicine would work better. To answer these questions, we have to look more closely at actual policies and their effects.

III. Income Tax

The federal income tax is the major housing policy in the United States. The income tax law encourages people, particularly richer people, to buy and hold more owner-occupied housing than they would if housing were treated like other large durable goods.

A. How the Tax Code Encourages Owner-Occupied Housing

Four major provisions of the tax code favor expenditures on owner-occupied housing: the exemption of imputed income from the services of owner-occupied housing; the deductibility of state and local property taxes; the almost complete exemption of housing from capital gains taxation; and the exemption of imputed income from spousal services.

The key concept for understanding these provisions is **imputed income.** Your imputed income is the value of things that you get that aren't cash and so don't appear in your paycheck. For instance, if your cousin owns a restaurant and lets you eat there for free every Thursday, the value of the meals you eat there is imputed income for you. That's because your cousin's letting you eat $100 worth of meals is just about the same for you as his giving you $100 cash and making you pay regular prices.

Similarly, if *you* own the restaurant and eat there for free on Thursdays, the value of the meals is still your imputed income. You could have taken more profit or paid yourself a higher salary and paid for the meals, and nothing would have been different.

With housing, the imputed income that matters is the value of services you get from a house that you both own and occupy. If you own and occupy a house that would rent on the market for $2,000 a month, your imputed income from that house is $2,000 a month. You could charge yourself $2,000 a month and collect it as your landlord, and nothing would be different.

1. Exemption of Imputed Income from Homeownership

Think about an investor with a choice between investing in a bond and investing in a house, which she will then occupy. Both cost $100,000. The bond pays interest, which is taxable, and the investor will pay the rent for an apartment. Suppose the bond pays 10 percent interest and the investor pays 30 percent of

her income in taxes (at the margin). If she buys the bond, her after-tax income is $7,000, which she can spend for a place to live, and for whatever else she wants.

On the other hand, she could take the $100,000 and buy a house instead of a bond. Then she wouldn't have any taxable income from her investment. What will she do? Essentially her choice is between getting a house to live in or getting $7,000 a year. If she values the services of the house at more than $7,000 a year, she'll buy the house; if she values the services of the house at less than $7,000 a year, she'll buy the bond.

But this decision rule sets up the possibility of a Pareto improvement, since the bond produces $10,000 a year. Suppose she values the services of the house at $8,000 a year. Then she'll buy the house and pay no taxes. But if the government doesn't tax the interest on the bond, she can buy the bond and receive $10,000 a year. She would be better off and the government would be no worse off, because in neither case would the government collect anything.

Alternatively, the government could include as part of her income the value of the services of the house—her imputed income from owning the house. Then she would buy the house if and only if it were worth more than $10,000 a year to her—which is the right decision rule for her to use. Any other rule leaves the door open to Pareto improvements. Exempting imputed income causes deadweight losses.

This scenario started with an investor with $100,000 in cash. Most people have to borrow—take out a mortgage—when they buy a house. It turns out, though, that even with borrowing, the argument for either not taxing bond interest or taxing imputed house income is the same.

To see this, suppose our investor doesn't have any money to begin with. To buy the house, she'll have to borrow $100,000. Since bonds are being offered that pay $10,000 a year, she'll have to offer $10,000 a year to anyone who lends her the money to buy the house. This payment of $10,000 a year is deductible under current tax law—it's called the mortgage interest deduction—and it reduces her taxes by $3,000 a year (assuming she has other income and enough other deductions). So once again she'll buy the house if and only if its services are worth more than $7,000 a year to her. But if the services of the house are worth only $8,000 a year to her, everyone could be made better off if she didn't do anything and the government paid her $2,000 a year. The government would be better off because it would be losing only $2,000 a year instead of $3,000 a year, and she would be better off because she would be getting $2,000 instead of paying $7,000 for services worth $8,000.

Taxing her imputed income from the house in this case would be a potential Pareto improvement. If her imputed income were taxable, and the annual services of the house were worth less than $10,000 to her, she would show a loss on the deal (equal to the difference between the interest payment and the imputed income). The mortgage deduction would mean her loss

would be partially offset by a reduction in taxes (the federal government would pick up 30 percent of it), but still there is no reason to suffer a loss. So she would buy the house if and only if it were worth more than $10,000 a year to her—once again using the right rule.

Notice that I'm not arguing that the mortgage interest deduction is the problem, even though this deduction is popularly cited as the major subsidy to homeownership. The first example showed that even people who pay cash have the wrong incentives when imputed income is not taxable. Concentrating on the mortgage income deduction gives the wrong idea of how much federal income-tax policy favors owner-occupied housing. Especially since the 1986 changes in the tax code, many homeowners do not claim the mortgage interest deduction because they are not leveraged enough, or their housing value is low, or because they have so few other deductions that they do better by taking the standard deduction instead of itemizing. In the early 1990s, only about 40 percent of homeowners took the home mortgage deduction, and even of those who used it, many didn't gain the full advantage of it because they didn't have enough other deductions (if the standard deduction is $5,000 and you have only $1,000 in other deductions, a mortgage income tax deduction of $10,000 reduces your taxable income by $6,000, not the full $10,000).

Follain, Ling, and McGill (1993) estimated that in 1989 taxpayers paid $109 billion less than they would have if imputed income were taxable and they behaved the same way; this amount is referred to as the tax expenditure. The tax expenditure for the mortgage interest deduction in 2002 was $64 billion; if the same ratio as 1989 held, the annual tax expenditure from not taxing imputed income would be around $215 billion (U.S Bureau of the Census 2002b). This amount may fall in the future, if marginal income tax rates fall.

2. Deductibility of Property Taxes

The deductibility of state and local property taxes is similar to the mortgage interest deduction: it wouldn't be preferential treatment if imputed income were taxable. If you treat owner-occupants as landlords who rent to themselves—that's what taxation of imputed income amounts to—then they, like other landlords, should be able to deduct all normal business expenses, including property taxes and mortgage interest. But if you treat a house as purely a consumption good, like a loaf of bread or a ticket to the circus, then deducting the tax you pay on it is strange, since you can't deduct an excise tax on a circus ticket. The result is that people buy too much housing and too few circus tickets, relative to a situation in which housing and circus tickets were treated the same way. So this preference for homeowners adds to the distortion that the imputed income problem causes.

3. Capital Gains Exemption

The final commonly cited tax preference is the virtual exclusion from taxation of capital gains in owner-occupied housing. **Capital gains** are the income you

derive when the price of something you own goes up. If stock you own increases in value from $20,000 to $30,000, you can spend $10,000 on an unanticipated vacation and still have just as much wealth as you did before the stock prices rose. Often you have to pay taxes on capital gains on stocks and most other assets (although because you owe taxes only when you sell a stock, not when it actually rises in value, capital gains are taxed differently from other kinds of income). For owner-occupied houses, though, capital gains are almost never taxable. Thus people may invest in houses even if they think they are actually going to appreciate less quickly than stocks, because they'll still end up doing better when they don't have to pay capital gains taxes.

This tax preference for housing is especially valuable in times of inflation. Owning a house is especially cheap when prices for everything are rising: the gain in house values is not taxed, and while inflation drives up the interest rates homeowners pay on their mortgages, the mortgage interest deduction and the exemption of imputed income make the government pick up a significant piece of this increase. Even in noninflationary times, though, the capital gains exemption is a fairly large tax expenditure, though not close in size to the imputed income exemption: the 2000 estimate was $20 billion.

4. Imputed Income from Unpaid Labor

One more provision of the tax code also encourages people to buy housing instead of other things: the exemption of imputed income from spousal labor. You don't pay tax on the household work that you do or your spouse does for you even though you value that work. If you live in an apartment and the superintendent shovels the snow, the federal government taxes that income. But if you own a home and shovel the snow yourself, the full value of your labor escapes taxation.

Similarly, if you hire a butler, he has to pay income tax on all of the wages you pay him. But if you marry him and he continues to act as your butler, he doesn't pay taxes on the allowance and room and board you give him, and you don't pay taxes on the value of the services you receive from him (even though the marriage allows you to forgo hiring a regular butler and to spend your money on something else instead).

What does this have to do with housing? People generally don't randomly marry other people they do business with in order to save on taxes; hardly anybody I know has married a dentist just to save taxes on getting their cavities filled. The process works the other way: people do more of activities that spouses do for each other because this labor is essentially untaxed. This matters for housing because most of these activities are complements to housing.

Raising children is the major activity that fits this description, and it is wives who do most of the work, not husbands. If households had to pay themselves taxable salaries (rather than the untaxable salaries they pay themselves now) for the time they spend raising children, then raising children would be more expensive. People would devote less effort to raising children and more

effort to getting their cavities filled and other goals. They would also raise their children differently—they would use less of their own time (since it would be more expensive for them) and more purchased inputs, like paid caregivers, day care, and toys. Housing is complementary to raising children in the sense that one reason to have a big house and a nice yard is that they help you to raise children yourself. If you plan to spend all your time in the dentist's office and keep your kids in child care 150 hours a week, you can make do very well in a studio apartment. Because the labor that goes into using big houses and nice yards is untaxed, the income tax code is once again encouraging people to buy more housing.

5. Asset Protection

While not part of the income tax code, Medicaid rules encourage people to invest in housing by protecting them against claims on their assets.

If you're old and need nursing home care, the federal government will pay for it under Medicaid if you've exhausted all your assets. Nursing homes cost about $8,000 a month in 2003, and so most people without houses enter nursing homes, use up all their savings, and then rely on Medicaid. But if you own a house and have a spouse or children who live there, the house is protected—the spouse or children can continue living there and inherit the house, even though Medicaid is paying the full tab for the nursing home.

So if you're likely to need Medicaid and have a spouse or coresident children, you and they are much better off investing in housing than in other assets. If you sell the house for $200,000, invest in stocks or bonds, go live in an apartment, and then have to enter a nursing home, the government will implicitly take the $200,000 and your heirs will wind up with nothing. But if you stay in the house and go directly to the nursing home, your heirs will wind up with $200,000. Yun (2003) shows empirically that the Medicaid incentives actually encourage this type of investment decision.

(Bankruptcy laws in many states work similarly, but they probably don't have the same effect. If you go bankrupt, your creditors can claim all your stocks and bonds and other real estate, but they can't force you out of your home. So the more money you invest in your house, the greater the wealth you can emerge with after bankruptcy. Many states cap the value of a house that can be protected in bankruptcy, but some large states have very high caps or none at all. But this protection after bankruptcy probably makes lenders more reluctant to accept houses as collateral and so may not increase the amount of owner-occupied housing.)

6. Overall

Although most of the tax preferences in the U.S. code are for owner-occupied housing, there are a few for rental housing as well (many were eliminated in the 1986 tax revision). The most important of these are accelerated deprecia-

tion, certain rules on passive losses, and certain credits for apartments where low-income households live. As tax expenditures, the total value of these provisions was about $12 billion. Even though the rental sector is smaller than the owner-occupied sector (twice as many housing units are occupied by their owners as were occupied by renters, and the value of owner-occupied units is more than three times the value of rental units), the tax preferences for owner-occupied housing are much, much bigger.

Together, these preferences imply that the before-tax rate of return on housing should be lower at the margin than the before-tax rate of return on other assets people could be investing in. (Here rate of return means the sum of any dividends or interest and capital gains.) We don't have any direct data on marginal rates of return, but economists have calculated the average rate of return to investments in housing and in other assets over long periods of time. Taylor (1998) found that the annual return to housing was about 5 percent between 1960 and 1995, while the return to nonhousing fixed capital was about 13 percent between 1975 and 1995, and the returns to high school and college education were around 11 percent and 8.5 percent, respectively. Mills (1989) found similar results for the period before 1985. If these average results hold on the margin, then a shift of $1 trillion of the $13.4 trillion housing investment to nonhousing fixed investment would increase the U.S. national income by $40 billion to $80 billion a year, forever.

However, if marginal income tax rates fall and taxes on capital gains and dividends from stocks are cut substantially in the next decade, the size of this distortion may decline (and with it the relative value of owner-occupied housing).

B. Consequences

1. Better Housing

What have been the effects of all these tax preferences? Since World War II, when the income tax first became important, U.S. housing stock has grown and improved immensely. In 1940, 45.3 percent of housing units lacked complete plumbing facilities (hot and cold piped water, a bathtub or a shower, and a flush toilet); in 1990 only 1.1 percent of housing units lacked complete facilities. In 1940, 20.2 percent of housing units were crowded in the sense that they were occupied by more than one person per room; in 1990 only 4.9 percent of units were crowded (U.S. Bureau of the Census, 1999).

By 1999, a large portion of the nation was very well housed. Of occupied year-round units, 81 percent had a porch, deck, balcony, or patio; 32 percent had a usable fireplace; 30 percent had two or more living rooms or recreation rooms; 42 percent had two or more complete bathrooms; 53 percent had central air-conditioning. Owner-occupied units were even better: 89 percent had a porch, deck, balcony, or patio; 43 percent had a usable fireplace; 41 percent

had two or more living rooms or recreation rooms; 54 percent had two or more complete bathrooms; 60 percent had central air-conditioning (U.S. Bureau of the Census 2001, tables 963, 954; U.S. Bureau of the Census 2002a).

Americans are better housed than most other people in rich countries. The average resident of a central city in the United States has 496 square feet available to him or her; the average suburban resident has 570 square feet. By contrast, the average square footage in Paris is 350; in Madrid, 260; and in Tokyo, a mere 150 (Glaeser and Kahn 2003).

Newly constructed private, single-family homes are of still higher quality. Of houses completed in 1999, 93 percent had two or more bathrooms, 62 percent had a fireplace, 84 percent had central air-conditioning, 34 percent had four or more bedrooms, and 51 percent had more than 2,000 square feet of floor space. (U.S. Bureau of the Census 2001, table 938). Clearly the income tax preferences are not solely responsible for this great and continuing surge in housing stock—rising incomes, improving transportation, and inventions like air-conditioning and television have contributed a lot, too—but they have played a role.

2. Sprawl

The tax preferences have also contributed to the decentralization and reduced density of U.S. cities. This has happened in several different ways.

The most obvious is the direct subsidy to the holding of residential land. Land is part of what you buy when you buy a house, and the rental value of that land and the capital gains from that land are both exempt from taxation, and protected if you enter a nursing home. So people buy more land than they would otherwise, and less of other goods. Because they buy more land, metropolitan areas are bigger and less dense. Voith and Gyourko (1998) speculate that this effect might reduce density by up to 15 percent, but they label this as an upper bound.

Less obvious is the subsidy to automobile use. What we and the tax code call housing provides a habitat for cars as well as for humans. In 1999 most owner-occupied houses included a garage or carport (73 percent); another 24 percent provided some other form of off-street parking (U.S. Bureau of the Census 2002a). The imputed income that owner-occupants receive from these parking facilities is untaxed. This tax preference lowers the cost of owning and operating a car—or multiple cars: probably more than 60 percent of the owner-occupant households have more than one vehicle. Since the actual cost of driving a car is well below the social marginal cost, as we saw in Chapters 3 and 4, once you have a car, you drive it too much. Thus by encouraging owner-occupants to have more cars, the tax code encourages decentralization and socially harmful driving. This encouragement is not small: my rough estimate of the tax expenditure on garages and off-street parking is more than $11 billion a year; see Box 13A. To put this in perspective, the total operating

Box 13A

The Garage Subsidy

☰ According to the American Housing Survey for 1999, 49.9 million owner-occupant households had garages or carports and 16.4 million had off-street parking (U.S. Bureau of the Census 2002a). Figuring the services of a garage at $1,000 a year and the cost of an off-street parking spot at $500 a year, we get an estimate of $58.1 billion annually in imputed income. At a 20 percent marginal tax rate, this implies an annual tax expenditure of $11.6 billion. Notice that this may be an underestimate, because it assumes one parking space per household; most households have multiple vehicles.

expense for passenger transit in the United States in 1999, both bus and rail, was about $20.5 billion (U.S. Bureau of the Census 2001, table 1104).

Of course, any sprawl that the tax code encourages would be unavoidable if the tax code were correcting real problems the best way possible. We've noted several reasons why housing markets, left to their own devices, might arrive at Pareto-inferior outcomes. The question of whether the tax code encourages excessive sprawl thus boils down to the question of whether the tax code is correcting these problems in the best way possible.

3. Better Household Sizes—or Worse?

One problem I noted was that the agreements between people that allow them to live together and enjoy the economies of scale that housing provides are difficult to negotiate and enforce; households may be too small. Homeowners' tax preferences, however, are not limited to households of any particular size. In 2000, 22.4 percent of owner-occupant households had only one member; the tax preferences these households received were clearly not discouraging fragmentation. In general, a tax code that makes housing cheaper across the board encourages fragmentation; it doesn't discourage it. The cost of living with others is loss of privacy and freedom; the benefit is saving money on housing. Reducing housing prices cuts the benefits from living with others, but doesn't cut the costs. So it exacerbates the problem of small households. From 1940 to 2000 the proportion of households with only one member rose from 7.7 percent to 25.8 percent (Woodward and Damon 2001). Of course, living by yourself is a luxury, and so rising incomes have also contributed to this increase.

On the other hand, the U.S. tax code is not really an across-the-board subsidy to housing. Renters mainly get small subsidies, and subsidies are larger for larger houses and richer households. To the extent that the tax code

differentially lowers the price on the types of housing larger households are most likely to occupy, it encourages larger households—a group of people may get a greater subsidy by living together rather than living separately.

In addition, the exemption of imputed income from spousal services is clearly an incentive for people who might be spouses to share housing. (We don't know whether the differential effect or the overall cheapening effect is stronger.)

Thus the tax code may or may not work toward encouraging large-household formation on net. If this were the chief goal of housing policy, however, different approaches might work a lot better, such as taxing imputed income of single-person households, or reducing the proportion of imputed income taxed as the size of the household increases.

4. Better Lives for Kids?

If housing markets were left to their own devices, child rearing might be relatively neglected. Here the beneficial effect of the tax preferences is clearer, both theoretically and empirically. Having space of their own, both inside and outside, is good for kids, and the tax code makes it easier for parents to provide that space. Schools help kids, too, and the deductibility of local property taxes makes it easier to pay for good schools. The tax code, moreover, encourages ownership, ownership may promote stability, and stability may help children develop. The exemption of imputed income from spousal labor also promotes child rearing activities. For all these reasons, then, we can think of the tax code as making up for some of the deficiency in care that might otherwise occur.

You still have to ask, though, whether the benefits could have been provided more cheaply and whether they are roughly of the right amount. Only a third of owner-occupied houses have children living in them (which is almost the same as the proportion of renter households with children) (U.S. Bureau of the Census 2002a). Even owner-occupied households with children have many housing-related expenditures that probably aren't closely related to children's well-being (fireplaces and marble bathroom fixtures, for instance)—but these are favored just as much by the tax code as those that directly help children. A tax code that restricted preferences to households with children and excluded the value of items like fireplaces and marble bathroom fixtures would promote child rearing much more cheaply and effectively than the current code.

5. Better Neighborhoods?

What the tax preferences might be addressing are traditional externalities. Perhaps people outside the household gain from the additional housing consumption. These externalities could not be matters that affected only one or two immediate neighbors; Coasian bargaining could resolve those without outside intervention. Nor could the externalities the tax code addresses be se-

rious observable matters that threaten a community's "health, safety, welfare, or general morals." Municipal codes—zoning, health, housing—are designed to deal with those kinds of problems. Indoor plumbing, for instance, is required in practically every densely inhabited location in the United States; homeowners' tax preferences do little to promote it.

How might the tax code alleviate externalities, then? By encouraging spouses to stay home during the day, the code may make neighborhoods safer or at least more congenial—it's good to have someone around with whom UPS can drop a package. To the extent that people stay home because they have great backyards and home entertainment centers, the tax code may reduce congestion, air pollution, and the spread of contagious diseases.

But all of these goals might also be pursued other ways, perhaps more cheaply and effectively. Keeping people in the neighborhood during the day could be encouraged by small tax breaks for people who work odd hours (better road and transit pricing would do this, too) or who telecommute. Congestion and air pollution can be addressed directly (Chapters 3 and 4 discuss how in detail). To promote stability, a tax on moving would work better than the current tax preference: it would discourage moving by renters as well as owners, and it wouldn't require people who would otherwise be renters to expend real resources in moving.

The tax code may also exacerbate many negative externalities. I have already argued that it promotes sprawl. Some random passers-by may enjoy looking at imposing mansions, but many others would find trees on a vacant lot more aesthetically appealing. And neighborhoods might be safer and more congenial if houses were closer together. Perhaps the best way to regulate externalities among neighboring houses is to join them under common ownership; that way the common owner internalizes all the externalities. We see such an arrangement frequently: it's called renting.

6. Better People?

Another argument for homeowner tax preferences is that they give people a stake in their community. Other mechanisms exist, though, to give people a stake in society—stock ownership, for instance. Ideas like the flat tax, the consumption tax, and the value-added tax exempt all income from capital, and so, their proponents believe, encourage saving across the board without singling housing out for special treatment. These reforms could give people the stake in society that owning assets brings, without distorting the choice between housing and other assets.

Stakes in local communities can also be created other ways. Communities could operate with Henry George–style land taxes and distribute any gains or losses among any group of people (for example, voters) they chose. This group would then be the stakeholders.

More generally, it is not the ownership of brick and mortar, or wood and

plumbing, that gives you a stake in a community or a society; it is the ownership of any asset that is supplied less than perfectly elastically. If you're the only piano teacher in an isolated town, then your income depends on that town's prosperity, no matter what you do for housing. If you rent an apartment in a place that you love, would feel horrible if you had to move, and were willing to pay far more for this apartment than anyone else in the world, then you would gain from any improvement in the neighborhood, so long as the improvement were worth more to you than to people who might bid against you for the apartment. Homeowners do care about their communities, but they are not the only ones who do.

7. Racial Disparities

Since minorities are less likely to own homes than whites are, and the homes they own are less expensive, homeowners' tax preferences go mainly to whites, even holding income constant. Some of this difference is due to discrimination by mortgage bankers and realtors. Ellen (2000), moreover, shows that the more renters there are in an integrated neighborhood, the more likely it is to stay integrated.

Reducing or eliminating homeowner tax preferences would therefore act like two kinds of different programs that have been advanced to reduce racial disparities. First, eliminating or reducing them would act like an integration maintenance program by making integrated neighborhoods more stable. Second, it would act like a program for making minority neighborhoods relatively more attractive, since net benefits would flow disproportionately to those neighborhoods. In Chapter 12, we saw that under the closed-city model such programs can increase minority well-being. Eliminating or reducing homeowners' tax preferences would thus be far more powerful (considering the size of the preferences) and more efficient than any known program that strives to improve minority neighborhoods.

IV. Rent and Tenancy Regulation

The income tax code, we have seen, is designed to encourage owners to stay put and to care about their community. The same goals animate most governmental initiatives in the rental sector, too. Most of these initiatives, however, operate at the state or local level, not the national.

A. Tenancy Laws

The main way that governments try to promote stability among tenants and give them a stake in the community is through laws governing leases and evictions. In most places, a tenant who has lived in an apartment for a certain length of time—generally a few weeks or a few months—must be given a

lease. A lease is a contract between the landlord and the tenant that runs for a year or more. During the term of the lease, the landlord can't raise the rent or evict the tenant unless he can prove to a court that the tenant behaved very badly. The tenant has to pay the rent throughout the term of the lease, even if she moves or decides she doesn't like the apartment any more or realizes that the landlord couldn't get anyone else to pay as much as she's paying. The only way the tenant can legally get out of paying the rent is if she proves that the landlord behaved very badly—withholding heat or necessary repairs, for instance. Leases also set some ground rules for interaction between tenant and landlord—for instance, the rules the landlord has to follow to gain access to the apartment.

Leases thus require that tenants be "mini-owners." Like owners, they have a degree of control over what goes on in their home. During the term of the lease, they have a stake in the community—if things get better, they can enjoy it completely without suffering any rise in rent; if things get worse, they bear the whole brunt without any compensating rent reduction. And they have sizable moving costs while the lease is still alive (even though landlords don't often successfully sue tenants who break leases, they do have the right to keep the security deposits that tenants have paid, and any furniture that's left behind). Tenancy laws don't make renters as stable and as tied to a community as owners are, but they do make them more stable than they would otherwise be.

The tenant and the landlord can negotiate over many provisions of the lease, just like any other contract, but different laws restrict what can be agreed on. No state allows leases that last for ninety seconds or that permit the landlord to enter the apartment whenever he feels like it and pour himself a cup of coffee.

Without tenancy laws, most tenants and landlords would almost certainly operate under some form of long-term lease (just as commercial tenants, who are not usually covered by such laws, do now). This is because leases avert a significant hold-up problem: searching for a new apartment and moving into it is costly for a tenant, and so a landlord could force a tenant to pay higher rents than she would otherwise have agreed to once she has sunk those costs and moved in. Tenancy laws, then, primarily affect what leases say, not whether they are written.

Tenancy laws have several other justifications, too. Prospective tenants differ in how much care they will give to an apartment and in how long they will stay. Landlords would like to rent to good tenants and avoid people who will become bad tenants, but their ability to differentiate between the two groups is probably poor, absent expensive psychological tests and extensive background checks.

A long lease with strong safeguards against eviction is more valuable to people who intend to be lousy tenants than it is to people who intend to be good tenants. Without tenancy laws, a smart landlord would conclude that a

prospective tenant who asked for such a lease was likely to be a lousy tenant, and so be reluctant to rent to her. Prospective tenants with perfectly legitimate reasons for wanting long, secure leases would therefore be reluctant to ask for them, for fear of the message that asking would send about themselves. Similarly, landlords would be afraid to try to carve out a niche for themselves by offering more-secure-than-average leases, as these leases would attract a disproportionate share of bad tenants. Tenancy laws, in this interpretation, help people get closer to what they might agree on if they weren't afraid to talk about it.

The problem of bad tenants manifests itself in other ways, too. Suppose leases are short and bad tenants reveal themselves quickly. Then bad tenants will always leave or be evicted at the end of every lease; they are better off in the pool of prospective tenants: a new landlord thinks they might be costly but the old landlord knows for sure. Then the pool of people searching for new apartments will be heavily weighted with bad tenants, and so landlords will think that the probability is high that a new tenant drawn from the pool of searchers is a bad tenant—and they will charge all new tenants high rents as a result. New tenants, both good and bad, would be better off if landlords got rid of bad tenants less quickly.

There is an externality at work here: a landlord who evicts a bad tenant makes every searcher already in the pool worse off, because the eviction makes landlords think worse of them and the average quality of the pool is lower. Because Coasian bargaining is unlikely to alleviate this externality, tenancy laws that force leases to be longer than they would otherwise be may create potential Pareto improvements. Borsch-Supan (1986) and Hubert (1990, 1995) show how in more detail.

But tenancy laws are not without problems. The great theoretical advantage of apartments over owner-occupied houses is that they "internalize externalities": if the owner controls everything that goes on in the apartment building and also reaps all gains and suffers all losses, then he will act efficiently and keep the annoyances that tenants impose on each other down to just the right level. By making tenants mini-owners, though, tenancy laws destroy this unity of control and ownership.

If the Coase theorem held for rental housing, however, this splitting of ownership rights wouldn't matter. Negotiation would realize all potential Pareto improvements, no matter how the original ownership rights were assigned. If one tenant's dog bothers another tenant, the bothered tenant can pay the dog owner to get rid of the dog—but only if getting rid of the dog is a potential Pareto improvement. If a tenant is thinking about throwing darts at the exposed plaster walls or selling the radiators for scrap metal, the landlord can pay her not to, even if the lease gives the tenant the right to do these things, provided that the harm these activities would cause to the building's long-run value would exceed their value to the tenant. Even if someone comes along who wants to pay more for an apartment than the current tenant, strong

leases present no obvious barrier to a Pareto improvement: either the current tenant can sublease or the landlord can pay the current tenant to leave.

But the Coase theorem doesn't always hold; negotiation and enforcement are sometimes too difficult. The dog may be bothering a hundred other tenants, not just one, and each of them would be happy to let the others take care of the problem. It's hard to keep the tenant who has been paid not to sell the radiators from smuggling them out in the middle of the night anyway. And many behaviors with possibly harmful effects are much more subtle than these and are essentially beyond any court's powers of verification: how diligently the tenants sweep crumbs off the floor to prevent infestation by vermin, how often they clean the grease out of the oven, how carefully and forcefully they instruct their children about the dangers of playing with elevators or matches.

To the extent that tenancy laws vest power over these decisions with tenants and make it difficult for landlords to influence tenant behavior, they make apartment buildings messier and less valuable than they should be. Everyone could be made better off if landlords charged lower rent and had more control, but the tenancy laws prevent this.

Tenancy laws also induce landlords to become more vigilant about whom they permit to rent their apartments, since eviction is difficult. Landlords become more reluctant to rent to people with children or with histories of mental illness or with spotty credit records—or to members of minority groups that are stereotyped as bad tenants. Without tenancy laws, they might want to take small gambles on questionable tenants by entering into short-term leases; then the tenants would get a chance to prove themselves. Small gambles like this, though, are impossible with tenancy laws, and so many landlords will opt for no gambles at all rather than take the big ones—full leases—that the law requires.

Could the goals of tenancy law be somehow achieved with a smaller number of problems? Possibly. Stability, for instance, could be promoted with a rent subsidy—or an income tax deduction—that increased with a tenant's tenure in a particular apartment (or neighborhood); the size of the subsidy could depend on the number and age of children in the household. Tenants could get a long-run stake in the community if they were required to buy some variety of security that was pegged to the town's or neighborhood's total property value. Neither of these policies would work perfectly, but then tenancy laws don't work perfectly either.

B. Rent Control

1. What It Is

While most jurisdictions require leases and regulate their content, a few go further and regulate what happens between successive leases on the same apartment. This is generally called **rent control.** Since many of the motivations

for rent control are the same as the motivations for tenancy laws, and sometimes rent control looks like a very long lease, you should think of it as a strong species of tenancy law. In practice, it's often hard to sort out rent control laws from tenancy laws.

Rent controls forbid changes in rent for much longer periods than leases do, and keep tenants in apartments much longer than they would otherwise stay. They give tenants a real stake in their community and so (like homeowner tax preferences) probably encourage civic participation.

Arnott (1995, p. 101) estimated that some form of rent control (invariably soft) covered 10 to 15 percent of U.S. rental stock. That would translate into 3 to 4 percent of the housing units in the United States. Rent control is more common in Canada and Europe.

Rent control comes in many different forms. The best known form is what is called **hard rent control** (also called first-generation rent control). Hard rent control is simply a rent freeze: the government forbids some or all landlords from increasing the rents they charge. This kind of rent control was most widespread in the United States during World War II; the only place it survived for more than a couple of years after the war ended was New York City. European cities—for instance, London—kept hard rent control in force much longer than North American cities did.

Soft rent control (second-generation rent control) is much more varied. These regulations generally permit rent increases but cap the amounts, cover some units but not all, and contain various provisions ending all rent regulation on some units under some circumstances. Most soft rent controls allow automatic rent increases equal to the rate of inflation and permit landlords to petition for additional increases based on special circumstances. These controls generally exempt buildings with only a handful of units and newly constructed housing (although no one can promise that controls won't be placed on these units after a few years have passed). Some jurisdictions remove an apartment from rent control when it becomes vacant or a new tenant moves in; others decontrol apartments when their rent reaches a certain level; and still others have no decontrol rules at all (Arnott 1995, p. 102).

2. Goals

Rent control is sometimes instituted to pursue goals not usually linked to tenancy laws. Prime among these goals has been maintaining the character of certain communities (recall that this is also the motivation behind much exclusionary zoning law). If a neighborhood becomes more attractive for some reason, the type of people who live there may change—yuppies replacing artists, students replacing senior citizens, immigrants replacing the native-born, conformists replacing nonconformists. The loss of consumer surplus to the original inhabitants is, of course, a real loss, and so are the externalities of moving that I have already described as the argument for stability, but some-

times proponents of rent control go further: they argue that the general public benefits from having these particular people, or this class of people, living in this particular location because of their contributions to the community's cultural life, because they provide cheap labor to certain industries, or because everyone is better off when the population is appropriately diverse.

This may be true, but rent control is not the only way to achieve this goal. Rent subsidies for residents with the appropriate characteristics, for instance, could also keep the current population in place, as could taxes on renting residents with inappropriate characteristics. Subsidies and taxes like this could be targeted more specifically than rent controls, and they would force jurisdictions to be clear about whom they wished to help or hurt. Since it is the general public that benefits from the presence of these tenants in these particular apartment buildings, there is also an element of fairness in placing the cost of keeping them there on the general public, through taxes, rather than on particular landlords through rent control. (This argument is stated eloquently in Kain 1969.)

Another goal of rent control, one more often stated by economists than by the general public, is to circumscribe the monopoly power of landlords. It may sound funny to talk about the "monopoly power of landlords," since in even a moderately sized city there are thousands of landlords and thousands more who could easily enter the business if the opportunity presented itself. Remember, though, that what we call an apartment is an incredibly complex bundle of goods, services, and attributes. Nobody rents "an apartment": they look for "a one-bedroom with good light on a low floor in a pre-war doorman building in the Village," or "a two-bedroom with a good school and a park nearby, not too far from South Orange," or some such specific description. Even though a city may have hundreds of thousands of apartments, few may be precisely what a prospective tenant is looking for, and finding even a reasonably good approximation of an ideal apartment is likely to take a lot of work. That's how landlords get monopoly power.

Think about a prospective tenant who has just inspected an apartment. To keep things simple, suppose that if this prospective tenant doesn't lease the apartment, the next best alternative is for the landlord to live in it himself. Then the social cost of the prospective tenant's leasing the apartment is the value of the apartment to the landlord with him in it. To achieve an optimal allocation of apartments, this is the amount that the landlord should ask for from the prospective tenant. But he'll ask for more.

Why? Because there's a chance he'll get more; this might in fact be precisely the apartment the prospective tenant wants, or be close enough to it that she would be willing to pay a great deal of money for it rather than keep searching. Or it might not. What matters is that by raising the asking rent by a tiny bit over his reservation value, the landlord doesn't destroy all his chances of leasing the apartment to this prospective tenant. Therefore it makes no

sense for him to ask her for no more than the value of the apartment to him—he could do just as well by ignoring her or telling her to get lost. If he doesn't tell her to get lost, he'll quote a rent above his reservation value and the social cost of the apartment. The rent he asks for, in other words, will be too high. Sometimes prospective tenants will continue searching when both they and the landlord would be better off if the landlord offered and they agreed to a lower rent. By decreasing the rent the landlord asks for, rent control can create potential Pareto improvements.

Of course, in most cases the landlord won't move into an apartment himself if a prospective tenant turns him down: he'll just keep waiting for tenants. Figuring out the value to the landlord of waiting for future tenants, and how this compares with the social cost of a tenant's moving in, is quite a complex problem, but it has been solved by Arnott and Igarishi (2000). The conclusion with the complex problem is the same as with the simple one: because landlords have monopoly power, they ask too much (and vacancy rates are too high as a result); lower rents would be a potential Pareto improvement.

The monopoly-power argument, however, supports only mild forms of rent control, since the monopoly power that individual landlords have is fairly weak. Measures that reduced tenants' search costs—for instance, extensive online catalogs of properties, with detailed information and pictures—would probably have a similar effect.

3. Disadvantages

Rent controls have all the disadvantages of leases, too—and since they are longer and stronger, they have these disadvantages more strikingly. This is especially true of hard controls and soft controls poorly administered. Indeed, American economists agree on no proposition more unanimously than that rent control has noxious effects (Alston, Kearl, and Vaughan 1992).

a. Misallocation   The primary problem with rent controls is that (like leases) they allocate apartments to the wrong people. Suppose Joe Jones is living in an apartment for which he pays $400 a month in rent, and he's willing to pay no more than $500 a month for it; meanwhile Jane Doe is willing to pay $800 a month to live there. Both would be better off if Joe moved out, Jane moved in, and Jane paid Joe something between $500 and $800 a month (while Joe continued to pay the landlord $400 a month). This would be a Pareto improvement. If the landlord raised Joe's rent to $700, threw him out, and let Jane move in, it would be a potential Pareto improvement: the landlord would gain $300 a month, Jane would be better off by $100 a month, and Joe would be worse off by $100 a month. Glaeser and Luttmer (2003) estimated that eliminating misallocations of this sort from New York City's rent control program would produce a potential Pareto improvement of around

$200 a month per apartment. Rent control makes tenants too stable (if the op-timal amount of stability is that which the market would produce).

Rent control laws as they are usually written, even most second-genera-tion ones, preclude most of the Coasian bargaining that could realize these Pareto improvements. Rent control is more than a simple, innocuous rear-rangement of property rights.

Not all rent control ordinances, however, create such serious misalloca-tion problems. Under some soft rent control ordinances with vacancy decon-trol, the landlord can successfully pay Joe to disappear and collect a higher rent from Jane, making misallocation a less serious problem.

But vacancy decontrol doesn't eliminate misallocation, because it's a risky and expensive way to get rid of Joe and bring Jane in. With vacancy de-control, the landlord has an incentive to harass Joe into leaving (by withhold-ing heat, for instance, or placing a dumpster under his kitchen window, or botching his repair requests, or acting rudely to him) instead of paying him to go. Because of this incentive, tenant groups have insisted that vacancy decon-trol ordinances contain antiharassment provisions. Since proving you didn't do something is more difficult than proving you did, paying Joe to go away is not always a clean and easy process.

The misallocation problem grows over time. Rent control tries to freeze the allocation of apartments on the date of its adoption. The more time passes, the further from optimal that allocation becomes. People get older; they find new partners; they develop new interests and habits; they change jobs and careers and hobbies; they die. Old rent control has greater problems than new rent control.

b. Quality   Rent control reduces the incentives for landlords to maintain buildings and apartments without correspondingly raising the incentives or abilities of tenants to do so. This happens in two different ways.

First, consider short-term maintenance issues like sweeping hallways, putting garbage out regularly, letting tenants in when they lock themselves out, and labeling the mailboxes informatively—the kind of activities whose benefits are unlikely to extend beyond the tenure of current tenants. So long as the current rent-controlled tenants are getting a great deal, relative to their al-ternatives, they will not move if the landlord provides these services less dili-gently, and even if they do move, other tenants will clamor to take their place immediately. Reducing the quality of these services will lose the landlord no revenue. Thus you would expect him to do so.

Second, consider maintenance issues like replacing an elevator or up-grading windows or training a new superintendent—activities likely to benefit tenants who move in after current tenants move out, as well as the current tenants. Under hard rent control, the maximum rent those future tenants can

legally pay has already been set by the rent control laws, and it's low enough that vacancies are not going to be a problem. Since doing this maintenance will neither increase the rent future tenants will pay nor decrease the rate of vacancies the landlord will encounter, he has no reason to do it. Frankena (1975) goes so far as to show that hard rent control may even increase the price tenants pay for an apartment of given quality.

Many varieties of soft rent control try to mitigate both of these under-maintenance problems, but they don't always succeed. Vacancy decontrol lets landlords charge future tenants what the market will bear, and so forces them to bear all the consequences of the maintenance decisions that affect future tenants. But vacancy decontrol doesn't alleviate the maintenance problems with current tenants.

Various "pass-through" and "hardship" provisions, on the other hand, are designed to deal with the maintenance issues of both current and future tenants. These allow the landlord to increase rents to "recoup the costs" of maintenance, repair, and renovation activities; they make the rents the landlord collects depend on what he spends and can document. But they don't make rents depend on the right costs in the right way. First, landlords don't get any reimbursement for costs they can't document—like staying up late to make sure a boiler is fixed right or forgoing hiring a needy cousin as superintendent when a more qualified stranger is available. Second, under many ordinances a landlord gets reimbursed for maintenance costs that the average landlord incurs, not the costs that he himself incurs. This procedure may be fair, and it's a good way to cut down on paperwork, but it gives the landlord no incentive to incur the expenses himself. Finally, and most seriously, what a landlord should get for making an improvement or performing a maintenance task is not how much it cost him but how much it benefits tenants (at the margin). You should get rewarded for providing benefits, not for incurring costs. If rent control pass-through provisions don't give the landlord a percentage of the cost as profit or overhead (or kickbacks from contractors), then he still has no incentive to make beneficial improvements. If they do give him a percentage, then he has no incentive not to make improvements that are not beneficial.

c. Quantity   Rent control also affects the quantity of apartments available by changing incentives to build or abandon units. Once again, hard rent control is different from soft.

Hard rent control unambiguously decreases the number of units available. The lower the prospective rents an owner can get, the less willing he will be to build a new apartment building, and the more willing he will be to abandon an old one. Hard rent control prevents obvious Pareto improvements: some tenants, as well as the landlord, would be better off if they could pay

higher rents and live in the building that would otherwise be abandoned or not be built, but the laws do not permit them to. Cities that have had many decades of hard rent control, like London, often have very small stocks left of private rental housing.

Soft rent control ordinances contain many provisions designed to alleviate the construction and abandonment problems. Because construction and abandonment are gross, obvious, easily observed, and easily verified, administrative rules probably work better on correcting these decisions than they do on the more subtle decisions about maintenance and repairs.

Most soft rent control ordinances exempt new construction and so, to the extent that the implicit promise never to impose controls on these units is credible, allow the sort of simple bilateral Pareto improvements that were the main difficulty with the construction implications of hard controls. Similarly, hardship rent-increase provisions are designed to keep any building from going under that would otherwise be able to attract some tenants willing to pay enough to keep it afloat. But this does not imply that soft controls result in the right kind and mix of new construction and abandonment. They may even promote too much construction and too little abandonment.

On construction, a little example is helpful. Suppose that originally, before rent control, a city has 100 apartments inhabited by 100 tenants. Each tenant is willing to pay $400 a month, pays $400 a month, and nobody else wants to live in an apartment in this city. Then the city becomes more attractive to a second group of 100 prospective tenants. The prospective newcomers are willing to pay $800 a month for an apartment, but the old-timers are still willing to pay only $400. New apartment construction breaks even at rent of $600 a month.

Without rent control, the old-timers leave or move to owner-occupied houses, and the newcomers move in and pay up to $800 a month in rent. No new apartments are built because nobody who doesn't have an apartment is willing to pay $600 a month or more. Alternatively, suppose the city enacts a soft rent control ordinance before the newcomers arrive, caps rent on existing apartments at $400 a month, and exempts new construction. Then 100 new apartments will be built for the newcomers, who will pay more than $600 a month for them because the old-timers will stay in place.

In this example, soft rent control means that *more* apartments are built, not fewer. But these are apartments that should not be built. It's cheaper to produce a place for a newcomer to live by paying an old-timer $400 to go away than by paying a contractor $600 to build. Everyone would be better off if the newcomers sublet apartments at $500 a month from the old-timers and nothing new were built. Construction is not the goal of life.

In summary, then, tenancy laws and rent control, both hard and soft, address real problems in the housing market. They also create problems. The

question about them is not whether the problems they create are bigger than the problems they solve, but whether there are better ways of dealing with the problems they try to solve.

V. Housing Finance

In section IV, we saw that tenants are really in many ways mini-owners: leases give them permanency and protection, tenancy laws make leases stronger than they might otherwise be, and in some cases rent control strengthens leases even further. The polar case of pure tenancy—living at the landlord's whim—isn't attractive to most people and has other social costs as well.

In this section we'll see that the same is true, on the other end of the spectrum, with the case of "pure ownership." Most owners are really in some way "mini-tenants." They don't really have the full stake in the community and the house and the full control of the house that I portrayed them as having in section II.

A. Understanding Mortgages

To understand why they don't, and why they usually don't want it, recall the description of housing as a bundle of radiators and doors, windows and roofs, driveways and school admissions. That description ignored the temporal aspect of homeownership. When you consider time and uncertainty, you realize a house is an even bigger and more complex bundle than you thought before. Owning a house means not only that you control the radiators and the driveway today, but also that you control them fifty years from now (if you and they are still around); you control them a year from now if it's sunny, and you control them if it's cloudy.

When you consider a house as a bundle of contemporaneous stuff, the owner has some flexibility about what's included in that bundle. If you don't want a Jacuzzi in the bathroom or a statue of the Blessed Virgin Mary on the front lawn, you don't have to buy one. If the house came with it, you can hold a yard sale and get rid of it. But there are limits on how much you can unbundle and there are costs from unbundling: it's really difficult to keep the first and third floors of a house without the second, or to buy only every other foot of the telephone wire. It's only because there are advantages to bundling things together that we have housing—but it's not always clear what should be in the bundle and what should not.

The same is true for housing as a temporal bundle. I could control the house in even-numbered minutes and you in odd, but slicing the temporal bundle that way would present huge problems (although some condominiums and hotel rooms in resorts are governed by time-sharing contracts not much different from this). The appeal of owning a house, no matter what, un-

til you sell it, is that it improves your incentives and gives you a bigger stake in the community.

But most people don't want to own a bundle that big—not all by themselves. Caplin and colleagues (1997, p. xii) note that if the market for breakfast operated like the market for housing, people would buy perpetual supplies of bacon and eggs before they ate breakfast, and then they would have to sell their supply of bacon and eggs and pay a high commission if they decided they preferred cereal.

Because each house is a perpetual supply of housing services, houses are expensive. Most people borrow money to finance buying a house. They borrow through a **mortgage**—a loan for which the house is collateral. If a homeowner defaults on a mortgage, the mortgage holder can force the sale of the house. Most homeowners have mortgages—63 percent in 1999, including 79 percent of those who had moved in the last year (U.S. Bureau of the Census 2001, table 960)—and so only a minority of houses are owned "free and clear."

What does having a mortgage mean? For one thing, it means that the homeowner must abide by rules that are included in the mortgage. For instance, under the "doctrine of waste" the homeowner is obligated to maintain the house in good repair; the mortgage holder can require immediate repayment of the full loan if he does not. The homeowner must also buy hazard insurance and pay property taxes. Mortgages often carry restrictions on subleasing, too, but these restrictions are not always enforced or enforceable.

A mortgage also implies that the homeowner does not have to bear the full burden of many bad events, or the full consequences of his actions if these bad events occur. Suppose you buy a $150,000 house with a $120,000 mortgage, and two weeks later the value of the house falls to $100,000. If you continue to pay the mortgage, you have lost $50,000. But you can cut that loss to $30,000 by walking away from the mortgage. You will lose the house, which is bad, but you'll lose the debt, too, which is good. Losing the house costs you only $100,000, but losing the debt gains you $120,000.

Thus, the option to default makes a mortgage something of an insurance policy. It's insurance against any big enough loss in the value of a house, not just those caused by (insurable) natural disasters. Like all insurance policies, this creates moral hazard problems. Building a new deck or campaigning hard against the nearby location of a toxic waste dump makes no sense if you're going to collect the insurance anyway by losing the house. Letting the paint fall off and the termites destroy the porch loses you nothing if you're going to default soon. Some of the rules, like the doctrine of waste, that are included in mortgage contracts are designed to reduce these moral hazard problems, but they don't eliminate them.

Thus, a mortgage means that you don't control the house, no matter what; you don't bear the burden of whatever you do, no matter what; and you

don't have an inescapable stake in the community, no matter what. Your own-
ing a home with a mortgage provides the rest of the world with less of many
of those beneficial externalities that the homeowners' tax preferences were de-
signed to promote and provides you with fewer of the efficiency incentives for
good internal management. This is the big trade-off that policies toward
mortgages must deal with: the more insurance mortgages provide, the less
each homeowner has of that robust independence that ownership is supposed
to be about; but the less insurance mortgages provide, the fewer the number
of people willing to become homeowners.

Potential homeowners want mortgages for two reasons. The first is the
obvious problem of cash flow: it's difficult to save enough money to buy a
house free and clear. The second reason is more subtle: houses are risky invest-
ments.

Caplin and colleagues (1997) give three reasons why houses are risky.
First, house prices are volatile: the average house in a large metropolitan area
lost value in a five-year period about 20 percent of the time, if you count turn-
over costs, in the years between 1975 and 1994, and since almost no one owns
the average house, the degree of risk on a particular house that you might own
was considerably larger.

Second, house prices are correlated with income: the correlation between
house price appreciation in a metropolitan area and wage growth was about
0.23 (Caplin et al.). House prices are likely to go down at the worst possible
time for you—when you lose your job, or when your wages are stagnant or
declining because other people are losing their jobs.

And third, a house is a risky investment because it's hard to diversify.
Normally you would want an asset that was risky and correlated with your
wage income to be a small part of your portfolio; that way, if something went
wrong you wouldn't lose too much. "Don't put all your eggs in one basket," is
standard advice in financial planning. But houses are so expensive that for
most people, owning a house and being diversified are incompatible goals.
"Across the age spectrum, the average proportion of household (non-pension)
assets tied up in housing is extremely high, exceeding 90 percent for younger
households and never dropping below 85 percent for older households . . .
[F]or a great many households, housing is the *only* significant asset in their
portfolio" (ibid., p. 80).

B. How Governments Promote Mortgages

Since most mortgages are provided by what looks like a normal private, com-
petitive marketplace, you might be tempted to conclude that the market must
have worked out the optimal trade-off between risk and responsibility by this
time. This conclusion would be incorrect. Governments have been actively in-
volved in mortgage markets since the Great Depression, and even if there were
no government involvement, private mortgage markets might not produce a

Pareto optimum. The main reason for this is adverse selection, just as in the rental market with leases and tenancy laws. Any single lender who starts offering more generous mortgage terms than what other lenders are offering is likely to attract a disproportionately large share of borrowers who will default and who will not maintain their properties well.

The history of government involvement in mortgage markets is the story of a slow but relentless push toward less risk and less responsibility.

1. Before Government Involvement

In the old days, before the 1930s, mortgages were short and small—"balloon loans" is the technical term. Loans were made for no more than 50 percent or 60 percent of a property's value, and they lasted no more than five years. During the term of the loans, borrowers didn't pay off any principal; they just paid interest. Most lenders specialized in particular geographic areas where they knew the housing market well, understood the risks, and believed they could assess the character of potential borrowers. At the end of a five-year term, the entire principal would come due, and borrowers typically scrambled around to refinance—they took a new loan, used it to pay the principal on the old loan, and repeated the process with the new loan.

This system placed a lot of risk and responsibility on the homeowner. You had to have saved a lot of money and developed a good reputation before you could buy a house, and once you did, you were, for the most part, stuck with it. House values had to go down a lot—at least 50 percent—for defaulting to start looking attractive, and they had to drop quickly, in less than five years. If you didn't maintain a house well, you had to pay more of your own cash when it came time to refinance. The same was true if you were unlucky and property values were low when it came time to refinance.

2. The Depression

The Depression put an end to this system. With demand for housing low, property values plummeted. When loans came due, they were hard to refinance because property values were low. Owners couldn't come up with cash because wages had fallen, and many were out of work. Vast numbers defaulted and lost their homes. When the borrowers went under, they took the lenders with them. As lenders went out of business or stopped lending, property became even more worthless, demand fell, more owners defaulted, and the downward spiral intensified. New construction ceased. The mortgage system made bad times worse for almost everyone concerned.

The major New Deal initiative designed to alleviate the mortgage crisis was the Federal Housing Administration (FHA), established in 1934. The FHA insures mortgages; it doesn't make loans. But it revolutionized the mortgage market.

It did this by changing the standards for mortgages. The FHA wouldn't insure just any mortgage; instead it insisted on a new kind of mortgage. FHA

mortgages were for long terms—up to twenty years at first, and later, thirty years—and they represented large proportions of the cost of a house, up to 80 percent at first. Equal monthly payments had to be charged for both principal and interest so that when the term of the mortgage ended, nothing was owed. The FHA added a fee to the monthly payments to cover the cost of the insurance it provided, and so the agency was designed to break even, without using tax dollars. (Most of the time it has done so and even shown a profit.) The FHA's goal was to get the mortgage market moving again.

At first lenders didn't like this new type of mortgage. The housing market remained in the doldrums, as lenders did not find credible the FHA guarantee that they would be reimbursed in the event of a default. So in 1938 the government set up another New Deal housing program, the Federal National Mortgage Association (FNMA)—better known as Fannie Mae. Fannie Mae's initial job was to buy all qualifying mortgages from the lenders who originated them. To get the money to do this, Fannie Mae sold notes and bonds to private investors and promised to pay them back. It would get the money to pay back those investors when the original borrowers made the monthly payments on their mortgages.

This is called creating a secondary market. Without Fannie Mae, if the First National Bank of XYZ City gave John Jones an FHA mortgage for twenty years, it would have to hold that mortgage for twenty years. It would not have to worry about default, since the FHA guaranteed the loan, but it would not be able to use the money it had lent Jones for any other more lucrative opportunity that might come along. If interests rates went up, the bank would be in big trouble: its depositors would expect to be paid high interest rates, but it would have no way to pay those rates because its portfolio would include nothing but mortgages paying low rates. (This was probably another reason why lenders were reluctant to write FHA mortgages before 1938: they didn't want to be saddled for twenty years with instruments that were paying only low, Depression-era interest rates; this was before adjustable-rate mortgages.) With Fannie Mae, though, the bank can sell the mortgage, get back almost as much money as it lent Jones, and let Fannie Mae worry about holding the mortgage. Fannie Mae made lenders want to write FHA mortgages, and after World War II they became popular.

Why didn't some smart bank start selling its mortgages on the secondary market to reduce its exposure to interest-rate risk before Fannie Mae came along? Once again the answer is probably adverse selection. If you're the only bank in the country trying to unload your mortgages, buyers would rightly question the quality of your judgment; your willingness to sell mortgages would be a signal that the mortgages were of low quality.

3. World War II and Its Aftermath

World War II also brought the institution of a second government program for the primary mortgage market. In 1944 the GI Bill of Rights set up a series

of ways for the government to help returning military personnel. One of them was the Veterans Administration (VA) mortgage guarantee.

Unlike FHA mortgage insurance, the VA guarantee was not designed to revive the housing market; instead its primary goal is to help veterans. The VA guarantees only a portion of a mortgage—originally $2,000 but it was almost immediately raised to $4,000, and it has increased many times since. In effect, it provides veterans with down payments; in many cases veterans are able to use the guarantee to buy a house with no money down. The VA mortgage guarantee made it easier for many people to buy houses, but it also reduced the robust independence that homeownership is supposed to foster.

When the housing market revived after World War II, mortgages—and not just FHA and VA mortgages—were different from those that were written before the Depression. The FHA and VA segments were large but limited—the VA served only veterans, for instance, and both government programs were effectively limited to houses below the median in price. The conventional market (everything else but FHA and VA), however, did not return to the short balloon mortgages of the 1920s. Instead, conventional lenders wrote long, big amortizing mortgages just like the FHA. The private market imitated the government—that's the real FHA revolution.

Why did this happen? One possibility is simply learning: maybe nobody had thought of big, long mortgages before, or nobody had tried them out. But in a country with tens of thousands of mortgage lenders, most of them looking for some sort of edge, this is not plausible. More plausible is the adverse selection story. Suppose the number of borrowers who would abuse a guaranteed, big, long mortgage is rather small relative to the entire market, but huge relative to any single private lender. Then, as we noted before, any single private lender who tried to write FHA-style mortgages would be swamped, but a huge operation like the FHA, or a huge operation whose clients were not self-selected, like the VA, could write such mortgages without being swamped by borrowers likely to default. And once the FHA and VA were in operation, they took most of the risky borrowers whom private lenders feared; thus private lenders could offer the big, long, amortizing mortgages that most borrowers preferred, without fear of being overrun by risky borrowers. In 2000, for instance, only 2.5 percent of conventional mortgages were delinquent, but 6.8 percent of VA loans and 9.1 percent of FHA loans were (U.S. Bureau of the Census 2001, table 1187). In this interpretation, then, the FHA and VA were a potential Pareto improvement.

4. The Rise of the Secondary Mortgage Market

Since the end of World War II, the major development in the U.S. mortgage market has been the emergence of a flourishing secondary market. This, too, was the result of government initiatives. Fannie Mae in 1938 was the beginning of a secondary market. But until the early 1970s, Fannie Mae was just

about the only buyer of mortgages, and practically nobody bought from Fannie Mae.

That changed in the early 1970s. Legislation enacted at about this time did several things: it split two other entities off from Fannie Mae, the Government National Mortgage Association (GNMA or Ginnie Mae) and the Federal Home Loan Mortgage Association (FHLMA or Freddie Mac); it gave Fannie Mae and Freddie Mac greater independence from the government than Fannie Mae had ever had before, making them in many ways private companies; and it authorized Fannie Mae and Freddie Mac to buy (and sell) conventional mortgages. The splitting was not an economically important move—Ginnie Mae was set up to concentrate on FHA and VA mortgages, and Freddie Mac to deal with savings and loans. The independence and authorization allowed all three entities—they're collectively referred to as government sponsored enterprises, or GSEs—to start issuing mortgage-backed securities, and when they did, the rest of the market followed suit.

Mortgage-backed securities work like this: Ginnie Mae, say, will buy a large number of mortgages. It will then pool them together and sell the resulting pool to private investors. Ginnie Mae will pass all payments of principal and interest to the ultimate investors; Ginnie Mae will also guarantee these payments on a timely basis even if the borrowers default. Investors face two kinds of risk—interest rates may change, and borrowers may repay principal more quickly or more slowly than anticipated—but they do not face default risk. Since investors buy parts of pools of mortgages, not individual mortgages, the first kind of risk is reduced. Issuing mortgage-backed securities has become an extremely profitable business for the GSEs.

Once the GSEs started issuing mortgage-backed securities, a market developed. People got used to buying and selling them, learned how to price them, learned who would buy what. When the market developed, private entities started issuing mortgage-backed securities, too, especially for types of mortgages, like "jumbo mortgages," that Congress had declared out of bounds for GSEs. Nowadays it's unlikely that the institution you get a mortgage from will hold it for more than a few weeks.

Once again, then, a government initiative led to private imitation, and probably to a potential Pareto improvement. (The story is probably about the same for commercial mortgage–backed securities, a market that developed in the late 1980s and early 1990s. When many savings banks failed in the 1980s and the government set up an agency, the Reconstruction Trust Corporation, to handle their assets, which included many commercial mortgages, there was one player big enough to overcome adverse selection problems and get the market started.)

The government is much less dominant in the private market now. Of mortgages outstanding in 1999, only 14 percent were insured by the FHA and 5 percent were guaranteed by the VA (U.S. Bureau of the Census 2001, table 960).

5. The Consequences

Since the Depression, then, Americans have moved a long way from the robust independence that used to characterize homeownership. Owners now make much smaller down payments and have much longer to pay; this gives them much smaller stakes in their communities. The secondary market has reduced the stake that lenders hold in their community, too. When a few local banks or thrifts held most of the mortgages in a particular neighborhood, they too had a stake, a reason to lobby to preserve and protect that neighborhood. With many mortgages guaranteed or insured, and with most mortgages incorporated into large and anonymous pools, this reason no longer has great force.

The evolution of the mortgage market has thus weakened the case for ownership tax preferences in two ways. First, when adverse selection limited the mortgages available and no secondary market was around to reduce lenders' interest-rate exposure, an argument could have been made that the preferences were needed to offset the distortions that kept money from flowing into the mortgage market. Such an argument is much more difficult to make today than it was in 1940 or even in 1970. Second, many of the arguments for the external benefits of homeownership relied on ownership "free and clear," or at least close lender supervision and concern. The trend of smaller down payments, longer terms, more guarantees and insurance, and an active secondary market has made that kind of big stake much rarer. Homeowners are much more like tenants today than they used to be (and the strengthening of tenancy law in many places has made tenants more like owners). The social gain of moving someone from tenancy to ownership has decreased.

The federal government also subsidizes homeowners and homeownership indirectly through the GSEs; this is a subsidy in addition to those in the tax preferences. Even though nothing in writing says that the federal government guarantees the debt of the GSEs, their close relationship to the government and their huge size has led many investors to conclude that the government would bail them out if they were likely to default. As a result, they can sell bonds at lower interest rates. Passmore (2003) estimates that these lower rates save the GSEs between $120 billion and $160 billion a year in interest costs. Most of these savings inure to the benefit of GSE stockholders, but homeowners end up paying around $60 billion a year less because of slightly lower interest rates. Although the increase in homeownership that this subsidy causes is probably small, as is the increase in investment that homeowners make as a result, still it causes a diversion of capital from higher return investments that don't carry an implicit government guarantee.

C. Policy Directions

Most of the proposals for future evolution of the mortgage market are for further moves toward lessening homeowners' risks. Owning a house is risky for many owners because they can't diversify their portfolios. A large portion of

the wealth of most homeowners is tied up with the fortunes of a particular small piece of real estate at times in the future when they won't even be living there. Hardly anyone would willingly invest that way, and so owners are directly worse off as a result.

This portfolio overconcentration causes more indirect difficulties, too. Fischel (1999) points out that it makes homeowners adamantly opposed to any proposals that stand even a chance of reducing future home values. Most NIMBY controversies are about projects whose consequences for property values are uncertain. No one knows for sure what effect a high school down the street or a halfway house next door or a garden apartment two blocks away will have on the value of my house five or ten years from now. But if I'm counting on the sale of that house to finance my retirement, losing a significant portion of that value would be truly horrible—much worse than a comparable gain would be good for me. I would not be willing to accept a fifty-fifty gamble between a 20 percent loss and a 20 percent gain. Thus portfolio overconcentration makes homeowners much more opposed to any projects that change the status quo than they would be if they held more diversified portfolios. Their stake in the community—the community as it now stands—is too big.

Two proposals have been designed to address these problems. Shiller and Weiss (1994) proposed developing a housing price index futures market. This would allow investors to place bets about changes in average housing prices in the nation or in a region. Homeowners could use this market for hedging rather than gambling. When you bought a house, you'd be able to (but free not to) place a bet simultaneously that some relevant average of housing prices would go down. Then if housing prices went down, and yours went down with the rest, you wouldn't lose so much because you'd win on your bet while you lost on your house. If housing prices went up, you wouldn't win so much because you'd lose on your bet while you won on your house.

The beauty of a price index like this is that it makes owning a house less risky without reducing the incentives for care, maintenance, and wise investment. The drawback is that it does nothing about NIMBY or purely localized price fluctuations.

The other proposal, housing partnerships, does more to address these latter problems but reduces incentives more. Caplin and colleagues (1997) provide a detailed explanation of this idea. As they phrase it (p. 6):

> We propose that housing be financed with not only a mortgage but also an institutional investor that provides equity capital for the house in exchange for a portion of the ultimate sale price . . . [T]he basic financial transaction would involve the [institutional investor's] supplying funds up front in exchange for a fixed proportion of the ultimate sale price of the house, with no other monetary payments made between the parties. The [household] would retain complete control and use of the space for an unlimited

amount of time and would be penalized only for decisions that harm the [institutional investor], such as failing to maintain the property.

This proposal would obviously mitigate portfolio overconcentration and both its direct and indirect effects.

The drawback of housing partnerships is that they reduce the occupant's incentives to make minor improvements that increase the house's long-term value—sealing the driveway, attending a board of adjustment meeting. (Major improvements can be handled by negotiations with the institutional investor—perhaps better than they are currently handled, because the institutional investor would be more savvy than the average homeowner.) Housing partnerships reduce these incentives, but they do not eliminate them. Such reduced incentives, though, have been the hallmark of mortgage policy for most of this century. Housing partnerships would just be another step on a road the United States has been traveling for quite some time.

Notice, though, that the income tax system as it now stands discourages partnerships and forces portfolio overconcentration. Debt financing is tax deductible; equity financing isn't. If you borrow $100,000, you can deduct the interest you pay on it, but if you give somebody a share of your house with an expected present value of $100,000, the federal government contributes nothing.

VI. Conclusion

In this chapter we have looked at two chief questions: whether people are buying enough housing, and whether the rules under which they acquire that housing are giving them the right incentives to be good citizens. Good citizenship includes such ideas as maintaining property, helping children grow and develop, participating in the community, and putting down roots in one place.

Quite a few different public policies promote the ends of ownership and community responsibility: income tax preferences that encourage ownership and housing consumption, tenancy laws and rent controls that make tenants more stable and give them a stake in the community, and a series of initiatives in the mortgage market that have made it easier for people to become owners in name but not completely in fact. After considering all these policies, it doesn't appear that there is any shortage of housing in the aggregate; in fact, a good case can be made that the Unites States has too much lawn, too many fireplaces, too much electrical wiring, too little use of community pools and movie theaters, and too much fear of new land uses in a neighborhood. In many cases, we identified better ways of accomplishing the ends that the housing policies are supposed to be working toward.

The income tax preferences, we saw, encourage excessive sprawl because they encourage use of land and garages. Tenancy laws could work either way: if they make the rental market work better (by providing security of tenure,

for instance, that landlords could not credibly commit to on their own), they allow more people to live in the relative density of apartment buildings and so reduce sprawl. But if they drive landlords out of the market, as hard rent control can, they make people live farther apart than they should.

Mortgage programs have a similarly mixed record on sprawl. The VA mortgage guarantee no doubt encourages excessive sprawl, though in a minor way, by subsidizing a portion of some of the defaults that occur on lawns and garages. The veterans would have been better off with cash instead of a guarantee, and more green space would be left. If you ignore income tax preferences, the major thrust of the mortgage policies—establishing long, big mortgages and a secondary market—would clearly be a potential Pareto improvement, but one that contributed to sprawl. The mortgages would be like televisions, refrigerators, and frozen food—other inventions that made the world a better place and more sprawl desirable. The presence of the income tax preferences tempers this conclusion: if you already have too much owner-occupied housing, alleviating the unnecessary borrowing constraints that hold down housing demand only makes a bad situation worse. But even if this is the case, it identifies the income tax code, not the mortgage initiatives, as the major problem.

One question we have not addressed is who gains and who loses from the housing policies we have examined. This is a difficult question: housing is such a large part of national wealth that changes in housing markets have big ramifications throughout the economy, and tracing through all these ramifications is hard. Most of the people who would pay more taxes tomorrow, for instance, if the homeowner tax preferences were removed by surprise, are pretty well off (Follain, Ling, and McGill [1993] estimate that 32 percent of the extra money under this scenario would come from the richest tenth of households); but this is not the important question.

The important—and really difficult—question is, who would be better off or worse off after markets adjusted? Would factories see more investment, and would this push up workers' wages? If so, which workers' wages? Would investment move to developing countries instead? Or would savings fall? Economists have tried hard to answer these questions, but no consensus has emerged. What may have been the right answers in 1980 probably aren't the right answers for 2010.

In the next chapter, I will turn to a much more limited version of this question: how the policies discussed in this chapter have affected the price, quality, and quantity of housing that poor people use. Housing for poor people has been a major concern of the public and of governments ever since large Western cities grew up in the nineteenth century. Many policies have been designed explicitly with poor people in mind, and these, too, will be examined. But you can't understand how those policies work unless you understand the much bigger programs we've analyzed here.

Questions

1. Suppose the marginal rate of return to capital invested in U.S. businesses K_b is

$$r_b = 20 - (2/3)K_b$$

where r_b is a percentage (that is, $r_b = 5$ means a 5 percent rate of return) and capital is measured in trillions of dollars. The marginal rate of return on housing investment is

$$r_h = 11 - (1/2)K_h.$$

The total amount of capital that can be invested in the United States is $25 trillion.

a. There is a 40 percent tax on capital income in the business sector and no tax in the housing sector. How much capital is invested in the business sector? In the housing sector? What is the pretax marginal rate of return in the business sector? In the housing sector? What is the pretax average rate of return in the business sector? In the housing sector? What is the post-tax marginal rate of return in the business sector? In the housing sector?

b. Suppose that a tax of 40 percent is instituted on returns to capital in the housing sector. How much capital shifts from housing to business?

c. What is the deadweight loss on an annual (flow) basis from the misallocation of capital stock?

d. How would the consequences of homeowners' tax preferences be different if capital were perfectly elastically supplied on the world market at a post-tax rate of return of 6 percent?

2. Discuss the effect of permitting tenants in rent-controlled apartments to sublet their apartments to anyone they want under any terms they wish. Would this be a Pareto improvement? A potential Pareto improvement? Who would gain? Who would lose? Would this be a good public policy? Why or why not? Would there be a way for the winners to compensate the losers? If free subletting were permitted, would economists have any objections to rent control?

3. Discuss the effect of establishing for each rent-controlled apartment a "right to rent control" certificate and awarding it to the incumbent tenant. Should the certificates be fungible? Who would buy them? What would the effect be? How should such a market be regulated? Should the imputed value of such certificates be counted as income for the holder, for income tax purposes?

4. There are ten houses on Pierson Road. Each is worth $400,000. Each homeowner acts as if she were maximizing the expected value of the natural logarithm of her terminal wealth. A new sewer for Pierson Road has been proposed. If approved, it will be built for free by a cult of sewer builders. If it is

built, nine homes will increase in value by $10,000 and the tenth will lose $80,000 in value because of accidental floods. No one knows which is the tenth and each home is equally likely to be the unlucky one. For the sewer to be built, unanimous approval is required.

a. Is constructing the sewer a potential Pareto improvement?

b. Suppose each homeowner owns her home free and clear and has no other assets. Will the sewer be approved?

c. Suppose each has a $300,000 mortgage. Will it be approved?

d. What institutional arrangements or markets would improve the collective decision making?

References

Aaronson, Daniel. 2000. "A Note on the Benefits of Homeownership." *Journal of Urban Economics* 47:356–369.

Alston, Richard M., J. R. Kearl, and Michael B. Vaughan. 1992. "Is There a Consensus among Economists in the 1990s?" *American Economic Review* 82(2):203–209.

Arnott, Richard. 1995. "Time for Revisionism on Rent Control?" *Journal of Economic Perspectives* 9:99–120.

Arnott, Richard, and Masahiro Igarishi. 2000. "Rent Control, Mismatch Costs, and Search Efficiency." *Regional Science and Urban Economics* 30:249–288.

Boehm, Thomas P., and Alan M. Schottman. 1999. "Does Home Ownership by Parents Have an Economic Impact on Children?" *Journal of Housing Economics* 8:217–222.

Borsch-Supan, Axel. 1986. "On the West German Tenants' Protection Legislation." *Journal of Institutional and Theoretical Economics* 142:380–404.

Caplin, Andrew, et al. 1997. *Housing Partnerships: A New Approach to a Market at a Crossroads*. Cambridge, MA: MIT Press.

DiPasquale, Denise, and William Wheaton. 1996. *Urban Economics and Real Estate Markets*. Englewood Cliffs, NJ: Prentice-Hall.

Ellen, Ingrid Gould. 2000. *Sharing America's Neighborhoods*. Cambridge, MA: Harvard University Press.

Fischel, William A. 1999. "Robert Nelson's 'Privatizing the Neighborhood,'" *George Mason Law Review* 7:881–903.

Follain, James R., David C. Ling, and Gary A. McGill. 1993. "The Preferential Income Tax Treatment of Owner-Occupied Housing: Who Really Benefits?" *Housing Policy Debate* 4(1):1–24.

Frankena, Mark. 1995. "Alternative Models of Rent Control." *Urban Studies* 12:303–308.

Glaeser, Edward, and Matthew Kahn. 2003. "Sprawl and Urban Growth." National Bureau of Economic Research Working Paper 9733.

Glaeser, Edward, and Erzo Luttmer. 2003. "The Misallocation of Housing under Rent Control." *American Economic Review* 93(4):1027–1046.

Green Richard K., and Michelle J. White. 1997. "Measuring the Benefits of Homeowning: Effects on Children." *Journal of Urban Economics* 41:441–461.

Guasch, J. L., and R. Marshall. 1987. "A Theoretical and Empirical Analysis of the

Length of Residency Discount in the Rental Housing Market." *Journal of Urban Economics* 22:291–311.

Harkness, Joseph, and Sandra Newman. 2002. "Homeownership for the Poor in Distressed Neighborhoods: Does This Make Sense?" Paper presented at conference, "Policies to Promote Affordable Housing," sponsored by the Federal Reserve Bank of New York and New York University School of Law, February 7.

Haurin, Donald, Tobey Parcel, and R. Jean Haurin. 2000. "The Impact of Home Ownership on Child Outcomes." Social Science Research Network Electronic Paper Collection, *http://papers.ssrn.com/paper.taf?abstract id=218969.*

Hubert, Franz. 1990. "Tenure Security for Tenants." Working paper, Freie Universitat Berlin.

———. 1995. "Contracting with Costly Tenants," *Regional Science and Urban Economics* 25:631–654.

Kain, John. 1969. "An Alternative to Rent Control." Policy Note 77-2, Department of City and Regional Planning, Harvard University.

Mills, Edwin S. 1989. "Social Returns to Housing and Other Fixed Capital." *AREUEA Journal* 17(Summer):197–211.

Passmore, Wayne. 2003. "The GSE Implicit Subsidy and the Value of Government Ambiguity." Working Paper 2003-64, Federal Reserve Board of Governors, Washington, DC. Accessed at *www.federalreserve.gov/pubs/feds/2003/200364/200364pap.def.*

Shiller, Robert J., and Allan N. Weiss. 1994. "Home Equity Insurance." Discussion Paper 1074, Cowles Foundation for Research in Economics, Yale University.

Taylor, Lori L. 1998. "Does the U.S. Still Overinvest in Housing?" *Economic Review* (Federal Reserve Bank of Dallas) 2:10–18.

Voith, Richard, and Joseph Gyourko. 1998. "The Tax Treatment of Housing: Its Effects on Bounded and Unbounded Communities." Working paper, Federal Reserve Bank of Philadelphia and University of Pennsylvania.

U.S. Bureau of the Census. 1999. "Housing Then and Now: 50 Years of Decennial Censuses." Accessed at *www.census.gov.*

———. 2001. *Statistical Abstract 2001.* Washington, DC: Government Printing Office.

———. 2002a. "American Housing Survey." Accessed at *www.census.gov/hhes/www/housing/ahs99/.*

———. 2000b. *Statistical Abstract 2002.* Washington, DC: Government Printing Office.

———. 2004. *Statistical Abstract 2004.* Washington, DC: Government Printing Office.

U.S. Bureau of Economic Analysis. 2005. "National Economic Accounts: Fixed Asset Table." Accessed on February 11, 2005, at *http://www.bea.gov/bea/dn/FA2004/TableView.asp.*

Wang, Ko, et al. 1991. "The Impact of Rental Properties on the Value of Single-Family Residences." *Journal of Urban Economics* 30:152–166.

Woodward, Jeanne, and Bonnie Damon. 2001. "Housing Characteristics: 2000." Accessed on September 22, 2002, at *www.census.gov/prod/2000pubs/ckbr01–13.pdf*

Yun, Heesuk. 2003. "Empirical Investigation of Dissaving Near the End of Life." PhD diss., Department of Economics, Columbia University.

Chapter 14

Housing and Poor People

*I*n *Chapter 13 we saw* how some housing policies have been designed to change the kind of people Americans are—to make them better parents, better citizens, and better neighbors. The same policies, however, generally don't have a direct effect on poor people. Different housing policies are designed to make poor people better parents, citizens, and neighbors. This chapter is about them.

Housing policies for poor people differ in a few ways from the policies we have already examined. First, health plays a much greater role: part of being a good neighbor is not spreading disease. The more people there are in close contact with each other in the same house, the faster various communicable diseases—like tuberculosis—spread, and the faster they spread, the more likely you are to get them, no matter where you live. So, for instance, cities have adopted regulations about indoor plumbing, ventilation, and sunlight. Before such regulations, housing conditions for the poor were dire, and the effects were harshest on the most vulnerable, the young and the frail. One of New York's early housing reformers, Jacob Riis, emphasized this when he wrote in 1890:

> In this block between Bayard, Park, Mulberry, and Baxter Streets, . . . the late Tenement House Commission counted 155 deaths of children in a specimen year . . . Here in this tenement, No. 59½, next to Bandits' Roost, fourteen persons died that year, and eleven of them were children; in No. 61, eleven, and eight of them not yet five-years-old. According to the records in the Bureau of Vital Statistics, only 39 people lived in No. 59½ in the year 1888, nine of them little children. There were five baby funerals in that house the same year. (Riis 1890, pp. 46–47)

The second difference in housing policies for the poor is that the behaviors they aim to correct are much more serious those we addressed in Chap-

---

To get the most out of this chapter, you should be familiar with these concepts: Pigouvian subsidies and the closed- and open-city cases. You can find these terms in the Glossary.

ter 13. Here, being a good citizen often means refraining from committing crimes—not attending block-watch meetings. Many of the policies are motivated by a belief that growing up in the wrong kind of house can make you likely to become a thief or a ne'er-do-well:

> [In] the tenements all the influences make for evil; because they are the hot-beds of the epidemics that carry death to rich and poor alike; the nurseries of pauperism and crime that fill our jails and police courts; that throw off a scum of forty thousand human wrecks to the island asylums and work-houses year by year; that turned out in the last eight years a round half million beggars to prey upon our charities; that maintain a standing army of ten thousand tramps with all that that implies; because above all, they touch the family life with deadly moral contagion. This is their worst crime. (Riis 1890, p. 23)

The third difference in policies for housing the poor is that helping poor people is a goal with a moral standing of its own, quite apart from efficiency. On average, poor renters paid 64 percent of their household income for rent in 2000 (Raphael and Quigley 2004, table 3); thus housing is a major component of the well-being of poor people. In most religious traditions, and in many secular ones, too, actions that help poor people are good in themselves, even if some nonpoor people lose out in the process. Therefore, assessing these policies requires that we pay a lot of attention to who gains and who loses from them, not just whether the winners could compensate the losers.

The first section of this chapter deals with housing codes, the earliest attempt to improve how poor people live. These are municipal regulations that require housing to meet certain minimum standards. We will also revisit rent control, since this often works in tandem with housing codes.

Section II is about public housing, which was first instituted in the United States in 1937. We will examine the successes of public housing as well as its failures.

Section III looks at some of the other supply-side programs that followed public housing, and how they have tried to address the management and incentive issues that public housing raises. By **supply-side housing programs,** I mean programs that try to influence directly the people who own the housing that poor people use and how they act, rather than influencing how poor people deal with the owners of the housing.

Demand-side housing programs, in contrast, try to influence directly how poor people act; they influence the behavior of landlords only indirectly. This distinction is traditional, and not precise. Section IV is about demand-side programs.

In section V we will look at how these programs (and two others we've examined elsewhere in this book) affect nonparticipants in public housing.

I will try to steer clear of two characterizations that have dominated many discussions of poor people's housing. I am not going to portray land-

lords as devils or tenants as pigs. All I'm interested in doing is arranging the incentives for people of all types and dispositions to cooperate with each other well. Deciding who should be admitted to heaven is beyond the scope of this chapter.

I. Housing Codes and Rent Control

When you see people living under deplorable conditions, regulation is the obvious response—we should make these deplorable conditions illegal. The history of government responses to poor housing begins with housing codes. New York City, for instance, banned pigs from houses in the main part of the city after the 1849 cholera epidemic; according to Burrows and Wallace (1999, p. 787), "Overcoming sometimes violent resistance by impoverished owners, the police flushed five to six thousand pigs out of cellars and garrets . . . that summer."

Today, most cities have extensive codes that regulate housing in many ways. In Newark, New Jersey, for instance, every room in a home (except the kitchen) must be at least 7 feet in every dimension, and every dwelling unit must contain at least one room of 150 square feet or more; hotel rooms must have at least 50 square feet if they are for transients, 80 square feet if they are for permanent residents. Every apartment must have off-street parking, and every hotel must have one off-street parking space for every four rooms (and one toilet, one shower, and one washbasin for every six to eight residents). Residences must have smoke detectors, and common areas must be monitored by the hard-wired variety. Chicago requires 70 square feet per room, with the window area equaling 10 percent of the floor area.

Many regulations that affect housing are not labeled "housing code" or "housing maintenance code," such as those concerned with fire safety, health, soundness of construction, plumbing, and electricity. Zoning matters, too, and it is often difficult to sort out whether a particular regulation belongs under "zoning," "housing," "building," or "fire." For instance, most cities mandate that landlords maintain minimum temperatures in all apartments. These regulations create potential Pareto improvements to the extent that they combat negative externalities. Minimum temperature regulations, for example, are designed to reduce fires and slow the spread of communicable diseases. Such regulations sometimes fail, however, because they are captured by groups that use them for other purposes (as we saw in Chapter 8).

The new question I want to ask in this chapter is, What effect do these regulations have on poor people? Since these are municipal regulations, not national ones, they can't change how well off poor people are in the long run. If Chicago has worse regulations for poor people than Milwaukee, enough poor people (not all) will move from Chicago to Milwaukee that poor people

in Chicago will be no worse off than those in Milwaukee. Ill-conceived regulations drive up wages, drive down land prices, and keep cities from realizing all of the economies of scale they should. Once again we must ask whether the regulations are potential Pareto improvements.

Even in the short run, and even if we are concerned only with poor people, it is hard to escape this question. Who loses if the externalities these regulations are supposed to counter go unchecked? Mainly, poor people lose. Poor people are at greatest risk of contracting communicable diseases from other poor people (or of being stereotyped as probable disease carriers), at greatest risk of seeing their houses catch fire because a fire spreads, at greatest risk of suffering from a delayed response because another fire is preoccupying the fire department, at greatest risk of being hurt by falling debris or of seeing their homes lose value or their addresses stereotyped by employers when other houses deteriorate. Poor kids lose when they eat lead paint or they can't sleep because there is no heat. If bad housing causes violent crime, poor people are much more likely to be its victims than rich people.

Thus, the questions we asked in Chapter 8 about regulations are still the relevant ones: Do they create potential Pareto improvements, and can better ways be found to do the same things?

Rent control is the other type of regulation often cited as a way of helping poor people. Rent control can help certain poor tenants at certain times by reducing their rent and allowing them to stay in apartments they couldn't otherwise afford. These gains come directly at the expense of particular landlords.

Suppose you identified these particular tenants and particular landlords and figured out how much each tenant gained from each landlord, then you abolished rent control but forced each landlord to pay each tenant the amount of the rent control transfer. No one would be worse off under this program, and some people would be better off because apartments would be better allocated and better maintained, as we saw in Chapter 13. The new program would be a Pareto improvement over rent control.

It would also be rightly criticized as unfair. Why should these particular poor people get money and not others? Why should these particular landlords bear the burden of alleviating poverty and not other wealthy people? But this arbitrary transfer program (which no one would think of as a good way of helping poor people) is Pareto superior to rent control. So rent control is probably not a good way of helping poor people.

II. Public Housing

Another obvious reaction to the plight of people living in deplorable housing is to build good housing for them, the same way governments build water systems and sewers. This turns out to be harder than it sounds. Sometimes the

government builds and operates the housing itself, and sometimes it subsidizes other private groups to do so. Housing that the government owns and operates itself is called public housing.

Public housing has never been a big program in the United States and it probably never will be. About 1 percent of the U.S. housing stock is public housing now, on the most expansive definition; stricter definitions that coincide more closely with the picture most people have of public housing would come close to cutting that figure in half. Public housing was first constructed in the United States during the Depression, and the government stopped building it in the 1960s. Since the federal government currently spends about $3 billion a year in operating subsidies and an equal amount for renovation (Quigley 2000, p. 60), the tax expenditure on garages is almost twice as large as the federal budget expenditure on public housing.

In Europe and some Asian countries, public housing is much more prevalent than it is in the United States. Almost half the population of Hong Kong, for instance, lives in public housing. These countries operate their public housing quite differently.

In the United States, public housing in each city is run by an independent corporate entity called a local housing authority (LHA). Only low-income people are accepted as residents, and they have to pay 30 percent of their income as rent. Rent receipts aren't enough to cover either the capital or the operating costs of these projects, and so LHAs receive a variety of subsidies from different levels of government.

A. What's Good about It

Although U.S. public housing has been unpopular as a policy for many years, it has accomplished several important goals. First, many poor people have had better housing and better lives than they otherwise would have. Housing authorities provide something poor people want, and even the most troubled LHAs have long waiting lists. Public housing sometimes makes adults into better parents, and often improves its residents' health.

These conclusions arise from a variety of studies. Wilner and colleagues (1962) suggest that moving into new housing projects in the 1950s improved residents' health. Currie and Yelowitz (2000) find that in the early 1990s, everything else being equal, families that lived in pubic housing projects were less likely to suffer from overcrowding. Children who lived in housing projects also did better in school than otherwise similar nonproject children; boys especially were less likely to have been held back in school. Newman and Harkness (2002) look at a longer time frame—they ask whether teenagers who lived in housing projects in the 1970s were more successful in their midtwenties than otherwise similar kids who did not. Their answer is a mild

yes: growing up in a project made them more likely to avoid welfare dependency and possibly more likely to earn a good wage and stay out of poverty.

(Notice that these studies ask and try to answer a difficult question, but the right question. No study maintains that children from projects do better in school or make more money than children who are not from projects. They don't. But the reason they don't is that they come from families that are poor and have many problems. The question is whether children from poor, multiproblem families who live in housing projects do better than children from poor, multiproblem families who don't, and the answer is yes.)

Like tax preferences for homeowners, public housing probably promotes stability and good citizenship. The theoretical argument that public housing promotes these attributes is the same as the argument for rent control: because public housing residents have a good deal (witness the long waiting lists) that they would lose if they left, they should be less likely to move, and more likely to take a hand in improving their community (since a better community won't increase their rent). Public housing tenants are in fact quite stable: only 11 percent moved in 1997, as compared with 33 percent of tenants generally and around 20 percent of households with income below $15,000 (National Association of Housing and Renewal Officials 2000; U.S. Bureau of the Census 1999, tables 30, 31). Less is known about civic activity.

Thus public housing is like homeowners' tax preferences and rent control: it addresses some real problems and real distortions in the housing market. Like rent control, it helps some poor people.

B. Problems

But like homeowners' tax preferences and rent control, public housing also has problems. I will look at three major areas of concern: location, quality, and fairness.

1. Location

If the government is going to build apartments for poor people, it must build them *somewhere*. It's hard to see how governments can get the location question right, and easy to see how they can get it wrong.

In a perfect land market, a profit-maximizing apartment operator commissioned to build a building for poor people will choose the right location: he will get the benefits of a better location in the form of higher rents his tenants will be willing to pay and lower construction expenses, and will bear the costs of the inconvenience his location imposes on other people by paying the price of land. These incentives to build in the right places disappear when the rents that an LHA collects have little or nothing to do with what tenants are willing to pay, as in U.S. public housing.

Recall, moreover, from Chapter 5, that decisions in which government officials have strong private concerns different from public well-being are also best taken out of their hands. The government officials making decisions about where to locate public housing have few incentives to put it in the right places and many incentives to locate it in the wrong places. This is because American elections are place-based. I may be much better off and much more productive living in New Zealand, but if I am a reliable Democratic Party voter, Democratic-elected officials gain by paying me to stay in New Jersey.

The actual locations of U.S. public housing have often been criticized, and are probably worse than the simple theoretical model implies. Public housing is disproportionately located in older cities in the Northeast and Midwest (and Puerto Rico). More than 12 percent is in New York City, another 7 percent of public housing apartments are in Puerto Rico and Chicago. In 1997 (after over a decade of imploding projects), Newark (population 273,000) had just about as many public housing units as Los Angeles (population 3.7 million). Within metropolitan areas, public housing is concentrated in the poorest, most heavily minority neighborhoods. There's reason to believe that these are the wrong locations.

Why? The intrametropolitan issue was discussed at length in Chapter 11. Locating projects in minority neighborhoods is costly in terms of commuting (the projects aren't near jobs their residents can hold), child rearing, and possibly employment and business development. Public housing is a major reason why minorities live in the wrong places.

On the intermetropolitan question, since public housing residents are being subsidized to live in, say, Chicago rather than Sacramento, and some would move from Chicago to Sacramento if they could take the subsidy with them, the losses to the losing areas exceed the gains to the winning areas. Public housing subsidies tied to living in Chicago distort intermetropolitan location. Public housing keeps people in the wrong places. They would be more productive in Sacramento, but they stay in Chicago because of the subsidy.

Why were public housing projects located so perversely? Probably because public housing was originally presented in the Depression as a program for slum clearance, not as a program for improving the lives of poor people. So the original sites for public housing were chosen because slums were located there in the 1930s, not because they would be good places for poor people to live in the twenty-first century.

The legal status of slum clearance led to the second reason for the perverse location of public housing. A Supreme Court decision in 1935 had held that public housing, or slum clearance, was not a purpose for which the U.S. government could exercise its eminent domain powers, and so direct federal operation of the program would have been difficult. The noxious effects of slums were obvious enough to the public at this time that states and cities could plausibly invoke their police powers for getting rid of them, and in time

most state courts ratified this interpretation. Police powers could not have been used if the programs were for bettering the lives of poor people. Thus, even though the federal government approved projects and funded them, only localities could initiate them. Housing projects were therefore built where local governments wanted them.

To see the consequences of local control, think about who stood to gain from a housing project, at least in the short run. One prime group was the new residents, who could reward their benefactors only by voting. Thus officials who represented substantial numbers of poor constituents wanted housing projects, and they wanted them built in their districts. Officials who represented wealthier areas, or undeveloped areas, did not want projects: they saw no reason to replace the voters who had had the wisdom to elect them with a group of unknown poor people.

Another group who could gain were landowners looking to unload their property, and hoping to get a higher price from the LHA than the market would provide. Slum clearance for them was an attractive "golden handshake," especially if the value of their property was declining because of suburbanization. The LHAs, since they were spending federal money, not their own, did not acquire a reputation for driving hard bargains when they acquired land, especially when they bought it from their friends.

This pressure from interested landowners worked in much the same direction as pressure from would-be tenants: it concentrated housing projects in the poorer neighborhoods of older cities. By the 1940s, Newark, for example, had several housing projects, each identified with a different ethnic group and political faction, but most New Jersey suburbs had none, and most Sun Belt cities had few.

Race added to this location problem, especially after World War II. As whites prospered and blacks moved to cities, the proportion of the poor and hence eligible population that was black grew. Because African Americans at this time were paying more for housing than whites and getting less, public housing was especially attractive to them, and so black politicians sought housing projects for their constituents. Where would these projects be built? In black neighborhoods, obviously. White politicians didn't want them in white neighborhoods—they wanted neither to outrage nor to replace the voters who had elected them—and neither did black politicians, who wanted the people they helped to be able to vote for them, and were not inclined to engage in fierce and prolonged battles with their white colleagues that would result in their losing those voters.

2. Quality

A government that builds housing for poor people must decide exactly what housing to build, as well as where to build it. The way the public housing program was set up, it was almost impossible to find a good answer to what hous-

ing should be built. The public housing program spent a lot more building housing than that housing was worth to the people who lived in it.

Think about the type of housing you would build in a program for slum clearance. You can't build absolutely terrible housing, because you're supposed to be getting rid of bad housing. You can't build truly great housing either, because you can't expect middle-class taxpayers to pay happily for housing for a select group of poor people that's nicer than what they're living in.

The natural response is to find some compromise type of housing—to pick some niche fairly low in the income distribution but somewhat above poverty and imitate the type of housing that people at that income level live in. But this natural compromise won't work, because there's nothing for you to imitate. Nobody at the income level you've identified lives in newly constructed housing, and the only kind of housing you can build is newly constructed.

Why don't lower-middle-class Americans live in new housing? I will address this question in more detail in section V, but the basic idea is that people who aren't well off buy used housing for the same reason they buy used cars: it's cheaper. Because poor and lower-middle-income people can get housing of the quality they can afford more cheaply by buying it used, the market doesn't build any housing like this (the same way the market doesn't produce any new water; it just recycles the water that's been here for aeons). If somebody built new housing for poor and lower-middle-income people, they wouldn't buy it, because they could get the same quality for less in the used market.

(The fact that they can get their housing more cheaply in the used market than in the new market makes life better for poor people, and so we should be happy that housing is so durable. We should not lament the fact that some evil force prevents construction of new housing for poor people, any more than we should lament the fact that our clothing lasts long enough that we don't wear new clothing every day. If houses were like toilet paper and had to be thrown out after every use, poor people would live in new housing, and everyone would be worse off.)

Thus public housing had to be of a quality never supplied by the market, and totally different from anything anyone had ever seen or built before (and something for which no one could observe what private developers were paying).

Why couldn't LHAs just rent existing apartments of the right compromise quality for the people displaced by slum clearance? Because in the Depression public housing was touted as an employment program, not just a way of reducing the noxious effects of slums. Moreover, finding existing apartments to rent would have been tough, because slums were being torn down.

So public housing began with serious problems concerning the quality of

the housing it was building. These problems were compounded in the 1950s when a new goal was added—urban renewal.

In the 1940s many large downtown property owners became concerned that the trend toward reduced density in cities and decentralization was working against them—their holdings were becoming less valuable as their workers and customers were moving further away. In many cities, moreover, they saw the Great Migration as exacerbating this trend, threatening to turn the great downtowns of the 1920s into stagnant white islands, cut off from their natural constituency by a vast sea of black poor people. (We saw in Chapter 11 that this perception may have been correct.) The newspapers, utilities, office building companies, insurance companies, hospitals, and universities that found themselves trapped in decaying downtowns were a formidable political force, and urban renewal was their first attempt to fight back.

Urban renewal programs were supposed to use federal dollars (with soft local matching funds) to buy up blighted properties, knock them down, and rebuild on them the vibrant, beautiful, useful buildings that would make cities thrive again and draw the middle and upper classes back downtown to shop, work, and live. Large federal subsidies would be needed at first to gather a critical mass, but once that critical mass was attained, increasing returns to scale would take over and downtown property values would be restored. Given enough federal dollars, urban renewal would undo the harm done by the automobile, the Great Migration, and frozen food.

At the same time, downtown interests felt they also needed better highway access. They knew that any downtown without good highways was going to be reduced to a backwater. In the late 1950s, the federal government was paying a large share of the costs of the interstate highway system. Highways to downtowns had to go through blighted areas. Building highways was another reason for tearing those neighborhoods down.

What downtown real estate interests (and most urban intellectuals) called blight, however, a lot of other people called home. The owners of blighted properties were no problem—like the earlier owners of land taken for public housing, many of them found the generous price subsidized by the federal government an excellent way of ending their association with a particular neighborhood. In principle, the people who lived in these neighborhoods could have been paid off in cash, too, and made no worse off by the revitalization of their city. Or in a more repressive regime, they could have been simply booted out into the street, shipped back to the South or to Puerto Rico, or resettled on reservations (like Soweto) somewhere in the far suburbs.

In the United States of the 1950s and 1960s, however, politics ruled out all of these possibilities. The only feasible alternative was to build housing projects. Housing projects built simultaneously with urban renewal projects would sop up a lot of the displaced families without threatening white neigh-

borhoods or black politicians. The trick, though, was to free up land for urban renewal or highways at the same time that land was freed up for new housing—without too large an outflow into white neighborhoods. The only way to do that was to build at higher density: you could get 50 acres for urban renewal by tearing down 100 acres and building enough housing for the people who used to live on those 100 acres on 50 acres. That meant high-rises.

Since most of public housing before urban renewal had been low-rise (three-story buildings being predominant), new high-rise construction for poor people introduced still another new form of housing never before seen in the market. Before this, high-rise apartments had only been for middle-class and rich people—they traded off low privacy and the expense of elevators and reinforced structural supports for great access, great locations, great views, or amenities like doormen. Architects of public housing high-rises had to include all the bad features of high-rise living (including severely restricted elevator access) and leave out all the good, as they tried to invent a new form of housing.

The result of this bizarre history is an architecture almost universally reviled. In many cities the high-rises built in this period have since been depopulated and blown up. The Pruitt-Igoe housing project in St. Louis was the first to go, in 1976, but high-rise projects in Newark, New Orleans, Boston, Chicago, and several other cities have also been knocked down. During the 1990s the federal government spent billions of dollars on a housing program, called HOPE VI (Housing Opportunities for People Everywhere), designed to replace existing public housing units. In many cases, a lot of the units demolished have been replaced by newly constructed low-rises and townhouses, but the trend in the number of occupied public housing apartments has been downward. Generally, the buildings that have been demolished are the newest, not the oldest. The projects built when public housing was an employment-generation scheme have mainly survived, but many of those built when public housing was an urban rejuvenation project have not.

Even before cities began to knock the projects down, economists found that they cost much more to build and operate than they were worth to the people who lived in them. The residents would be a lot better off if, instead of spending the money on public housing, the government gave it to them directly. Mayo and colleagues (1980), for instance, estimated that recipients would be just as well off if they got, in cash, around 50 cents for every dollar the government spent on public housing. Several other studies done at about the same time reached this conclusion for an amount ranging from 70 to 80 cents per dollar (for a review, see Olsen 2001). This deadweight loss comes from several different sources: because some recipients are forced to consume a higher quality of housing than they would choose on their own; because LHAs do not spend their money wisely; and because the quality of housing that LHAs provide could be supplied more cheaply through downgrading

than through new construction. The last reason is probably the biggest source of loss.

3. Fairness

Public housing has never been an entitlement program like social security, FHA loan insurance, homeowners' tax preferences, or fire protection. In an **entitlement program,** everyone who qualifies gets the service. For public housing, only fairly poor people are eligible—the exact rules are complex, vary by locality, and have changed over time—but only a few of these people are admitted. No matter how poor you are, for instance, if you live in a city that doesn't have any public housing you can't get into public housing.

Congress could have made the program big enough to accommodate everyone eligible, but that would have been very expensive. Or it could have restricted the category of those eligible to fit the size of the program, but that would have necessitated hard choices, limited the discretion and power of LHAs, and reduced the number of people who thought they might benefit. Or it could have reduced the amount it spent per household. But it rejected all of these options.

Since the subsidy that public housing residents get is fairly big—witness the long waiting lists—you have to wonder how fair it is to single out a fraction of a large group of equally deserving people for special treatment. This inequity also makes it important that we ask how public housing affects people who qualify but don't get in. Does public housing indirectly help poor people who live in private housing, or do housing markets compound the unfairness by making life tougher for the unlucky poor people who don't get in? Answering these questions requires some careful analysis, which I will postpone until section V.

In summary, then, the U.S. experience with public housing has left much to be desired (though the popular view of housing projects as hell is probably considerably overstated). Some of the problems arose from particular policy decisions—local control and new construction, for instance. Others may be endemic to the idea of public housing.

III. Other Supply-Side Programs

By the late 1960s, policy makers were looking for alternatives to traditional public housing. Because public housing was perceived as a failure, they wanted to reverse some of the policy decisions that had gone before, but not all of them. The new programs established at this time sought to involve private rather than public actors in various aspects of housing operations. Slum clearance also disappeared from the agenda. The consensus at the time was that the problems with housing projects were caused mainly by how LHAs operated

them or how architects built them, not by anything deeper. Mainly the new programs did not tinker with policy decisions regarding new construction, the fact that more people were eligible than would be served, local influence, and high-but-not-too-high building quality.

A plethora of new programs were initiated, each named after the section of legislation that authorized it, so that discussions of housing started to sound like prognostications about a state lottery. The two biggest supply-side programs that were developed at this time were Section 236, which provided 448,000 new units of housing in 1997, and Section 8 "New and Rehab" (as distinguished from Section 8 "Existing"), which provided 895,000. Together these programs have created slightly more units than public housing has, but fewer people live in them because their proportion of elderly residents is higher. Section 236 stopped making new commitments in 1974, but the years in which it flourished—the Nixon administration—saw the greatest expansion in federal housing commitments in U.S. history.

Under both of these programs, a private entity acquires land (generally not through eminent domain), borrows money, and builds a building. The private entity is either a nonprofit corporation or a limited-dividend corporation (a corporation constrained to pay annual dividends of no more than 6 percent of shareholders' investment). The government subsidizes the mortgage payments, and often the limited-dividend sponsors get transferable tax advantages: sharply accelerated depreciation and tax credits. The units may be rented only to low-income tenants, and tenants can pay no more than 30 percent of their income as rent—the same rent they would pay in public housing. The government establishes a "fair market rent"—what it thinks an apartment of that quality should cost in that area—and pays the sponsor the difference between that amount and what the tenant pays.

Since sponsors are almost guaranteed a profit, can't make a big profit, don't have to be elected by tenants, and can't collect higher rents if they provide better services, it seems at first that their incentives would be just as bad as those of LHAs, and to some extent this is true. Sponsors have strong incentives to award construction and operation contracts to related companies and close friends, to reduce quality, and to channel the resulting funds back to themselves, directly or indirectly. These programs have been far from scandal-free (that was why Section 236 stopped new construction in 1974), and in recent years some buildings built under these programs have been abandoned and torn down, too, just like the high-rise public housing projects.

But sponsors are not monopolies like LHAs. They have to compete with LHAs and against other sponsors for eligible tenants, especially the best eligible tenants. This competition need not be particularly fierce, since the number of eligible tenants in most cities far exceeds the number of available subsidies, but it places a check on the activities of both LHAs and sponsors. A truly hor-

rible building will draw only the most undesirable of tenants. (This introduction of competition with newer and hence more desirable buildings may have hastened the demise of some housing projects in the 1970s and 1980s and induced LHAs to try to replace them with their own new construction.)

Nonprofit sponsors also are often highly diversified, unlike LHAs, which are usually constrained to do nothing but run housing projects (although sometimes they have run urban renewal programs as well). Many nonprofit housing sponsors started as churches, and some have grown into multipurpose community development corporations that offer an array of services: job training, recreation, social services, child care, supermarkets, and so on. Often these other arenas are highly competitive, and reputation matters a lot. Worshipers can switch to another church; philanthropists can look to other charities; governments can fund other providers. A reputation for low-quality housing would hurt housing sponsors severely in their other endeavors, and so sponsors have incentives beyond those of an LHA to site their buildings in good locations and to maintain them well.

The final large supply-side program dates from 1986 and is called the Low Income Housing Tax Credit (LIHTC). About a million units had been constructed under this program by the turn of the century, and the number is expected to keep growing. Each state gets a fixed dollar amount of credits to distribute to new projects each year; the amount depends on the state's population. The states select the projects and developers to receive the tax credits. The projects are financed with tax-exempt bonds, and the holders of these bonds get tax credits equal to 70 percent of the cost of building the projects. The typical project built with LIHTC assistance receives subsidies from other sources as well; the packages are usually quite complex. For instance, about 40 percent of LIHTC units receive Section 8 assistance (Olsen 2001, p. 28). Only tenants with household income below about 60 percent of the local median can live in LIHTC units, and the maximum rent an owner can charge is 30 percent of this upper income limit. Rents within a project don't vary with income. Because of this, only 28 percent of households in LIHTC projects are "very low income" as the Department of Housing and Urban Development defines this term—as opposed to around 80 percent in other supply-side programs (Wallace 1995). States usually have many more developers eager to build LIHTC projects than they have credits available—evidence, as Olsen (2001) points out, that the profits this program offers are larger than they have to be.

Some recent supply-side programs have also concentrated on neighborhood revitalization. Here the purpose of the subsidies has been to increase the value of nearby properties—to make poor people better neighbors. New York City's large local supply-side program in the 1980s and 1990s apparently accomplished this objective (Ellen et al. 2003), but results in other cities have

not been so positive. LIHTC projects do appear to increase property values in the surrounding neighborhoods, but other federal housing programs for families probably have the opposite effect (ibid.).

IV. Demand-Side Programs

The basic ideas behind demand-side programs are that poor people can find adequate housing in buildings that have not been constructed explicitly for them, and that profit-maximizing landlords can provide this housing if they have the right incentives to do so. Supply-side programs give poor people adequate housing; demand-side programs give them the means to acquire it.

The major U.S. demand-side program for poor people is usually referred to as the Section 8 Existing Housing Program (which I'll call simply as "Section 8"). It began in 1974. Section 8 now serves more households than any other housing program for poor people, though not more than all the supply-side programs combined. (The current official name is the Housing Choice Voucher Program.)

In many ways Section 8 is a successor to public housing, and it has continued many of the policy decisions on which public housing rested. More people are eligible for Section 8 housing than Congress authorizes money for; LHAs run the program and select the lucky tenants. Housing has to meet quality standards dictated by the federal government, but it can't be all that much better than the housing many voters consume.

The rent and selection formula, too, for most Section 8 tenants, is a pure carryover from public housing. The formula depends on "fair market rent" (FMR), the same amount that sponsors of supply-side buildings get. Households have to find their own apartments. They pay 30 percent of their income in rent, and the government makes up the difference between that amount and the FMR. If a household lives in an apartment that rents for less than the FMR, the government pays less; if the apartment rents for more, the household is responsible for the amount above the FMR. Section 8 Existing is therefore just like Section 236 or Section 8 New and Rehab, but without the capital costs. Thus it's somewhat cheaper per household from the point of view of the government's budget.

LHAs have fixed numbers of certificates to distribute, and they give them to applicants on their waiting lists. If an applicant receives a Section 8 certificate, she has several months to find an apartment that qualifies: it has to meet the program's standards, and the landlord has to be willing to rent it at the predetermined FMR, file the necessary paperwork, and submit to periodic inspections. The tenant can also stay in her current unit if it qualifies. Often, though, the time elapses before the certificate holder can find a qualifying apartment that she is willing to move to—moving isn't free, people have all

sorts of reasons for wanting to live in particular locations, and sometimes no landlords are willing to rent a qualifying apartment at the FMR. When that happens (and it happens about a third of the time), the tenant forfeits her Section 8.

Section 8 has some clear advantages over supply-side programs. Being cheaper is a big plus; producing new housing when old housing is adequate for the job is wasteful. Tenants can choose their location under Section 8, subject to real constraints of apartment availability; they can receive a housing subsidy without being bound by LHA location decisions. Tenants can put pressure on landlords by moving out or never moving in, and landlords can profit if they figure out a way to provide adequate housing that both tenants and inspectors like, at less than the FMR (although they also have a stronger incentive to try to fool or confound inspectors, since they can then keep their savings directly, unlike LHAs, nonprofits, and limited-dividend corporations).

Despite these advantages, Section 8 has a number of problems. One of the goals that some analysts had hoped to further with Section 8 was **horizontal equity**—the equal treatment of equals. Although a housing allowance program could in theory provide horizontal equity, the Section 8 program doesn't. The number of vouchers is far smaller than the number of households that qualify, and so many people just as deserving as Section 8 recipients don't participate.

It's not clear, either, that Section 8 does a good job on the central tasks of making poor people better citizens, parents, and neighbors. So far as citizenship goes, Section 8 tenants may have a stake in the community because to some extent the rents they pay don't change if the neighborhood deteriorates or improves. But a significant improvement in neighborhood desirability could raise market rent for their apartment above the FMR and force them either to move or to pay more from their own pocket. Section 8 tenants are also not so stable as public housing residents: vouchers are mobile, although moving them may be challenging. I'm not aware of any evidence on whether Section 8 tenants are better citizens than public housing residents or unsubsidized renters.

Whether Section 8 is good for kids is also an open question. Experiments, natural and controlled, that have transferred randomly chosen families from public housing to Section 8 housing haven't seemed to make a difference in kids' lives. For instance, children's academic achievement was unaffected when families were "vouchered out" of troubled high-rise projects in Chicago (Jacob 2004).

The most extensive evidence comes from the Moving to Opportunity (MTO) experiment that I discussed in Chapter 12. One experimental group in the MTO program received regular Section 8 vouchers, and outcomes for those families were compared with the outcomes for families who received no special assistance but were allowed to remain in public housing.

In the families given Section 8 vouchers, girls were less likely to be arrested and to experience behavior problems, but only the arrest difference was statistically significant. Girls' mental health also improved. Boys in Section 8 housing had more behavior problems and more arrests than public-housing boys, but the difference wasn't statistically significant. There were no statistically significant differences in terms of academic achievement, high school graduation, drug use, or pregnancy. Generally, girls in the families with Section 8 vouchers seemed to behave better (relative to middle-class norms) and achieve more, while boys behaved worse and achieved less, but the effects were small (Kling, Ludwig, and Katz 2004; Kling and Liebman 2004). Thus Section 8 seems to be no better than public housing for kids, but no worse either. Of course, there is a little evidence that public housing is better for kids than private unsubsidized housing, but that evidence is also weak and may be out of date.

As for making people better neighbors, it's not known how Section 8 affects communicable diseases or property values. Selection, however, may work against big positive effects, because many of those households whose housing is worst don't participate in the program, and those who do participate use a great deal of the money for other purposes. In housing allowance experiments that the federal government conducted in the 1970s with programs that looked like Section 8, the households living in the worst housing were least likely to participate in the experiment; disproportionately, but not exclusively, these were the poorest households. Many of the households that decided to participate already lived in housing that met the standards, and most of those that did not lived in housing that fell short of the standards in ways that a few cheap repairs could correct. The average cash outlay for repairs was only $60 (Allen, Fitts, and Glatt 1981, p. 29).

While the Section 8 program as it operates today is not the same as any of the experiments from the 1970s, there is not much evidence to contradict this conclusion, and there is some to support it. Crews (1996) found that people who participated in subsidized housing programs—all kinds, not just Section 8—on average had a much greater desire for housing, relative to other goods, than families who didn't participate.

A final reason for low-income housing programs is to assure that everyone has certain minimum capabilities, whether they want them or not—being well-nourished, for instance, sheltered, free of avoidable diseases, able to take part in society without shame. Some elements of the housing bundle are crucial to several of these important capabilities. But Section 8, because it misses the people in the worst housing, doesn't do much to assure minimum capabilities, even if the program were available to everyone who sought it out. I will discuss these issues in greater depth in Chapter 15, which is about homelessness.

We can see that Section 8 seems to lack many of the traits that a good

low-income housing program should have. But Section 8 is not the only form a demand-side program can take; the failures of Section 8 are no more endemic to demand-side programs that the failures of U.S. public housing are endemic to supply-side programs. European housing-allowance programs, for instance, are vastly different from Section 8, as well as from each other (see Howenstine 1986).

One alternative might be to pay fixed sums for physical characteristics that have value for society at large. This would be a simple system of pure Pigouvian subsidies: ten cents a month, say, for every square inch of window space, or for every square foot of floor space in a household with children; ten dollars if there's indoor plumbing. Finding out the characteristics of a large number of apartments would be difficult, but no more difficult than checking to make sure that a long list of standards is met and that a correct amount of rent is charged—the information Section 8 requires. Perhaps this information could be collected when a housing unit is built and then spot-checked only occasionally thereafter. Bonuses could also be paid for stability.

Figuring out what should be paid for and how much should be paid would be extremely hard. But the U.S. government spends close to $30 billion a year on housing programs for poor people and well over $100 billion (including tax expenditures) on housing programs for everyone else. Surely an effort to decide precisely what it is that these programs are supposed to be producing and to measure it would be a reasonable objective.

Pigouvian subsidies should not be tied to income; an outdoor privy owned by a millionaire is just as noisome as one owned by a pauper. For upper income groups, Pigouvian subsidies would be a well-designed improvement on homeowners' tax preferences. For lower-income groups, they would reward small improvements in different relevant aspects of a house, even if those small improvements together didn't bring the whole house up to some arbitrarily high government standard.

This system would leave income distribution to be determined by taxes and an income-maintenance system. Horizontal equity would improve—those who were receiving greater housing payments would be those who were providing the rest of society or their children with greater benefits—and so would work incentives. Households would no longer have to contend with two separate tax rates on their earnings, one imposed by the housing authority and the other by the Internal Revenue Service. There are reasons to want to improve housing, and there are reasons to redistribute income, and they are not the same.

Olsen (1981) proposed a less radical reform: a universal, small subsidy of a fixed amount, no matter how much rent a household pays, plus some fraction of rent, up to a maximum amount. There would be no housing standards that would have to be met (other than whatever local codes were normally enforced). Lower income households would get greater amounts. This program

would probably create greater horizontal equity than Section 8 and do a better job of reaching those most in need. Even the poorest and worst housed people would be subsidized for improving the quality of their housing, and so they might do so more than they do under Section 8.

Whether the added rent would go to the right aspects of housing—those that affect kids and outsiders—is a more difficult question. The system could even be susceptible to collusion between tenants and landlords—landlords could raise the rent, provide no better services and quality, but kick back enough of the cash to the tenant to make both richer. At the same time, the Section 8 standards could be written to force recipients to concentrate on those relevant aspects of the housing bundle, although it is by no means clear that they do so now.

V. What Happens to Nonparticipants?

Despite all the programs I've talked about in this chapter, most poor and near-poor people live in private, market-provided housing. Therefore we need to figure out how these programs affect those who don't live in publicly funded apartments or get rent-subsidy certificates. Programs that cut the rents of nonparticipants would look a lot fairer than programs that raised those rents. Moreover, we saw in Chapter 13 that homeowners' tax preferences were by far the largest federal interventions in the housing market. We ought to explore whether they affect poor people, even though few poor people benefit from them directly.

Both questions are really about how changes in demand and supply for one type of housing affect the market for some other type of housing. Such interactions are complex and hard to decipher, but we need to think about them. First I will outline how economists think about housing and markets when they want to answer questions like these, and then we will apply this method to homeowners' tax preferences, growth controls, Section 8, and public housing.

A. Housing Qualities and How They Fit Together

The simplification that many economists use when they want to think about how the markets for different kinds of housing fit together is to think of houses as differing along only one dimension: some are "better," some are "worse," and everyone agrees what's better than what. This simplification is based in some truth—almost everyone thinks mansions are better than shacks, and houses with great plumbing usually have above-average electrical systems, too. But it clearly doesn't represent reality perfectly—some people love lawns and others hate them; some crave sunlight and others want security. That's why I call it a simplification, not a truth.

At any moment a city contains an array of housing, some of very high quality, some of very low, and much is in between. There are demands for the different qualities of housing, and there are supplies.

1. Demand

On the demand side, each household wants to have only one house, to take advantage of scale economies and to save on transportation. For now, I want to work with a closed-city model. Each household decides what quality of house it wants to live in, based on its preferences and the prices of the different qualities. How many households want to live in houses of a certain quality depends not only on the price of that particular quality but also on the prices of all the other qualities of housing. When the price of a particular quality goes up, demand for that quality goes down and demand for every other quality goes up. The effect is much greater on qualities immediately above or below the quality in question than on houses of very different qualities. Houses of similar quality are good substitutes for each other; houses of very different qualities are much poorer substitutes. If prices at Trump Tower rise, not many people are going to decide to spend the night in a flophouse as a result.

If we think of a short period of time in which the supply of each quality of housing is fixed, then prices in that period will adjust to make demand equal supply for each quality of housing: the number of households who want to buy houses of each quality has to be the same as the number of houses of that quality available. This is called a short-run equilibrium. In a short-run equilibrium, it's easy to see that prices of higher qualities of housing have to be higher than prices of lower qualities (nobody would want to buy a type of housing that was both worse and more expensive than another type).

Richer people will usually end up buying higher qualities of housing. Richer people buy more of almost every good thing, and housing quality (especially as we've defined it) is a good thing. Not every richer household will have better housing than every poorer household—some people care about housing an awful lot, while others like spending their money on other kinds of goods—but the general tendency is for greater wealth to be correlated with better housing.

What happens to a short-run equilibrium (and a closed city) if the supply of one particular quality of housing increases? First, the price of that quality would have to go down, because to make supply equal demand more households have to be coaxed into choosing that quality. Where will these households come from? From the closest qualities above and below the quality for which supply increased. But then supply will exceed demand at those immediately adjacent qualities, and price will have to fall for them, too. These secondary price decreases will intensify the need for a lower price at the original quality, and the effect will spread like ripples scooting across a pond. Lower prices at a quality a little distance from the original quality will draw

households from qualities a little bit further away; then prices will have to fall at those qualities a little bit further away to bring demand back up to supply; then demand will fall again a little further away than that. The cycle will repeat itself but get weaker every time. Thus, more supply of one quality will cause lower prices for all qualities, but the effect will be smaller the more distant a quality is from the original disturbance. Figure 14.1 illustrates.

This demand-side analysis gives us a picture of what happens over a short time period, say a few months. To get a longer-term view, we have to consider the supply side and then integrate supply and demand.

2. Supply

You can produce a house of a particular quality at a particular time in several different ways. Obviously, you can build a new house of that quality. The cost is the cost of labor, materials, and land. Or you can take a house of lower quality and renovate it—put in new air-conditioning, tear out a few walls, and rebuild the kitchen with the latest appliances. In this case the cost is labor and materials, just like building from scratch, plus the loss of a house at the lower quality. The same physical facts that make housing a bundle rather than a series of disconnected purchases often will make building from scratch cheaper than renovation: installing new wires in a house that's being built is a lot easier, for instance, than trying to snake them through crevices that weren't designed to hold them. When the stock market rose and demand for high-quality houses surged in the New York area, "knock-downs" became common: people bought perfectly fine houses and, instead of upgrading, knocked them down and rebuilt from the ground up.

Figure 14.1 The effect of more supply of a particular quality of housing on short-run equilibrium.

But in any year, the number of housing units supplied by building and renovating during that year is small—generally between 1 percent and 2 percent of the total units available (this is hard to measure because good records are not kept on renovations). Two other processes and decisions supply the other 98 percent of houses: maintenance and downgrading.

Maintenance is easy to understand. If you have a house of a particular quality now, and you want to have that same quality a year from now, you have to maintain it, and that's costly. The furnace may break, the leaves have to be removed from the gutters, the driveway has to be resealed, the sink needs to be fixed, new tenants have to be sought and screened, the windows have to be washed, and periodically the whole exterior needs a new coat of paint. The list of basic repair tasks is endless; this is almost guaranteed by the second law of thermodynamics. One way to respond to the challenges of natural deterioration is to accept them and defeat them. I call this a policy of maintenance. If you do this, the house will continue at the same quality.

But you don't have to accept every challenge that comes your way. Sometimes letting nature take its course and letting the house deteriorate is the best policy. (Most of us who are over eighteen are becoming more decrepit every day; why should our houses do better?) If you don't maintain a house this year, it will be a lower-quality house next year. Downgrading in this sense is a passive option, the opposite of maintenance.

Downgrading, however, can be more active, too. The best example of active downgrading is subdividing a large house into several smaller apartments. Nor is downgrading an all-or-nothing proposition—you can spend some money on maintenance, but not enough to maintain the house at the same quality. Downgrading is costly in another sense, though: what you lose when you downgrade is the opportunity to collect a higher price at the old quality.

Every year, then, owners of houses and of buildable lots decide what quality to supply next period (or to supply no housing at all). They can build, renovate, maintain, or downgrade. They do what maximizes their long-run profit. If the price of a certain quality of house is high relative to the cost of construction, more houses of that quality will be built; if the price is high relative to the price of slightly lower-quality housing, houses of that quality will be maintained; if the price is high relative to the price of slightly higher-quality housing, houses of the higher quality will be downgraded.

Because building from scratch is usually cheaper than renovating, and because downgrading is also cheap, in many cases the cheapest way to provide moderate or low-quality housing is to build high-quality housing and let it deteriorate. Suppose you want to produce one moderate-quality house five years from now, and the rate of interest is 5 percent. Suppose high-quality houses cost $200,000 each to build and moderate-quality houses cost $150,000. So, to build the moderate-quality house in five years you have to set aside $117,529 today; at an interest rate of 5 percent a year, this will grow to $150,000 in five years. On the other hand, if you spend $200,000 today, you

will have a moderate-quality house in five years, and in the meantime you will have the opportunity to collect rents (or live in the house). If the present value of five years of rent (or five years of living in the house) is more than $82,471 (= 200,000 − 117,529), then it's cheaper to build the high-quality house and let it deteriorate than to build the moderate-quality house. You won't see any construction of moderate-quality houses. On the other hand, if the present value of five years of rent is less than $82,471, houses of both qualities will be built.

In general, if the difference in cost of construction between two qualities is fairly small, but the amount of rent you can collect as the house deteriorates from the higher to the lower quality is fairly large, then only the higher-quality house will be built. It will be cheaper to supply the lower-quality house through downgrading. In the extreme, if every quality of house cost the same to build—building a shack cost as much as building a mansion—then only mansions would be built; shacks would just be old mansions.

3. Long-run Equilibrium

So in a long-run equilibrium, when supply can adjust as well as demand, some types of housing will be produced by building, some by downgrading, and some by both. (For the most part, we can ignore renovation here, since its actual impact is small; generally building from scratch is cheaper.) Clearly there has to be building at some quality higher than every quality supplied by downgrading.

Moreover, if any quality is supplied only by downgrading, then the price households pay for using that quality will be less than the price they would pay for supplying that quality by building. Otherwise some builder would come along and lower the price for households who wanted that quality.

This result has an important implication. If qualities are packed close enough together that the difference in construction cost between any two adjacent qualities is infinitesimally small, then no quality supplied only by downgrading is better than any quality supplied by building.

Suppose this were not the case and some qualities supplied only by downgrading were higher than some qualities supplied by building. Consider the lowest such quality supplied only by downgrading. Call this quality *D*. There must be some quality immediately adjacent to quality *D* and lower than quality *D* that is supplied by building. Call this quality *B*. Since qualities are packed infinitesimally close together, qualities *B* and *D* are as close together as imaginable, and so the cost of building at these two qualities must be as close together as imaginable.

Since housing is actually built at *B*, the price consumers pay for *B*-quality housing is the cost of construction at *B*. Since housing is supplied only by downgrading at quality *D*, the price consumers pay for *D*-quality is less than the cost of construction at *D*, which is approximately the cost of construction

at *B* (it's as close to the cost of construction at *B* as you could ever want to get). Hence the price consumers pay at *D* is less than the cost of construction at *B,* which is the price consumers pay at *B.* But this outcome is impossible at an equilibrium, either long-run or short-run: consumers are not going to pay more for a lower quality of housing *(B)* than they pay for a higher quality *(D).*

Therefore our original assumption—that some qualities supplied only by downgrading could be better than some qualities supplied by building—cannot be correct. All qualities supplied by building have to be better than all qualities supplied only by downgrading.

In a long-run equilibrium, then, you can divide qualities into four different segments. Figure 14.2 illustrates. The highest imaginable qualities have no supply at all; they're just too expensive for anybody to be willing to pay for them. We won't pay attention to these. The next lower interval contains those qualities supplied by building (and possibly also by downgrading); I'll call this the "building interval." The lowest quality in the building interval—the worst kind of house actually built—I'll call the "building boundary."

The next lower interval represents those qualities supplied only by downgrading. I'll call this the "use interval," and I'll call the lowest quality in this interval the "abandonment boundary."

The lowest interval contains those qualities of housing that are not supplied because they are too lousy. There are many reasons why really shoddy housing is not around: perhaps it violates building and housing codes; perhaps no one is willing to pay enough to draw land and other resources into producing it. You can't buy shovels that are guaranteed to break on first use or clothes that smell like three-day-old vomit either, and regulation has nothing to do with this. Markets don't (and shouldn't) produce every single quality of every single good.

This model of the housing market also implies that, under certain conditions, prices in the building interval (except at the building boundary) will be determined totally by supply-side considerations. The important condition is that the cost of building houses of each quality is the same no matter how many houses of that quality are being built; the supply of houses by construction is perfectly elastic. Quigley (1979) gives some empirical evidence in favor of this proposition.

Figure 14.2 How housing is supplied in long-run equilibrium.
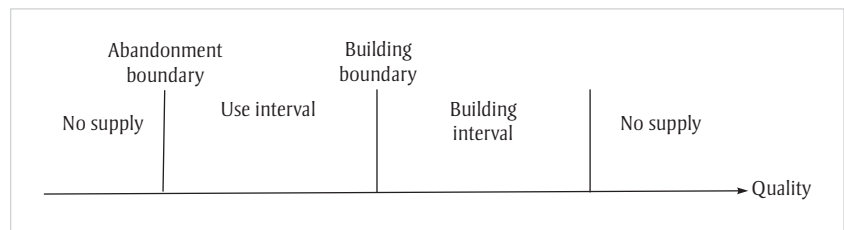
If this is the case, then the market value of any house in the building interval has to equal the cost of building a house of that quality. (If the market value were less, no one would build that variety and so that quality would not be in the building interval; if it were more, everyone would build and supply would exceed demand.) The market value of a house—what people are willing to pay for the privilege of owning it—depends on the price that its services command and the options of downgrading to a slightly lower quality or renovating to a higher quality. The value of those options depends on the market value of houses at those other qualities. Since the other qualities are in the building interval, their market values equal construction cost. Thus, if you know the construction cost of every quality in the building interval, you can figure out what the price of services for those qualities has to be.

In the long run—a time long enough to work out all construction bottlenecks—prices that households pay for qualities in the building interval depend only on construction cost (possibly in some complicated way). Demand doesn't influence these prices. Of course, this result holds only in the long run. But it also holds approximately in the short run if the supply of new houses is very elastic.

Most of the results I've discussed in this section hold for the open-city case as well as the closed-city case. That's because little depends on the exact structure of demand. Thus even in the open-city case, price increases as quality increases, the four supply intervals line up in the same order, and only supply determines prices in the building interval in the long run.

B. Homeowners' Tax Preferences

This setup will allow us to figure out the effect that programs like homeowners' tax preferences, the FHA, and the secondary mortgage market have had on poor people. Since these are national policies, a closed-city analysis will give the right answers.

A popular answer to this question is that these programs have helped poor people and improved the quality of their housing. We will have to assess how good this answer is. As one pretty rich household moves from one house to a better house, the argument goes, a somewhat less rich household moves into their old house, which frees up another house for a household a little less rich, and so on, until, at the bottom of the chain, the poorest household moves up a notch too, and everyone is better off. Notice that this argument pays no attention to prices, and so claims about well-being are not directly supported. Notice also that the favorable consequences are accompanied by an increase in abandonment at the bottom of the house chain.

This process is sometimes called filtering, but the word *filtering* has been used in so many different and imprecise ways over the past several decades that I don't want to use it here. (Grigsby et al. 1987 and Weicher and Thibodeau 1988 offer more detailed discussions of the terminology used.)

Figure 14.3a Short-run effect of more demand for higher quality houses.
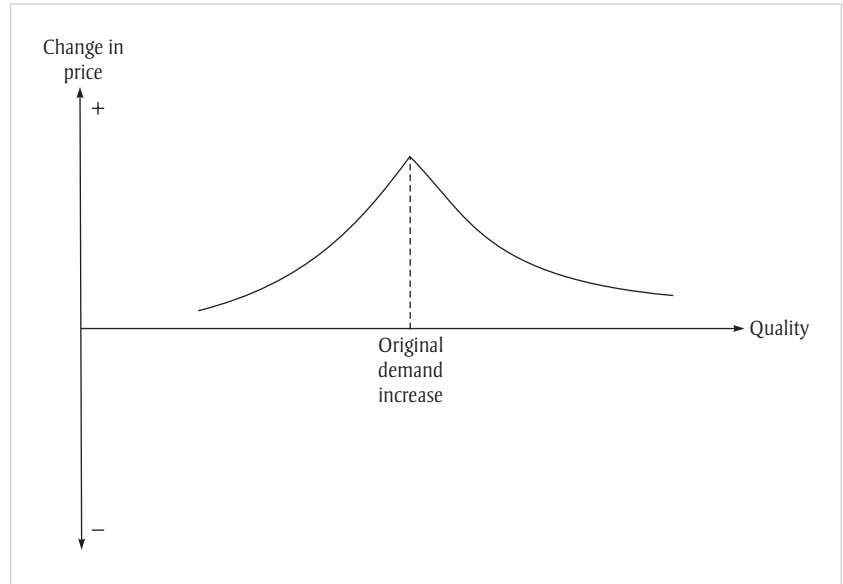
To analyze this line of thinking, let's start with the short-run model in which supply is fixed. Pro-homeownership policies increase demand at high qualities of housing and decrease demand at somewhat lower qualities. The increase in demand at the high quality increases prices there, and at every other quality, with the size of the increase depending on the distance from the quality where demand increased, as in Figure 14.3a. The decrease in demand at the somewhat lower quality decreases price there, and at every quality, in the same fashion but in the opposite direction, as in Figure 14.3b. The total short-run effect of the policies is the sum of these two effects, as shown in Figure 14.3c.

Since poor people are buying qualities closer to the location of the demand decrease than the location of the demand increase, the price-reducing effect will be stronger for them than the price-increasing effect. The price of the lowest qualities of housing will fall, and poor people will be better off. But the effect may be quite small, especially with programs like homeowners' tax preferences, where most of the direct recipients would otherwise be buying rather high qualities of housing.

In the long run, when supply can change, this picture changes, too. All of the programs increase demand at qualities in the building interval, but the effect depends a lot on whether the decrease in demand comes above or below the building boundary.

First, suppose that both the decrease and increase occur for qualities in the building interval. In the long run, prices don't change (or they change little). So the direct recipients of the programs are better off, and the naive sort of impact analysis we did in Chapter 13 turns out to be accurate. Since prices
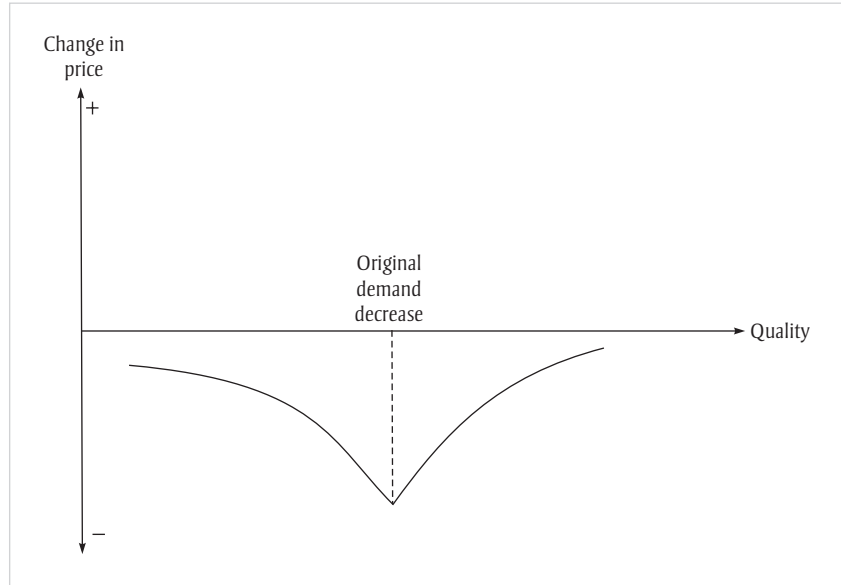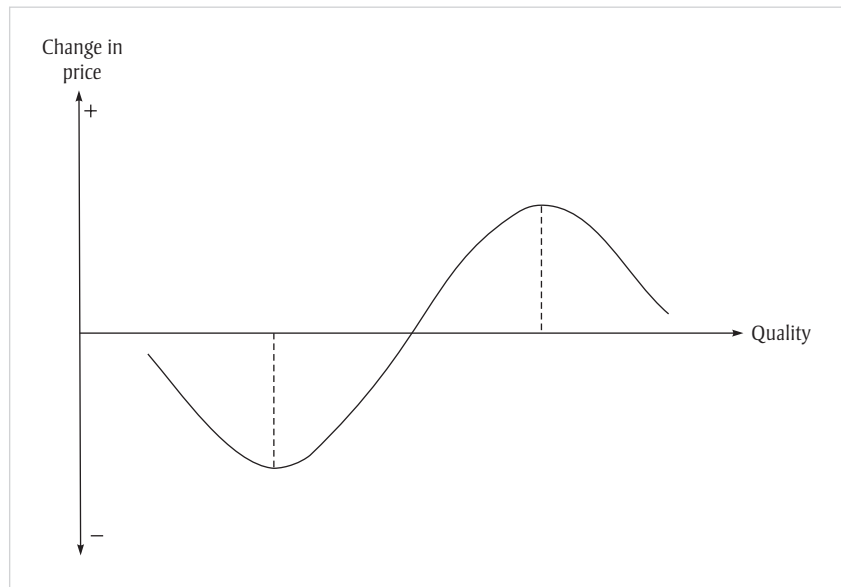
don't change and demand doesn't change anywhere where it matters, nothing else changes. The programs have no effect on poor people.

Alternatively, suppose that demand decreases below the building boundary (but above the qualities most poor people buy) and increases above the building boundary. This may better represent the effects of FHA insurance or

VA guarantees than of current homeowners' tax preferences. If prices fell throughout the use interval the way they fell with fixed supply, housing units would not be downgraded out of the building interval. Lower prices in the use interval would make owners in the building interval and at the building boundary more eager to maintain their houses because they would lose more money if their housing quality deteriorated. So prices have to rise throughout the use interval; average prices in the use interval must return to about what they were without the program.

In this case the effect on poor people is ambiguous. If the FHA or VA programs reduce demand for housing similar to the housing poor people consume, the first effect will be big, and poor people may be better off overall. But if the demand reduction is close to the top of the use interval, the pure demand effect will be small, the supply effect may be significant, and poor people could be compelled to pay higher prices and so become worse off.

In summary, then, careful analysis does not provide a definitive answer about the effect that the dominant U.S. homeownership policies have on poor people. They could be helpful or hurtful or do nothing. The answer depends on parameters that are not well known and that are likely to change often. The strongest theoretical conclusion we can draw is that whatever effect these programs have on poor people and the housing they buy, it's likely to be small.

C. Growth Controls

Growth controls are another program for richer people that might affect the rents poor people pay. We saw in Chapter 8 that suburbs often zone restrictively and make building new houses more difficult and expensive. Newer ideas like urban growth boundaries and smart growth would also raise the cost of new houses.

While homeowners' tax preferences raise the demand for new, high-quality houses, growth controls cut the supply. The difference is crucial. Prices in the building interval are determined by supply in the long run, not demand, and so growth controls can raise the price of high-quality houses while homeowners' tax preferences cannot. Higher prices for high-quality houses reduce the incentive to downgrade, and so cause prices for all qualities of housing to rise. In this model, growth controls hurt poor people.

Somerville and Mayer (2003) provide some evidence that growth controls reduce the rate at which houses are downgraded and become available to poor people. Malpezzi and Green (1996) estimate that moving from relatively unregulated metropolitan areas to fairly heavily regulated ones raises rents for the bottom quartile of renters by more than a fifth, everything else being equal. Raphael and Quigley (2004) provide a summary of the empirical literature on this issue, which is thin but all points in the same direction.

Of course, the result doesn't hold in an open-city model. In the long run, growth controls in one particular metropolitan area wouldn't hurt poor peo-

ple in that metropolitan area—some of them would just decide to live else-where. But a national movement toward growth controls would hurt poor people everywhere.

D. Section 8

Like homeowners' tax preferences, the Section 8 program reduces demand for some low qualities of housing and increases demand for some better qualities of housing. But there are two major differences.

First, with Section 8 both levels of housing affected are likely to be in the use interval, and, as we saw in the earlier discussion of Section 8, they may not be terribly different. The housing that most Section 8 recipients occupy isn't a lot different from the housing a lot of other fairly low-income people occupy.

Second, Section 8 makes households smaller—partly by restricting the number of people who can occupy an apartment, partly by offering larger per capita subsidies to smaller households, and partly by providing enough money to let people live on their own. Sinai and Waldfogel (2002) and Ellen and O'Flaherty (2004) show this result empirically, but with different methods and data.

For both of these reasons, it's likely that Section 8 raises rents for consid-erable numbers of low-income people who live in private housing. Susin (2002) estimates that, on average, Section 8 has raised rents for low-income people by 16 percent. "Low-income" here means the poorest third of renters. Annually, low-income renters gained $5.8 billion in federal subsidies from Section 8 but lost $8.2 billion to higher rents, according to Susin's calculations. Because he was searching for an *average* effect, Susin can't say how Section 8 affects different groups of low-income private-sector renters. Our basic model predicts that the poorest of the poor will gain, while private-sector renters al-ready choosing housing similar in quality to Section 8 housing will lose. If this is true, then Section 8 is fairer than Susin's basic calculation makes it seem, be-cause it helps the poorest of the unsubsidized poor. On the other hand, if this is the pattern that produces the average, it's even more unfair to unsubsidized renters in Section 8–quality housing, because their rents will go up by far more than the 16 percent average.

This problem arises from the Section 8 housing program as it exists in the U.S. today; it doesn't necessarily apply to every demand-side program. A universal entitlement program for poor people, for instance, wouldn't raise questions of fairness—but it might end up giving a lot of money to landlords, with little gain to poor people. Increases in housing allowances in France, for instance, which are received by about half of renters, appear to increase the rents recipients pay (Laferrere and Le Blanc 2004). Market reactions can't be ignored.

E. Public Housing

Since public housing is concentrated in a few cities and the decision to have public housing projects is a local decision, the market effects of public housing should be analyzed according to the open-city model.

In the open-city model, public housing can't change the prices for different qualities of housing in the private market; any price cut would trigger an influx of migrants, any price rise an exodus. So people who don't get into public housing are no better off and no worse off than they would be if public housing didn't exist; these externalities don't exist.

If prices stay the same, then, supply in the long run stays the same too. If households leave private housing for public housing, they will be replaced by increased migration. This will happen even if the buildings they were living in are knocked down, because it will pay to let other buildings deteriorate to take their place. So in an open city, public housing, even coupled with slum clearance, does nothing to reduce the amount of low-quality housing in the long run. There are no external benefits or costs here either.

Public housing makes its tenants better off (and possibly more stable and civic minded); otherwise, it increases only the number of poor people in a city.

Although the long-run open-city story may sound a bit preposterous, most evidence supports it against the alternative scenarios. Studies by Vitaliano (1983) and Weicher and Thibodeau (1988) both found that the supply of public housing units in a metropolitan area had no effect on the incidence of substandard units. Sinai and Waldfogel (2002) found that public housing had almost no effect on the total number of housing units. Only the open-city story carries this implication.

VI. Conclusion

Traditional public housing, the newer supply-side programs, and the Section 8 Existing Housing Program account for only about 5 percent of the total housing in the United States. Blank (1997, p. 109) estimates that only 22 percent of poor families in the United States receive any form of housing assistance. There's little reason to believe that these programs contributed much, if anything, to the amazing improvement in U.S. housing quality that has occurred over the past sixty years. So why should anyone care about this tiny speck of the housing market?

One reason is that publicly assisted housing plays a disproportionate role in stories about how large northeastern and midwestern cities got into their current situation and how they can get out of it. It's instructive that the cabinet department that's supposed to worry about cities is called Housing and

Urban Development, and that most of its budgetary appropriations are devoted to the programs I discussed in this chapter.

Another reason is cities in developing countries. By U.S. standards, most of the people who live in these huge cities are poor and poorly housed. Getting them into better housing is going to be one of the great projects of the twenty-first century, a project that will cost tens or hundreds of trillions of dollars. The U.S. experience is not going to be exactly duplicated in these cities, but certain ideas—the need for incentives, the merits of separating housing and income, the dangers of programs' being hijacked to serve other goals—are probably inescapable.

Finally, bad housing matters a lot to the people who live in it, to their neighbors, and to their children. That's reason enough to think about it.

Questions

1. With the benefit of hindsight, outline a policy that downtown business interests in northeastern and midwestern cities should have followed after World War II, instead of the policies of urban renewal and high-rise construction.

2. In light of the discussion in Chapter 13 about mortgages and homeownership, what are the arguments for and against programs to subsidize homeownership by poor people? Consider, for instance, the use of Section 8 vouchers for mortgage payments, and any other way you can think of that encourages more poor people to own homes. Are these programs equitable? Are they potential Pareto improvements?

3. To what extent should the government subsidize car ownership by poor people rather than housing? To what extent are car ownership subsidies a substitute for housing subsidies? What advantages do they have? Disadvantages? Draw on the information from Chapter 8 about zoning and Chapter 11 about spatial mismatch.

4. In the closed-city model, what's the effect of public housing on the market rents that poor people pay?

References

Allen, Garland E., Jerry J. Fitts, and Evelyn S. Glatt. 1981. "The Experimental Housing Allowance Program." In Katherine L. Bradbury and Anthony Downs, eds., *Do Housing Allowances Work?* pp. 1–32. Washington, DC: The Brookings Institution.

Blank, Rebecca. 1997. *It Takes a Nation: A New Agenda for Fighting Poverty.* Princeton, NJ: Princeton University Press.

Burrows, Edwin G., and Mike Wallace. 1999. *Gotham: A History of New York City to 1898.* New York: Oxford University Press.

Crews, Amy. 1996. "Do Housing Programs for Low-Income Households Improve Their Housing?" Occasional Paper no. 178, Metropolitan Studies Program, Syracuse University.

Currie, Janet, and Aaron Yelowitz. 2000. "Are Public Housing Projects Good for Kids?" *Journal of Public Economics* 75(1):99–124.

Ellen, Ingrid Gould, et al. 2003. "Housing Production Subsidies and Neighborhood Revitalization: New York City's Ten-Year Capital Plan for Housing." *Federal Reserve Bank of New York Economic Policy Review* 9(2):71–86.

Ellen, Ingrid Gould, and Brendan O'Flaherty. 2004. "Do Housing and Social Programs Make Households Too Small?" Working paper, Department of Economics, Columbia University.

Grigsby, William, et al. 1987. *The Dynamics of Neighborhood Change and Decline.* New York: Pergamon Press.

Howenstine, Jay. 1986. *Housing Vouchers: A Comparative International Analysis.* New Brunswick, NJ: Rutgers University Center for Urban Policy Research.

Jacob, Brian. 2004. "Public Housing, Housing Vouchers, and Student Achievement: Evidence from Public Housing Demolitions in Chicago." *American Economic Review* 94(1):233–258.

Kling, Jeffrey R., and Jeffrey B. Liebman. 2004. "Experimental Analysis of Neighborhood Effects on Youth." Working paper, Woodrow Wilson School, Princeton University. Accessed at *www.wws.princeton.edu/~kling/mto/recent.html.*

Kling, Jeffrey R., Jens Ludwig, and Lawrence F. Katz. 2004. "Youth Criminal Behavior in the Moving to Opportunity Experiment." Working paper, Woodrow Wilson School, Princeton University. Accessed at *www.wws.princeton.edu/~kling/mto/recent.html.*

Laferrere, Anne, and David Le Blanc. 2004. "How Do Housing Allowances Affect Rents? An Empirical Analysis of the French Case." *Journal of Housing Economics* 13(1):36–67.

Malpezzi, Stephen, and Richard K. Green. 1996. "What Has Happened to the U.S. Housing Market?" *Urban Studies* 33(10):1807–1820.

Mayo, Stephen K., et al. 1980. "Housing Allowances and Other Rental Assistance Programs: A Comparison Based on the Housing Allowance Demand Experiment." Cambridge, MA: Abt Associates.

National Association of Housing and Redevelopment Officials (NAHRO). 2000. "1997 Picture of Subsidized Households." Accessed at *www.nahro.org/reference/stats_picture1997.html.*

Newman, Sandra, and Joseph M. Harkness. 2002. "The Long-Term Effects of Housing Assistance on Self-Sufficiency." *Journal of Policy Analysis and Management* 21:21–44.

Olsen, Edgar. 1981. "Comment on Kain paper." In Katherine L. Bradbury and Anthony Downs, eds., *Do Housing Allowances Work?* pp. 366–374. Washington, DC: The Brookings Institution.

———. 2001. "Housing Programs for Low-Income Households." National Bureau of Economic Research Working Paper 8208.

Quigley, John M. 1979. "What Have We Learned about Urban Housing Markets?" In

Peter Mieszkowski and Mahlon Straszheim, eds., *Current Issues in Urban Economics.* Baltimore: Johns Hopkins University Press.

———. 2000. "A Decent Home: Housing Policy in Perspective." *Brookings-Wharton Papers on Urban Affairs* 1(1):53–88.

Raphael, Steven, and John M. Quigley. 2004. "Is Housing Unaffordable? Why Isn't It More Affordable?" *Journal of Economic Perspectives* 18(1):191–214.

Riis, Jacob. 1890. *How the Other Half Lives: Studies among the Tenements of New York.* Repr., New York: Hill and Wang, 1957.

Sinai, Todd, and Joel Waldfogel. 2002. "Do Low-Income Housing Subsidies Increase Housing Consumption?" National Bureau of Economic Research Working Paper 8709.

Somerville, C. Tsuriel, and Christopher Mayer. 2003. "Government Regulation and Changes in the Affordable Housing Stock." *Federal Reserve Bank of New York Economic Policy Review* 9(2):45–62.

Susin, Scott. 2002. "Rent Vouchers and the Price of Low-Income Housing." *Journal of Public Economics* 83(1):109–152.

U.S. Bureau of the Census. 1999. *Statistical Abstract 1999.* Washington, DC: Government Printing Office.

Vitaliano, Don. 1983. "Public Housing and Slums." *Urban Studies* 20:173–183.

Wallace, James E. 1995. "Financing Affordable Housing in the United States." *Housing Policy Debate* 6:785–814.

Weicher, John, and Thomas Thibodeau. 1988. "Filtering and Housing Markets." *Journal of Urban Economics* 23:21–40.

Wilner, Daniel M., et al. 1962. *The Housing Environment and Family Life: A Longitudinal Study of the Effects of Housing on Morbidity and Mental Health.* Baltimore: Johns Hopkins University Press.

Homelessness

<div style="border-bottom: 2px solid black"></div>

We began our study of housing in Chapter 13 with a long list of purposes that housing serves: it offers shelter from the elements; security from thieves and predators; a place to store possessions; a venue for preparing and eating meals; refuge for study, privacy, and entertainment; a forum for sociability; prestige; the basic tools to maintain personal cleanliness and hygiene; and the base from which people are rooted in a community through voting, use of schools, and access to other municipal services. These purposes are so varied and so fundamental that it is hard for most people to imagine how they would live without a house.

Yet large numbers of people, even in developed countries, are homeless. Some homeless people, to be sure, find innovative and nontraditional ways to accomplish some of the ends housing usually accomplishes, but many fail to do so. This chapter is about the problems that those failures cause for homeless people and for others.

We begin by thinking about definitions. "Homeless" means many different things and has been used many different ways, probably because housing encompasses so many different attributes. In Section I of this chapter I will try to give some order to these many usages, and to establish some conventions for their use here.

In the next two sections we will look at the problems homelessness causes and the reasons why the lack of homes should be a more compelling object of public and charitable concern than, say, the lack of toothpaste or vegetables. Section II is about why homelessness is bad for homeless people themselves, and section III is about why it's bad for people who aren't currently homeless. Section IV will draw on these sections for a discussion of policy options.

To get the most out of this chapter, you should be familiar with these concepts: efficiency, externalities, and potential Pareto improvement. You can find these terms in the Glossary.

I. What Is Homelessness?

The word *homelessness* has meant many different things to different people over the past century, and so it's a topic that can't be discussed sensibly without some careful preliminary attention to definitions. Different definitions emphasize different aspects of the bundle called housing; essentially, each definition chooses some parts of the bundle and says that you are homeless if you lack those parts.

It is one of the oldest traditions in the United States to view a home as a place through which you are rooted in a family and a normal community. According to this tradition, if you're rootless and unattached, you're homeless, no matter what access you have to physical amenities. Thus in 1973 a leading sociologist could write : "Homelessness is a condition of detachment from society characterized by the absence or attenuation of the affiliative bonds that link settled persons to a network of interconnected social structures . . . [T]he man who occupies the same lodging on skid row for forty uninterrupted years is properly considered homeless. The essence of the concept goes beyond residential arrangements" (Bahr 1973, pp. 5, 7).

In this tradition, homeless men were people who might otherwise be called hoboes (in the early twentieth century), vagrants, or derelicts (in the late twentieth century). They slept outdoors or in lodging places in dilapidated parts of town referred to as "skid row"—neighborhoods with cheap housing ideal for transients, cheap and abundant liquor, and no family institutions. Homeless women were rare, and homeless families impossible—if you were living as a family, you had enough affiliations that you couldn't be considered homeless, even if your "home" was a cardboard box underneath a highway bridge.

This tradition survives in the United States today in the belief that you can identify homeless people by sight on the street in daytime. This is possible if homelessness is not about where you sleep but about how you are tied in with the rest of the world. A few years ago, when I was robbed of five dollars in a subway station, the police officer I reported the incident to asked me whether the perpetrator was a homeless man—whether he was disheveled, smelly, perhaps drunk, perhaps a little wild-eyed. (I said no.)

A considerably different use of the term *homeless* has predominated in Britain. The aspect of the housing bundle that the British tradition has emphasized is the legal right to occupy residential space. The homeless are people who don't have a right to occupy the place where they are sleeping and storing their possessions. During the 1960s and the early 1970s, most discussions of homelessness in Britain focused on squatters (who were plentiful in London at the time). Since the Homeless Persons Act of 1977 was adopted, attention in Britain has shifted to "doubled-up families"—families who are sharing living space with another family, owing to the first family's (often strained) goodwill.

This British usage places far greater attention on what people actually do in the housing market than does the traditional American usage, which looks instead at what housing is supposed to produce (community ties). In the traditional British use of the term, you have to know something about how an individual interacts with housing suppliers to determine whether he or she is homeless; in the traditional American use of *homeless,* you need no such information.

The way scholars and government officials in this country use the word *homeless* today (and since about 1984) is considerably different from either of these traditional usages, but it draws pieces from each. On the one hand, like the British tradition, it focuses on what actually happens to people in the housing market, not on whether they achieve some sort of relationship with the rest of society that housing is supposed to promote. On the other hand, it follows the old American usage by singling out as those housing arrangements that make one homeless the arrangements that vagrants traditionally used. You're homeless if you live like a vagrant.

Specifically, according the modern official definition, the homeless are people who sleep either in shelters or in places not intended for human habitation. The definition thus emphasizes two aspects of the housing bundle: physical access to conventional facilities like beds, doors that lock, and toilets; and participation in the right kind of community setting. You're homeless if you don't have physical access to the basic facilities—because you sleep on the subway, for instance. Even if you do have physical access to these facilities, you're still homeless if you live in the sort of place traditional vagrants lived in—a municipal shelter, for instance, or a skid-row religious mission.

Which of these three definitions is the best one to use? Probably none of them. Rather than emphasizing any particular piece of the housing bundle as crucial, the best procedure is probably to consider several different pieces of the bundle, and to identify several different groups as lacking important aspects of housing. Different groups have different problems, and different policies should be designed to address those problems. Families with small children facing eviction from their aunt's front room have real problems, and so do middle-aged men who sleep in cardboard boxes underneath freeway overpasses. To lump their problems together and pretend they are somehow the same does not do a lot to advance intelligent discussion.

Nobody knows how many people lack important parts of the housing bundle in the United States today. Every night, several hundred thousand people sleep in facilities that are generally called homeless shelters, and an unknown number of similar magnitude, probably somewhere between 100,000 and 500,000, sleep in unconventional places like subways and abandoned buildings. The Census Bureau estimated there were 230,000 homeless people in this country in 1990; 280,000 in 2000. Since some people remain in this condition for only a few days or a few weeks, the number who experience

some form of homelessness in a year is several times greater than the number who experience it on any given night. These numbers, while small in a nation of more than 280 million people, appear to be several times larger than they were in the 1960s and 1970s.

II. Why Homelessness Is Bad for Homeless People

Should you care about something that happens to such a small number of people? The attention and resources devoted to homelessness over the past several decades indicate that many people do care, and so to begin analyzing policies that try to deal with the different varieties of homelessness, we need to start by asking why they are bad. There are two broad classes of problems that the different varieties of homelessness can cause: problems for homeless people themselves, and problems for people who aren't currently homeless. In this section we look at the problems for the homeless.

The most obvious problems homeless people have are those arising from exposure to weather. Those who lack elementary shelter have elevated chances of suffering hypothermia, frostbite, and other cold-related conditions. Johiel (1992) finds that skin infections and thermoregulatory disorders are also more common among people who sleep on the streets, as are chronic respiratory conditions. And being forced to be outdoors in foul weather is just extremely painful, even if you suffer no direct immediate medical effects.

Lack of kitchen facilities and a place to wash are other problems facing people who live on the street. Malnutrition and food poisoning can result, and poor hygiene makes infectious disease more prevalent. Wright (1989) argues that large amounts of walking, and sleeping standing up or sitting up, lead to arterial and venous disease, especially in the legs. Homeless people on the street are also susceptible to traumatic injury, from both accidents and assaults. Security, after all, is one of the main services that housing provides, for inhabitants as well as their possessions.

Homelessness can also cause, exacerbate, or create conditions that look like mental illness. Hopper (1988) summarizes research that suggests that, at that time, about a third of the mental illness found among homeless people arose after they became homeless, not before.

The physical consequences of living in a shelter, as opposed to living on the street, vary greatly with conditions in the shelter. Some shelters offer little in the way of space, privacy, security, or autonomy—for instance, shelters that consist of nothing more than an array of cots in an armory or a church basement. Such shelters are also sometimes implicated in the spread of infectious diseases, like tuberculosis. On the other hand, many shelters do provide much better measures of security, privacy, and hygiene than those that are merely

cots on an armory floor. These shelters do not present the same health and safety concerns.

Even the most secure and hygienic shelters, however, interrupt children's schooling, separate them from friends, disrupt their routines, and often expose them to ridicule from their classmates. Entering or living in shelters is not good for children, although in some cases it may be better than the alternatives.

For adults, too, the social and psychic costs of living on the streets or in shelters are also likely to be large. Getting a good job, for instance, often requires an opportunity to bathe, a clean set of clothes, an address, and a telephone number where messages can be left. None of these requirements is absolutely impossible for even street homeless people to meet (thanks, in part, to some ingenious programs designed to help the homeless), but all of them present serious difficulties—especially when employers who know that an applicant's address is a homeless shelter tend to think that he or she will be less productive and less reliable than others. (See Glomm and John 2002 for a discussion of the labor market implications of homelessness.)

Nor are the social implications of homelessness for adults confined to the labor market. Both street and shelter homelessness make it difficult to take part in the life of the general community, to be trusted, and to "appear in public without shame," to use Adam Smith's phrase. A stigma of moral deficiency attaches to homelessness (of all varieties, partly because many people conflate them), and this stigma imposes a cost on anyone who is homeless.

Since these medical, physiological, and social costs (with the partial exception of those borne by children) are borne by people who, if they had as much money as most of us do, would choose not to be homeless, one possible objection to this list of problems of "costs of homelessness" is that they are really costs of extreme poverty. In this view, extremely poor people are doing as well as they can with the meager opportunities available to them. The terrible costs of homelessness simply represent how bad it is to be extremely poor.

The policy implications of this view, that homelessness is merely a reflection of extreme poverty, are fairly straightforward. The best way to combat homelessness is the best way to combat extreme poverty: with cash. Homelessness is bad because homeless people are unhappy, they are unhappy because they lack money, and money is the most direct and effective antidote for lack of money. If people with money, or with the means to earn money, choose to be homeless, then that may be an unfortunate lifestyle choice, but—except as it affects children and other people who are not compensated—it is not a problem we should worry about. Some people choose to watch World Wrestling Federation on cable TV; other people choose to live in cardboard boxes. So long as each group has the opportunity to do otherwise, their choices cannot be a problem.

An alternative view of homelessness with much the same structure holds that homelessness results primarily from mental illness and substance abuse. The costs of homelessness, then, are really costs of mental illness and substance abuse. In this view, homelessness is just an indicator of mental illness or substance abuse. The way to treat homelessness is to treat mental illness and substance abuse.

Notice that both of these views of homelessness—as a symptom of something else—imply that a precise definition of homelessness is unimportant. If the real problem is extreme poverty or mental illness, the important task is identifying and addressing extreme poverty or mental illness, not figuring out who is homeless.

The views of homelessness as a symptom, however, pose some serious empirical difficulties. The vast majority of people who are extremely poor, who are severely mentally ill, or who are serious substance abusers manage to be conventionally housed. On an average night, maybe half a million people (including children) in the United States are homeless by all definitions, almost certainly fewer than a million. But in 1999, 12.7 million people were severely poor, with income below half of the poverty level (Dalaker and Proctor 2000, p. xii). In the same year, 14.8 million used illicit drugs, 12.4 million were heavy drinkers (five or more drinks on five or more occasions in the last month), and 10.3 million used so much alcohol or drugs and had enough history of use that the government labeled them "dependent" on illicit drugs or alcohol or both (U.S. Substance Abuse and Mental Health Services Administration 2001). (Both poverty and substance abuse data refer to the household population, which excludes homeless people, nursing home residents, and prisoners.) The U.S. surgeon general estimates that about 15 percent of adults suffer from a diagnosable mental illness in a year, and that 2.6 percent suffer from "severe and persistent mental illness"—about 5 million adults in 2000. (U.S. Department of Health and Human Services 1999). Homeless people are a very small fraction of those who have these problems.

Alternatively, you can look at most varieties of homelessness as bad in themselves for the people involved, not just as indicators of extreme poverty, mental illness, or substance abuse. But this requires some departure from the conventional way that neoclassical economists look at peoples' well-being (which is the way we have been looking at well-being throughout this book). The departure is to think of a person's well-being as more or less objectively related to her functionings and capabilities—how well she can carry out some basic pieces of life—rather than as related to her ability to satisfy her desires. This approach was pioneered by Amartya Sen as a way of thinking about very poor people in developing countries—women, for instance, who may have extremely low aspirations for themselves or believe they should not live well. Peoples' desires may be warped by oppressive social systems or be subject to systematic change in the process of development; in either case, their satisfac-

tion would be unattractive as the sole measure of well-being. In substituting an outside observer's idea of what a person's well-being consists of for a person's own idea of what she wants, this approach is clearly paternalistic, but so are many other approaches to housing and homelessness.

The problems with homelessness, in this view, are that those on the street are not sheltered from the elements; homeless people are not secure in their persons or in their possessions; many are subject to excess morbidity and premature mortality; often they are not well nourished; and they are often prevented from appearing in public without shame or from taking part in the general life of the community. All of these are capabilities commonly cited in the economic development literature. Even those homeless people who choose to be homeless or are reasonably happy with being homeless are poorly off, in this view. Homelessness is a problem for homeless people, whether they think so or not.

III. Why Homelessness Is Bad for Other People

Cataloging the problems that homelessness causes for other people is philosophically less challenging. Homelessness of different varieties causes externalities that affect people who are not homeless.

Street homelessness often makes people use large amounts of resources that are usually free but can become expensive if they are used intensively (like anything else). A park bench is a good example. A stroller who sits on the bench for ten minutes on a pleasant spring afternoon is probably imposing no cost on anyone, because no one else wants to sit on it at exactly the same time, and even if someone else did want to sit on it just then, enough good substitutes are probably available that the cost she would have to bear in using another bench would be trivial. Thus, considering the costs of collection and payment, it makes sense under normal circumstances for park benches, like uncongested roads, to be free.

But a homeless person who camps on the bench for days at a time is different from the stroller. He creates congestion where none would otherwise be present. Substantial numbers of people are barred from sitting on the bench, and this presents real costs. The same reasoning applies to the use of benches in transportation terminals, floor space in corridors, bathroom stalls, sidewalk space, and, of course, subway seats. All of these uses can impose costs on other people.

One might argue that, because would-be users of park benches could pay homeless people to leave them and don't, the use of these benches by homeless people is efficient. But even if this claim is true (and it may not be true for some venues like sidewalks and corridors, where thousands of passersby are slightly inconvenienced, or even for park benches because strollers are unsure

of the costs of negotiating), it does not change the fact that the more homeless people there are camping on park benches, the worse off are strollers who would want to sit on those benches.

The story with medical care is similar to the story with park benches, except that the resource being overused is not otherwise free for so large a class of people. Homeless people use disproportionate amounts of medical care, both in emergency rooms and as inpatients in general and psychiatric hospitals. They pay little or nothing for this care. The accounting costs, which include things like overhead, are huge, but the real costs, especially of emergency room treatment, are also probably large. If I am being treated in an emergency room, you may be forced to wait for treatment, and if you are really sick, every minute you have to wait is exquisitely painful. Triage reduces some of these costs, but it doesn't eliminate them. Queuing for a bed in a general or psychiatric hospital is less common in the United States these days than queuing for emergency room treatment, but staying in a hospital still incurs real costs in staff time, food, laundry, and medicines.

Since considerable proportions of homeless people are substance abusers or mentally ill, it may be they would have used a good portion of these medical services even if they were not homeless. Still, since we saw in the preceding section that street and shelter homelessness are both bad for health, it is not implausible that homelessness per se, especially street homelessness, increases these costs. Homelessness also makes hospital stays longer, since homeless people have nowhere to be discharged to.

Other externalities of homelessness are more direct. To the extent that homelessness makes people more susceptible to infectious diseases (this may be especially true of low-quality shelters), people who are not homeless also face greater chances of infection. And unhygienic practices of homeless people living on the street directly affect the quality of life of people who live, work, or travel nearby.

None of these external costs of homelessness has been well measured, but there is plenty of reason to believe that they are real.

IV. Policies

What you think is bad about homelessness should determine how you measure it and what policies you think would combat it. For instance, if you think the major problem homelessness causes is interference with the daily peregrinations of other people in parks, transportation terminals, and on sidewalks, then what you are interested in is a one-night (or, better, one-day) count of street homeless people and others who hang around in the day, no matter where they sleep. If, on the other hand, you worry mainly about the effects of homelessness on children's development, then you will count families in shel-

ters, and probably give more weight to long stays than to short, if it can be shown that damage to children is more than proportional to length of shelter stays. If most of the long-term damage occurs on the first day of a shelter stay, and thereafter it doesn't make much difference to a child's development whether she stays in a shelter for ten days or ten months, then you want to know the number of children who enter a shelter during, say, a year, rather than the number who stay in a shelter on any particular night.

Different policies to alleviate homelessness correspond to different views about what's wrong with homelessness. Since there seems to be little agreement about what's wrong with homelessness, the best way to consider various policies is to start with each of the different conceptions and ask which policies alleviate the problems that that conception identifies.

A. Homelessness as an Indicator of Something Else

If homelessness itself is not a problem but only an indicator of some underlying condition like extreme poverty, mental illness, or substance abuse, then there is no reason why policies should be directed at alleviating homelessness. Instead, they should be directed at the underlying condition.

If the real problem is extreme poverty, then cash is the obvious answer, as I've already mentioned. Giving cash to extremely poor people, though, can create incentive problems, since it's hard or impossible to identify which people would still be extremely poor even if they made reasonable efforts to help themselves. Sometimes policies mix cash (or food) with work or work requirements (see, e.g., Jencks 1994; Dreze and Sen 1990) to mitigate these incentive problems. Alleviating extreme poverty is a difficult and important problem, but it's an issue addressed in labor economics and in public economics, not in urban economics (see the literature in those fields for more on the topic). Even the most successful programs aimed at alleviating extreme poverty will leave some people homeless—those who prefer living in the open, for instance, and those who have significant earning power but are unwilling to do much work. Such homelessness would not present a problem, according to this view.

Similarly, if homelessness is bad because it's an indicator of mental illness, then the only policies you need to consider are those that treat or prevent mental illness. Just as the methods physicians use to treat other types of illness have changed tremendously over the past half century, so too have the methods they use to treat mental illness. The primary shift has been from long-term confinement in large state psychiatric hospitals to a combination of outpatient treatment with medication for most illnesses and confinement in nursing homes for elderly patients with Alzheimer's disease. Antibiotics have reduced the prevalence of some mental illnesses by preventing syphilis, and because people drink less, the diseases associated with alcoholism are also less

common. As with most varieties of medical progress, these changes have probably been for the better, but some patients fared poorly in the transitions and in the new systems.

I would be willing to pay very large amounts of money to avoid becoming mentally ill; so, too, I believe, would most other people. I would also be willing to pay large amounts of money to prevent mental illness among my family and friends; again, I don't think I'm unusual in this regard. Therefore, many better ways of preventing and treating mental illness are likely to be potential Pareto improvements—provided they work reasonably well and don't cost extraordinary amounts.

But whatever treatment methods work best, they work on mentally ill people whether they are homeless or not, and they don't work at all on homeless people who are not mentally ill. Because, as we have seen, homelessness is not empirically a very good indicator of mental illness, it's not a very good way of targeting mental health services. Much the same story holds for substance abuse. Substance abuse policies will be considered in some detail in Chapter 17.

B. Homelessness as Bad in Itself

If homelessness is bad in itself, for any of the reasons I discussed in the preceding section, then policies ought to aim at getting people housed—and housed in a way that allows them to live appropriate lives.

There are a number of different general strategies for getting people housed. You can make some of the personal characteristics that predispose people to be homeless (extreme poverty, mental illness, substance abuse) less common; I'll call this "personal characteristics policy." You can make varieties of housing cheaper and more attractive; I'll call this "housing policy." Or you can make homelessness less attractive; I'll call this "enforcement policy." We will consider each in turn.

1. Personal Characteristics Policy

If fewer people were extremely poor, or mentally ill, or substance abusers, then almost certainly fewer people would be homeless (and many people would be better off). So personal characteristics policies do in fact work in the right direction.

The questions for these policies are about costs and effectiveness. Changing any of these personal characteristics—"curing" someone of poverty, mental illness, or substance abuse—is very, very hard. Some programs work but many do not, and those that do work tend to be expensive.

It's also very difficult to target these services to reduce homelessness. As we have seen, the number of people with these characteristics who are *not* homeless far exceeds the number with these characteristics who are homeless. Within these groups, it's very difficult to predict who will be homeless and

who will not (see O'Flaherty 1996, pp. 242–243, for an analysis of how well various predictive systems do). Programs that treat people with these characteristics will inevitably treat many who would not otherwise be homeless. (This is true even if current homelessness is an eligibility criterion. First, people leave homelessness all the time on their own. Second, it's easy to become "currently homeless" for a short time, and if these programs were truly effective, they would be so attractive that some people would be willing to become homeless for a short period of time to gain entry. See Cragg and O'Flaherty [1999] for some evidence on this process in New York City.)

Even though personal characteristics policies are expensive and poorly targeted, however, they may be the best available way to reduce homelessness. So we will have to examine alternative policies before reaching any firm conclusion about personal characteristics policies.

Before that, though, we need to examine the claim that homelessness is so closely linked to personal problems that only personal characteristics policies are capable of reducing it—that some people just can't manage conventional housing unless their lives get straightened out first. Considerable evidence indicates that this is wrong, and the first piece of evidence is that most people with these characteristics are conventionally housed.

The second piece of evidence against this claim is the large variation over time in the proportion of people with these characteristics who are homeless. By almost any definition, many more people were homeless on any night (both on the streets and in shelters) in the 1980s and 1990s than were homeless in the 1960s and the 1970s—probably several times as many. But between the early 1970s and the 1990s, the number of daily users of illicit drugs hardly changed at all (and may have decreased), while the number of heavy alcohol users decreased considerably. The number of people in extreme poverty increased, but not nearly so much as the number of homeless people.

The relevant population of mentally ill people, too, was approximately stagnant during the period when homelessness rose. The number of severely mentally ill working-age adults who were not in institutions was only a little larger in the 1990s than it was in the early 1970s. The population of state and country psychiatric hospitals fell, but a lot of this decrease was among the elderly. Among mentally ill adults less than sixty-five years old, fewer were in psychiatric hospitals but more were in jails, prisons, and nursing homes. (For more details, see O'Flaherty 1996, chapters 12 and 13.)

Thus personal characteristics by themselves do not uniquely determine who is homeless or how many people are homeless. Other policies can make a difference, too.

2. Housing Policy

If people have easier access to protection from the elements, from marauders, from shame and morbidity, then more people will be able to achieve the basic level of functioning we are concerned about, even if they have mental health

or substance abuse problems. That is the basic premise of housing policies designed to combat homelessness. The various housing policies differ in how they promote this easier access.

The most straightforward way of promoting easier access is to give it to people: to build shelters where anyone who wants it can get basic protection and security without paying anything for it. This has been the predominant policy for dealing with homelessness in the United States. (In Britain and Canada, these establishments are usually called hostels rather than shelters.)

Shelters as a policy have two serious drawbacks. First, providing the basic kinds of protection and security that the capabilities approach calls for turns out to be surprisingly difficult and expensive—much more expensive than a naive engineering view might suggest. Protection from weather is simple and cheap: all you need is a big room with heat and some places to sleep. The bigger, the better because of the two-thirds rule: doubling the number of people sleeping in a big room doesn't double the cost of the walls, floors, and ceilings.

But providing cheap protection from the elements comes at a price in terms of protection from marauders, shame, and morbidity. The more people there are sleeping in a big shelter, the greater the dangers they pose to one another—from the spread of disease; from ordinary criminality; from simple disturbances, good-natured horseplay, and loud talking that can keep people from sleeping. And the more homeless people there are in a single place, the greater the stigma attached to that place. Most of these difficulties can be overcome by spending more money—to provide smaller rooms with better security, more different establishments, more toilets and showers, more intensive cleaning and health care. As you spend more money, shelters become less and less of a cheap and simple policy.

Only very, very large amounts of money, moreover, can resolve the dilemma of enforcing rules in shelters designed to support basic daily-life capabilities. Almost all homeless shelters have to operate with some rules for their residents—no loud partying late at night, no stealing, no assaults on fellow residents, no drug sales, no eternal showers. Rules make sense because fines might be unenforceable with a population generally as impoverished as homeless people, and because for most of these offenses the marginal external cost of the first infraction at any time is much greater than the marginal external cost of more infractions. (This is the criterion I discussed in Chapter 8.) What happens to shelter residents who break rules—or refuse to pay fines? Ultimately the only serious threat is expulsion. Although expulsion deprives the expelled individual of basic capabilities, it may be needed to preserve basic capabilities for the other residents of a shelter.

The second serious drawback to a policy of providing shelters is that they attract people who otherwise would not be homeless. Since shelters that meet basic capabilities need to be safe, hygienic, and orderly, they must be at least as good as much conventional housing, if not better. Who would want to pay

rent for conventional housing not much better than what's available for free in a shelter? This attraction need not take the obvious form of families breaking leases and moving into homeless shelters. It can work in a variety of more subtle ways: men released from prison may not search intensively for a place to live, boyfriends may move out and into a shelter at the first sign of relationship difficulty, extended families may be less willing to patch up quarrels and live with each other. Even more subtle—and probably more serious—is the effect on the supply side of the market: landlords may become more reluctant to supply low-quality housing because they can't compete with the rents shelters charge.

Good shelters will inevitably attract people who would not be homeless if the shelters didn't exist. Screening may mitigate the problem some, but since information about counterfactuals is impossible to observe, screening will either turn away substantial fractions of needy people or admit substantial fractions of non-needy people, or (most likely) both. This is the same difficulty we noted with policies designed to treat personal characteristics.

The other kinds of housing policies don't even try to target people who would otherwise be homeless. Instead, they attempt to make conventional housing more attractive for large classes of poor people, and in doing so, to make homelessness relatively less attractive for large numbers of people who might otherwise be homeless. These are the policies discussed in Chapter 14, on housing for poor people: building code relaxation, supply-side programs like public housing, demand-side programs like Section 8.

Relaxing building codes may decrease homelessness by making cheaper housing available in the conventional market. Even if this cheaper housing is of lower quality, it can still reduce homelessness by expanding the set of options available in the conventional housing market. Of course this policy has costs too: relaxing building codes can cause harmful externalities—more fires or more serious fires, for instance, or faster spread of diseases. And large changes in such codes can make conventional housing no better at promoting basic capabilities than poorly run shelters.

Supply-side housing programs—building more housing for poor people—work in two ways. Some people who would otherwise be homeless receive comfortable, hygienic apartments at modest rents, and, in the short run at least, prices fall for all varieties of low-quality housing (as demand decreases). Some skepticism about the size of both effects is in order. A study by Early (1998) has shown that only a very small proportion of the people receiving housing subsidies (supply-side or demand-side) have personal characteristics like those of homeless people; thus if future recruitment resembles past, new subsidized housing would draw in very few people who would otherwise be homeless. In addition, in Chapter 14 we saw that supply-side programs were expensive, requiring considerably more than a dollar in government funds to produce a dollars' worth of benefit for their beneficiaries.

The primary existing demand-side housing program, Section 8, also reduces homelessness by housing people who would otherwise be without a home. Theoretically Section 8 also lowers the price of the lowest qualities of housing, at least in the short run, by reducing demand for them and increasing demand for better qualities, but we have seen that empirically this effect is tiny, if it exists at all. Early's finding of low participation in subsidized housing programs by people who look as if they might be homeless thus casts doubt on the effectiveness of Section 8, as well, as a tool to prevent homelessness.

Why don't a lot of homeless people get Section 8 vouchers? One reason is bureaucratic hassle: Section 8 vouchers are rationed by a bureaucracy, and those who find bureaucracies difficult to negotiate are least likely to succeed in the struggle to acquire them. Many homeless people—because of mental illness, limited education, or just a predisposition to independence and self-sufficiency—find bureaucracies challenging.

The rules for Section 8 certificates place another handicap on participation by many homeless people: participants must secure an apartment that both meets program standards and rents for the fair market rent or less. Many landlords are reluctant to rent to people whom they suspect are mentally ill or substance abusers; if they can receive the same rent from other prospective tenants, they may not rent to people they believe will be bad tenants. Since the public associates homelessness with mental illness and substance abuse, it should be no surprise that New York City, for example, pays large cash bonuses to landlords who accept families leaving shelters, even when those families do carry Section 8 certificates with them.

Of course, the difficulties of reducing homelessness through the Section 8 program don't imply that all possible demand-side programs would face similar challenges. Mansur and colleagues (2002) argue that, even without major modification, a universal Section 8 program would probably cut homelessness by between a third and a half in California cities (a result not inconsistent with Early's findings). In Chapter 14, we examined some alternatives to Section 8, such as a universal percentage-of-rent formula or a universal lump-sum formula. By reducing bureaucracy, eliminating quality standards, and permitting additional payments, both of these variations would give more assistance to people who would otherwise be homeless. Universal programs, however, would also serve many people who would not otherwise be homeless.

Notice that this is the same conclusion we have reached for the other plausible housing policies and personal characteristics policies we've considered: they all serve many people who would not otherwise be homeless. To judge among these policies, then, we need to consider both how many extra people are included and what the social cost of that inclusion is. If the benefits that each extra person receives are very great, or if there are sizable external benefits, then the net social cost of serving extra people may not be large.

3. Enforcement Policy

Fewer people sleep on sidewalks, in transportation terminals, and beneath bridges when police roust them from those sleeping places, arrest them, or confiscate their belongings. Enforcement policies thus can be effective in reducing some varieties of homelessness, but not all (they may even increase shelter population).

Enforcement policies make homeless people worse off, not better off, and so we would almost automatically rule them out if we were doing a standard economic analysis that equated well-being and desire satisfaction. But the capabilities approach allows a person's well-being to be separate from her desires, and so we must ask whether these policies can expand a person's important capabilities, even if they thwart satisfaction of her desires.

Sometimes a very affirmative case can be made for enforcement: to protect the mentally ill person, for instance, who refuses to take refuge from a blizzard or an ice storm. At other times, enforcement policies may have serious drawbacks:. They may further stigmatize the homeless, for instance; push them further from the life of society by making their every movement illegal; make them less secure by denying them basic police protection and discouraging them from reporting crimes to the police; force them into ever more obscure nooks and crannies, where their health is even more seriously impaired; make them paranoid. Whether enforcement policies induce more homeless people to have a better relationship with the rest of society (by becoming conventionally housed or entering shelters) or a worse one is an empirical question for which few answers are known. Different policies probably have different effects at different times and places.

Enforcement policies have costs, too. The most obvious one is police time and resources: often police could be preventing or investigating other infractions, always they could be relaxing or getting to know their public better. Many police officers find this sort of activity distasteful, either because they feel sorry for homeless people or because they don't like being in contact with them. Laws and procedures designed to keep homeless people out of certain areas must be enforced uniformly, and so people who are not homeless also lose out when playgrounds are locked up at dusk or benches are removed from public plazas. Plus there are the costs that homeless people bear when they are rousted.

C. Homelessness as Bad for Other People

Finally, we need to consider policies directed at reducing the external costs of homelessness, the costs that other people bear because homeless people get in their way, or draw too much on common resources, or spread diseases. We have already done much of the analysis required, because the primary way to reduce externalities is to reduce homelessness, and the primary tools for doing

this are the three types of policies described above. Anything that reduces homelessness also reduces its externalities. What we need to do here is note some differences when the focus is on how homelessness affects others.

One difference is in the groups to whom attention is paid. Except possibly for the spread of infectious diseases, shelter homelessness produces far fewer noxious externalities than street homelessness. Indeed, the external costs of panhandling by people who sleep in conventional housing may well be larger than the external costs of sleeping in reasonably hygienic shelters. So housing policies that move people from street to shelter are important for reducing externalities, even though they're probably less important for promoting basic capabilities for homeless people themselves.

The argument for enforcement policies is also probably stronger than it was when we considered only the well-being of homeless people. Clearly enforcement policies reduce the externalities of homelessness; whether they promote basic capabilities is much less clear. Shooing a homeless person off a park bench may or may not be good for that person, but it's almost surely good for the stroller who wants to sit on that bench (although just because it's good for the stroller doesn't mean it should happen).

V. Conclusion

Homelessness denotes many different things; the word has become mired in a plethora of different, imprecise meanings. Starting with problems associated with homelessness, and then identifying groups with those problems and policies that address them, is a more productive approach than starting with an emotionally charged word, trying to divine its "true, deep" meaning, and then looking for the policies that flow from that meaning.

In looking at the different kinds of homelessness and the problems associated with them, we have revisited several different topics covered earlier in this book: public health and externalities, rules and public order, building codes and housing markets, programs to help low-income renters. One can't understand homelessness and the policies associated with it without a good understanding of these issues. But two issues that come up repeatedly in discussions of homelessness have not been addressed yet in depth: crime and security, and substance abuse. I will take up these topics in the next two chapters.

Questions

1. How would you define and measure homelessness in a developing nation like China or Brazil?

2. Suppose you could identify a large sample of individuals who as children had spent varying lengths of time in different family shelters and were now young adults. What would you try to find out? How would you go about trying to find it out? What would the weaknesses be in your conclusions?

3. In Chapter 8 I discussed Ellickson's proposal to establish different zones in which different kinds of street behavior would be permitted or prohibited. How would this proposal affect enforcement policies against homelessness? How would this affect the external costs that homelessness imposes? How would it affect the well-being of homeless people? Should zones be established where "anything goes"? What should be prohibited in the least restrictive zones?

References

Bahr, Howard M. 1973. *Skid Row: An Introduction to Disaffiliation.* New York: Oxford University Press.

Cragg, Michael, and Brendan O'Flaherty. 1999. "Do Homeless Shelter Conditions Determine Shelter Population? The Case of the Dinkins Deluge." *Journal of Urban Economics* 46:377–415.

Dalaker, Joseph, and Bernadette D. Proctor. 2000. "US Census Bureau, Current Population Reports." Series P60-210, *Poverty in the United States: 1999.* Washington, DC: U.S. Government Printing Office.

Dreze, Jean, and Amartya Sen. 1990. *The Political Economy of Hunger.* New York: Oxford University Press.

Early, Dirk W. 1998. "The Role of Subsidized Housing in Reducing Homelessness: An Empirical Investigation Using Micro-Data," *Journal of Policy Analysis and Management* 17:687–696.

Glomm, Gerhard, and Andrew John. 2002. "Homelessness and Labor Markets." *Regional Science and Urban Economics* 32(5):591–606.

Hopper, Kim. 1988. "More Than Passing Strange: Homelessness and Mental Illness in New York City." *American Ethnologist* 15:155–167.

Jencks, Christopher. 1994. *The Homeless.* Cambridge, MA: Harvard University Press.

Johiel, R. 1992. "Health and Health Care of Homeless People." In M. J. Robertson and M. Greenblatt, eds., *Homelessness: A National Perspective.* New York: Plenum Press.

Mansur, E. T., et al. 2002. "Examining Policies to Reduce Homelessness Using a General Equilibrium Model of the Housing Market." *Journal of Urban Economics* 52(2):316–340.

O'Flaherty, Brendan. 1996. *Making Room: The Economics of Homelessness.* Cambridge, MA: Harvard University Press.

U.S. Department of Health and Human Services. 1999. *Mental Health: A Report of the*

Surgeon General. Rockville, MD: Substance Abuse and Mental Health Services Administration.

U.S. Substance Abuse and Mental Health Services Administration, 2001. "National Household Survey of Drug Abuse." Accessed at *www.samhsa.gov.*

Wright, James D. 1989. *Address Unknown: The Homeless in America.* New York: Aldine de Gruyter.

Chapter 16

Crime

$\mathcal{N}$*o one has ever* been raped by a tree, assaulted by a mushroom, or robbed by a sheep. People are the ones who do these things, and for the most part they are in close physical proximity to their victims when they do them. You can't get punched by e-mail or murdered by fax (phone sex and Internet fraud get a lot of media attention precisely because they are exceptions to this general rule). If you want to avoid crime, you can go live in the woods by yourself. I met a ranger once in the George Washington National Forest who said he had never taken the key out of his car's ignition.

Since cities are places where a lot of humans are in close proximity to one another, they are places where a lot of crime occurs. We have already seen that in cities people harm each other unintentionally—by driving cars, spreading germs, occupying subway seats, building ugly hotels—but they harm each other deliberately, too. You can't have a city without the threat of crime. Nor can a city prosper unless crime is held to manageable levels.

This chapter is about crimes that have victims. Fear of becoming a victim keeps people from fully taking advantage of city life, and so these are the crimes we should start our examination with. The serious crimes of this sort, the ones we will concentrate on, are called "index crimes" because they appear on the FBI's index of serious crime. Several of the more important crimes without obvious victims—those involving drugs, alcohol, and guns—will be discussed in the next chapter.

I will also omit from this discussion crimes for which proximity between offender and victim isn't particularly important—the so-called white-collar crimes like fraud, embezzlement, and antitrust violations. The amounts of money involved in these crimes are much larger than those involved in crimes for which proximity matters. David Anderson (1999, p. 625) estimates the an-

To get the most out of this chapter, you should be familiar with these concepts: Coasian bargaining, externalities, flight-from-blight hypothesis, Pareto improvement, potential Pareto improvement, sprawl, and transfer. You can find these terms in the Glossary.

nual amount transferred through occupational fraud ("the use of one's occupation for personal enrichment through the deliberate misuse or misapplication of the employing organization's resources or assets"), unpaid taxes, and health insurance fraud ($436 billion) to be around thirteen times as large as the total amount transferred through burglary, household larceny, personal theft, motor vehicle theft combined ($33.5 billion). Coupon fraud alone transfers more resources than robbery does. But white-collar crimes are not an issue for urban economics in particular; they're probably best addressed in studies of finance, industrial organization, and health economics. When tourists express fear of Manhattan, or when neighbors gather in someone's living room to try to "do something" about crime, it's not health insurance fraud they're talking about.

The problems that index crimes cause are multiple. They are bad in themselves—they deprive innocent people of things they are entitled to, and they do so inefficiently. They may discourage some people from living in cities, and among those who do live in cities, they may discourage living and working in the most densely populated and vibrant areas. People may invest in goods and skills they otherwise wouldn't have purchased in order to avoid crime—karate lessons and Doberman pinschers, dead bolts and The Club for their car, mace and bicycle locks, private security guards and burglar alarms. People may drive instead of riding a bus, take a cab instead of walking, or stay home at night. When crime fell in many American cities in the 1990s, life got better in many different ways.

Crime and the efforts to reduce crime also exact a toll on racial relations; they make it more difficult for people to cooperate harmoniously and productively. Minorities are disproportionately the victims of several of the crimes we will consider in this chapter—murder and robbery, especially—and they are disproportionately arrested, convicted, and incarcerated for these crimes. The perception that minorities are more likely to commit these crimes has had wide and serious consequences. It has contributed to racial and social segregation by making many whites more reluctant to live, work, or visit in minority neighborhoods, and by encouraging barriers that keep minorities out of largely white neighborhoods, schools, and gathering places. It has become an integral part of racist stereotypes. And it may exacerbate styles of policing that many minority-group members perceive as unfair.

Successful efforts to reduce the threat of crime can therefore have big payoffs, and so can policies that change how people respond to the threat of crime. Understanding what some of these policies are and how they work is the goal of this chapter.

I will begin in section I by describing the crimes that will occupy our attention in this chapter—what they are, how many of them are committed, who the offenders are, who the victims are, where they happen, what the di-

rect costs are. In section II we will look at the indirect costs of crime and try to assess some of the empirical evidence for these costs.

In the remainder of the chapter we will look at policies to combat crime. Section III is about policing; section IV, incarceration; and section V looks at crime prevention through programs that help people who might become criminals.

I. What Crimes?

Index crimes are gross violations of a victim's bodily integrity and simple property rights. In Coasian terms, it is hard to think of viable property-rights regimes that did not give people the entitlements that are taken by these crimes. Historically, almost all societies have thought of such takings as crimes, although sometimes some actors have been exempted from prosecution. By and large, then, the list of index crimes we today use is not controversial or biased by Western or modern perspectives. (The most conspicuous exceptions are probably marital rape and assaults by husbands on their wives, neither of which is universally considered to be a crime; in some societies wives are thought to have surrendered to their husbands certain entitlements to bodily integrity.)

There are seven index crimes. Murder, rape, robbery, and aggravated assault are called crimes of violence because the offender physically confronts the victim and harms him or threatens to harm him (the vast majority of victims are male). Theft, burglary, and motor vehicle theft are called property crimes: the offender takes something valued by the victim without inflicting or threatening to inflict bodily harm.

More specifically, murder means taking someone's life. Rape is a sexual assault, threatened or completed (it's the only violent crime in which most victims are female). Assaults are attacks, whether completed or attempted. Somebody jumps you, punches you, beats you up, slashes you, or shoots you. Barroom brawls, vengeance beatings, and domestic violence all qualify as assaults, as do threats that are not carried out. (No one steals anything from you in an assault. If they do, or try to, it's a robbery, not an assault.) Aggravated assaults are assaults in which weapons are used (whether completed or just attempted, whether or not an injury occurred), and assaults in which serious injuries occur (whether or not weapons are used).

The other four index crimes all involve attempts to take something of value from the victim. In a robbery, the offender threatens or inflicts bodily harm. Muggings and gas-station holdups are robberies. Burglaries require entering someone's house or place of business without authorization. Motor vehicle thefts are attempts, successful or not, to steal cars, trucks, or buses. Theft

is everything else: your laptop computer disappears when you leave it unattended in the library, or your house guest walks off with your stamp collection. Shoplifting is classified as a theft.

These seven major crimes are not the only crimes with victims. Of the others, though, the more serious ones, like arson and kidnapping, are rare, and the more common ones, like pocket picking, graffiti, and disorderly conduct, are not serious. The seven major crimes I've described are the crimes that people worry about.

All of these crimes are urban; the more people there are in a small area, the more likely they are to occur. The urban bias of some crimes, such as murder and motor vehicle theft, is greater than the bias for others, like rape and burglary, but living in a city makes you more likely to be a victim of each of these crimes.

African Americans and Hispanics are more likely to be victims of these crimes, and to be incarcerated for them. Again, this is most emphatically true for murder and robbery, and considerably less true for burglary and assault, but it holds for all major crimes. Even though poor people are more likely than rich people to be victims of most crimes (theft and motor vehicle theft being the exceptions), income explains little or nothing of the racial differences in victimization (and not much, either, of the racial disparities in incarceration).

All of these crimes are also disproportionately committed by men. We don't know why. Households who own their own homes and who live in single-family homes are less likely to be victims of property crimes, even theft, than households with other living arrangements. Since one purpose of having housing is to protect your possessions, part of this advantage may arise because owner-occupied single-family homes afford better protection. The difference in theft is more puzzling. Perhaps people with larger and better homes don't have to carry as much valuable stuff around in their backpacks and purses or leave it on the backseat of their car. Perhaps they spend more time at home and less time in public places where they could be victims of theft. Or perhaps people who own single-family homes have different habits and dispositions from those who don't, and so would have a lower victimization rate even if they lived in identical physical settings. The evidence on theft indicates that better physical protection probably isn't the sole reason why those who live in single-family and owner-occupied homes are safer from burglary and motor vehicle theft.

On the other hand, methods and motives for committing crimes, and detriments to and ways of avoiding crimes, differ greatly. To pretend that there is some monolithic activity called "crime" is just as silly as pretending that all legal activities, from mining coal to watching baseball, eating grapefruit to writing novels, are somehow the same, have the same incentives, and follow the same patterns. Crime is motivated by passion, lust, greed, the need to feed

a habit, and the desire to show off in front of high school classmates; its victims lose lives, integrity, health, watches, and cash; they avoid crime by staying home, leaving home, buying pit bulls, and keeping their money in their front pocket. Barroom brawls are different from break-ins in the middle of the night, wife beatings from joyrides; assassinations of rival drug sellers from thefts of classmates' Game Boys.

II. Benefits of Reducing the Threat of Crime

Victims of crime would be better off if the crimes weren't committed and everything else were the same; that's why they're called "victims." They take steps on their own to avoid becoming victims, and governments also initiate costly activities intended to reduce the number of crimes that occur. Most discussions about crime are about the wisdom, or lack thereof, of these government activities, and ours will be, too. As usual, we'll be looking for potential Pareto improvements.

First we need to think about why reducing crime is a good thing; that's what this section is about. But we need to be more precise: because victims and potential victims can change their behavior, the relevant goal is reducing the *threat* of crime, not crime per se.

Suppose that most of the people in a town are afraid to go out at night because those who do go out are often mugged. Then the town government installs brighter streetlights and beefs up police patrols at night. The threat of crime falls; the probability that a person will be mugged if he goes out at night goes down. As a result, more people take evening strolls, go to the movies, eat at neighborhood restaurants. More people on the streets means more muggings. Suppose that so many more people go out at night that, even with the reduced individual probability of being mugged, more muggings occur than before the safety improvements. Were the safety improvements a failure?

Of course not. All of the law-abiding citizens of the town are better off (the muggers may be, too, but that's a different issue). They have a new way of spending their evenings—strolling with a low probability of being mugged—but they still can sit home if they want to. That some of them take advantage of this new opportunity indicates that they're better off. They would be willing to pay something for this new opportunity, even though taking advantage of it increases their probability of being mugged, relative to cowering at home in the evenings. It's the reduction in the threat of being mugged that ultimately matters, not the number of actual muggings.

An alternative way of coming to the same result is to think separately about the benefits and costs of strolling at night after the safety improvements have been made. The benefits are the enjoyment of the activity—what you would be willing to pay to stroll at night if mugging were impossible. The cost

is your expected loss from being mugged, taking into consideration the probability of being mugged if you stroll in the evening. You will go strolling only if your benefits exceed your costs, and so everybody who goes strolling has benefits greater than costs. People with costs greater than benefits remain at home; they lose nothing. People who would be strolling anyway gain because they are less likely to be mugged. So aggregate benefits are greater than aggregate costs (not counting the costs of lighting and patrol).

Therefore we can think of two groups of people as gaining from a reduced threat of crime: those who don't change their activities and those who do. The former gain because they are less likely to be victims of crime; this is a direct benefit of reducing the threat of crime. The latter gain because they choose to undertake riskier activities; this is an indirect benefit. In this section we will look first at direct benefits and then at indirect benefits.

A. Direct Benefits

The direct benefits of crime reduction are fairly easy to think about. Most Americans treat small reductions in the probability of death as if they valued their lives at about $5 million to $10 million, and so, say, a reduction of 100 per 100,000 in the murder victimization rate of young black men—which would bring it into the same range as that for young white men—would be worth $5,000 to $10,000 to each young black man. We know less about willingness to pay to avoid the pain and suffering associated with rapes and assaults, the fears associated with robberies, the injuries and deaths caused by stolen cars, but they would all be significant amounts.

With property crimes, the value to victims of property lost is not only easy to calculate but fairly well known. Burglary victims often say they feel "violated" or "attacked," and this is an additional cost; having someone break into your house and take $200 out of your dresser feels a lot worse than losing $200 in the stock market.

On the other hand, reducing property crime reduces offenders' income, and this is a loss to them. Are these costs that should be offset against the benefits of crime reduction? A straightforward and simple-minded application of the potential Pareto improvement criterion says that they are. But a more nuanced interpretation leads to different conclusions.

Remember that I have never argued that there is anything particularly morally compelling about the potential Pareto improvement criterion itself; its value comes from pointing the way to actual Pareto improvements. The reason we should do something that creates a potential Pareto improvement is because it can lead to an actual Pareto improvement, and usually that actual Pareto improvement is a morally good thing. But sometimes that Pareto improvement is impossible, and when it is, the argument for the potential Pareto improvement is vitiated.

To be concrete, suppose that installing a video surveillance camera would stop Mr. X from stealing $10,000 from Mr. Y. The video surveillance camera would cost $4; $2 each in taxes would come from Mr. X and Mr. Y. Clearly, the potential Pareto improvement criterion would reject the camera; it uses real resources but only prevents a transfer. The argument against the camera has to appeal to an actual Pareto improvement: if no camera is installed, and Mr. X pays Mr. Y $10,001 after he steals $10,000 from him, then Mr. X is $1 better off than he would be if he had to pay $2 for a camera that kept him from stealing, and Mr. Y is $1 better off than if he had to pay for his share of the camera. The situation without the camera and with these transfers is Pareto superior to the situation with the camera.

But no one would accept this argument against the camera. The transfers that make getting rid of the camera a Pareto improvement are clearly impossible. Mr. X, who stole the money, is not going to give it back, and more, without detection and compulsion. Without the corresponding actual Pareto improvement, there is no compelling case for a potential Pareto improvement.

Essentially, then, to the extent that criminal's gains cannot be tapped to compensate victims as they were in the simple example without the camera, then they should not be considered part of the gains of crime; their loss should not be considered a cost of reducing crime. This reasoning applies to whatever perverse pleasures criminals may gain from violent crimes, as well as to their pecuniary gains from robbery and property crimes. (In a few special cases in which offenders are ordered to pay restitution or when their gains are subject to income taxation, this blanket conclusion may not hold true, but those special cases can be considered on an individual basis.)

B. Indirect Benefits

The indirect benefits of crime-threat reduction are realized because potential victims alter their behavior: they undertake activities they would otherwise have forgone, and they forgo activities they would otherwise have undertaken. As the story of the streetlights and mugging showed, these altered activities may cause more crime, and the direct costs of this added crime must be subtracted from the benefits of altered behavior.

In studying indirect effects, economists have focused most of their attention on the actions people take because they fear crime; reducing the threat of crime would release the time and resources that go into these activities. "Twenty dollars spent on a door lock is $20 that cannot be spent on groceries" (D. Anderson 1999, p. 616). Because they fear index crimes, people spend money on security guards (about $18 billion a year); alarm systems ($6.5 billion); computerized pass systems for business access ($4.7 billion); locks, safes, and vaults ($4.4 billion); small arms and ammunition ($2.3 billion); surveillance cameras; safety lighting; protective fences and gates; nonlethal weaponry,

like mace; guard dogs; and library theft detection devices. David Anderson (pp. 623–624) estimates that Americans spend $90 billion worth of time a year locking and unlocking homes, cars, offices, gym lockers, bicycles, and so forth—and looking for the right keys. Security concerns add substantially to the cost of windows and doors; and much of the work that receptionists do is directed more toward keeping people out than helping them come in.

All told, the cost of these items ($130 billion, according to Anderson), is about the same as the combined cost of murder (about $90 billion) and the losses of property crimes (about $40 billion). The costs of pain, anguish, and humiliation from nonmurder violent crimes are large, and so there is little doubt that the direct costs of crime are greater than the indirect. But the indirect costs are large, and so you would expect that a large proportion of the benefits of reducing the threat of crime come from reducing the resources that flow to these activities. Or more precisely, since most of these activities are meant to deter crime by strangers, a large part of the benefit of reducing the threat of crime by strangers comes from reducing these activities.

The positive effect of crime-threat reduction—freeing people to undertake new activities—has been much less studied and is much more difficult to quantify, but it is real nevertheless. Less threat of crime by strangers lets people go more places, stay out later, work later, walk instead of taking cars or cabs, meet more new people, enjoy parks and public spaces more, carry and use valuable items like laptop computers in more locations, wear more expensive jewelry more often, ventilate their houses better, and in general live freer, gentler, more productive and easier lives.

Reducing the threat of domestic violence and other crimes by intimates also gives people valuable new opportunities. People want to live together for many reasons—love, child rearing, and the economies of scale that come from the two-thirds rule are probably the prime ones. Fear of domestic violence can keep them from doing so. A woman and a man may derive great advantages from raising their kids together rather than incurring the expense of maintaining separate households (and the kids benefit, too), but if no safeguards are present to make the woman believe the man's promise that he will never lay hands on her again, they will live apart. Letting more, bigger, and better households form and stay together is the big indirect benefit of reducing the threat of domestic violence. Since people value a lot the person they live with (how many million songs are there on this topic?), this indirect benefit may be very large.

Notice that, in the case of domestic violence, children benefit as well when a couple is able to live together; the actions allowed by reducing the threat of crime create external benefits. Other crimes have this effect, too: when people walk instead of taking cabs, they cut congestion; when they leave their windows open instead of installing an air-conditioner, they cut pollution and slow global warming.

The most frequently cited externalities of reducing the threat of crime deal with safety. Here there are costs as well as benefits. Ayres and Levitt (1998), for instance, show that the more cars there are with LoJack systems (hidden locator devices that allow police to track stolen cars unbeknownst to their possessors), the less likely are cars that don't have LoJack to be stolen. The LoJack system discourages car theft in general because potential thieves don't know which cars have it. Every three cars with LoJack reduce car thefts by one a year, although LoJack owners don't change their victimization probabilities at all, since LoJack is unobservable.

On the other hand, visible deterrents, like barbed wire, The Club, and conspicuously advertised burglar alarm systems may divert criminals from my house or car to yours; you are better off if I don't have them. The more people there are walking on a certain street, the safer each is, since there are witnesses, interveners, and people to call the police. The external effects of gun possession, however, are a controversial topic that I defer to the next chapter.

Except in a few instances, such as with the LoJack product, few data are available on the size of these effects. They may go in opposite directions or cancel out: if reductions in the threat of car theft cause me to forgo buying a LoJack device for my car, that's bad for other drivers; if they cause me to forgo buying The Club, that's good for other drivers. So the net effect of safety externalities is unknown.

Whatever the impact of victim precautions on others, many types of precautions are successful in protecting the people who take them. Bus robberies in New York City stopped when an exact fare system was adopted; the presence of bank guards in the Philadelphia area deterred bank robberies there; homes in Portland, Oregon, displaying stickers indicating that possessions had been engraved with identification numbers were less likely to be burglarized. (For a review of this literature, see Cook 1986). When bank robberies fell by 31 percent between early 2003 and early 2004 in New York City, the police commissioner attributed the fall to greater use of bandit barriers, dye packs, security guards, and conspicuous video cameras (Wilson 2004).

Since the value of resources devoted to private protection efforts probably exceeds substantially the value of resources devoted to the public criminal justice system (especially when you consider the resources the public system devotes to victimless crimes), and since these resources have some effectiveness, you would expect that changes in the precautions victims take have major effects on the behavior of potential criminals—effects at least as great as those caused by changes in policing and imprisonment.

C. Effects on Location and Sprawl

We have seen from the data on individual crimes that living in a metropolitan area is a risky activity, and living in a central city is riskier than living in a sub-

urb. One of the indirect benefits of reducing the threat of crime, then, may be that it lets more people live in cities, and in the denser parts of cities. Fighting crime is fighting sprawl.

Notice I am not arguing that if there were no crime, there would be less sprawl, and so some part of sprawl is inefficient. Crime is inevitable; spatial arrangements that work only in the absence of crime are about as relevant as spatial arrangements that work only in the absence of gravity. Nor am I arguing that if the probability of becoming a crime victim were uniform across the landscape there would be less sprawl, and so sprawl is bad. Just as the probability of contracting many communicable diseases and the probability of meeting an expert piano-tuner are higher in cities than elsewhere, there are many good reasons to think that cities are inevitably more dangerous.

Part of this danger probably arises from victim behavior. If you're afraid your spouse will occasionally beat you up, you probably don't want to live where you're two miles from your nearest neighbor and thirty miles from the police. If you go out for dinner rather than eat at home, you raise the probability of burglary (because your house is unattended), theft (because someone could take your wallet in the restaurant's restroom), robbery (because you could get mugged on the way home), and motor vehicle theft (because you parked on the street near the restaurant rather than leaving your car in a relatively safe garage you control). If you go to a bar afterward instead of having a beer on your front porch, you raise the probability of assault. And going out for dinner is something you're more likely to do if you live in Manhattan than if you live on a farm in Nebraska.

Offender behavior matters, too. You have a much better chance at finding someone to mug if you lurk on a street in San Francisco than if you stalk passers-by on a country lane or try to jump on cars on an isolated stretch of I-80. If you're looking for a house to burglarize and are going to keep looking until you find an open window or an unlocked door, you're better off looking where you can try fifty doors an hour rather than five. Cities teem with wealth—with stores, businesses, and wealthy individuals; Glaeser and Sacerdote (1999) estimate that between a sixth and a third of the urban crime difference is due to the greater attractiveness of the targets. You may also have a better chance of getting away with a crime against a stranger in a city. City life is more anonymous, and most crimes have many plausible suspects. Glaeser and Sacerdote argue that apprehension probabilities in big cities are in fact lower than elsewhere, and that this accounts for another part of the difference.

Finally, people in cities may be different. Female-headed households are more common in cities (housing units are smaller and supports are closer at hand), and kids who grow up in such households may be more likely to become criminals (or victims, because they don't have fathers present to protect them).

Thus, higher crime rates in cities are not indications of a problem by themselves. You can't get the advantages of city life without getting some of the disadvantages, too.

My argument is much less grand. Suppose there is some program that would reduce the threat of crime, either generally or especially in cities. Then one of the benefits of that program is that it will allow more people to enjoy city life and reduce the external costs associated with sprawl. Moreover, if initiating that crime-reduction program is a potential Pareto improvement (or a potential Pareto improvement if you ignore losses to criminals) and it is not done, then you can say that sprawl is excessive. (But of course it would be a whole lot more straightforward just to say that the crime-reduction program is a good idea.)

This is a sophisticated version of the idea of flight from blight. For this version of flight from blight to be important and relevant, two conditions have to be met. First, good crime-reduction programs have to be available but not implemented. And second, the level and distribution of the threat of crime has to make a big difference to where people live and work in metropolitan areas.

Economists have paid more attention to the second question than to the first. The first condition is in the province of criminology, and while I'll say something about it in the next three sections, I'm not qualified to give much of an answer. On the second question, most empirical findings have indicated that crime does matter, and that crime in central cities drives residents and jobs out. Cullen and Levitt (1999), for instance, find that increases in reported index crimes decrease the population of central cities, holding suburban crime rates constant; Grubb (1982) and Sampson and Wooldredge (1986) reach similar conclusions.

The effects of crime that these empirical studies find, however, are rather small. If you use the rough results of Cullen and Levitt, for instance, for 1999, and ask what would happen if the reported crime rate in cities with populations over a quarter million (6,557 per 100,000 inhabitants) fell to the reported crime rate in suburban counties (3,129), the answer is nothing earth-shattering—a 3.4 percent rise in population (FBI 2001). And no one knows of a policy that could produce such a big drop in central-city crime.

The major difficulty with these studies is that the measure of crime they use is probably a pretty poor indicator of the ways that crime affects people's decisions about where to live and work. Crime rates have several problems. First, they count all crimes equally, even though people worry a lot more about murder than about losing a calculator. Theft accounts for about 60 percent of the index crimes reported to police, burglary and motor vehicle theft for another 30 percent (since theft is the only index crime that rich people are more likely to be victims of than poor, it's not surprising that Cullen and Levitt (1999) and Grubb (1982) both find that rich people are more sensitive

to crime rates than poor). Reporting bias—people are more likely to report serious crimes to the police—mitigates this problem but doesn't eliminate it. Second, the threat of many crimes, particularly those involving intimates, may not depend much on where you live. The concentration on theft, burglary, and motor theft, however, alleviates this problem considerably (although it creates problems of its own). And finally, the *threat* of crime is what matters, not actual crime. A town where everybody goes home at five o'clock and sits in the window all night with a shotgun may have the same burglary rate as a town where everybody leaves doors and windows open all the time and visits one another frequently, but crime is a much bigger problem in the first town.

Since the crime rate doesn't accurately measure what we want to measure—the threat of crime that could be avoided by moving—the empirical studies probably understate the effect of crime, and so probably understate the benefits in sprawl reduction that crime reductions could cause. But it's hard to say how much.

III. Police

Because many benefits accrue from reducing the threat of crime, governments do many things to reduce that threat. In the next three sections I will survey some of these activities and review the evidence about how effective and how costly they are.

For taxpayers, the most costly governmental crime-reduction activity is policing. State and local governments spent more than $57 billion operating police departments in 2000 (U.S. Bureau of the Census 2003, table 443). This is not a huge amount, however: Americans spend about as much on police protection as they do on cigarettes—and they spend much more on private protective measures.

A. Why Police?

Why have police at all? Even today many small towns and rural areas have no police force, and no place had one before 1800, when London established the first police department (more precisely, when the East India Company's security force, which had been established a few years before, was made part of the government). People could hire their own security guards and bodyguards, take precautions, and offer rewards for the capture of lawbreakers. We see private security guards and private rewards today, and so we know such institutions are feasible. (Indeed, sworn police officers often work as private security guards, sometimes in uniform, either during their time off or after their retirement, and in the early nineteenth century New York City police officers often earned much of their pay in rewards from private individuals for solving

crimes.) There's always much to be said for private markets: they minimize shirking, force people to bear many of the costs of their decisions, encourage innovation, allow for competition.

But private policing is replete with externalities, both positive and negative. We've already discussed LoJack and The Club, both examples of private policing. A regime of pure private policing would add new externalities: my watchmen and Doberman pinschers would deter burglars from trying my next-door neighbor's house (and so be undersupplied in the private market) but divert them to houses two blocks away (and so be oversupplied). Since people who commit one crime are often likely to commit many more, if you lock up or exile the person who robbed you, you are helping many potential future victims.

Even more serious is the problem of force and coercion. If my bodyguards can beat up people who steal from me, or threaten to do so, then they can beat up people I don't like, too, or threaten to beat up people I'd like to get money from. Or they can beat me up or threaten me. Unless some ultimate authority has a monopoly on legally sanctioned force, contracts are worthless, Coasian bargaining is futile, and markets in security as well as everything else are impossible.

But the argument against anarchy is an argument for courts and for some body of agents to enforce the courts' decrees—a bunch of jailers and people who round up the suspects and criminals the courts order them to round up. This is not the type of full-blown police department we see today. Those who round up criminals could even be paid on a piece-rate basis, as are process servers and the bounty hunters that bail bondsmen employ.

The arguments for full-blown police departments rather than armies of bounty hunters are largely practical and apply mainly in densely populated areas with considerable transience and anonymity. To find someone who should be punished, it's often a bad strategy to wait until after a court has handed down a conviction; this gives the criminal plenty of time to learn he's in trouble and flee or hide. Placing alleged criminals under the court's control immediately, as the crime is committed or soon thereafter, is often a lot cheaper and more effective. That's what arrests are about. Those who make arrests need to be able to exert legal force at their own discretion, since the whole idea of an arrest is to act before a court does.

Courts also need information to operate, and information isn't always cheap. Once a crime has been committed, especially a crime of violence, it can't be undone, and neither the victim nor anyone else has an incentive to provide the court with accurate, expensive information. Hiring people to provide information is tough, because if you could tell for sure whether they were delivering good stuff, you wouldn't have needed to hire them in the first place. When the information they provide will be important in determining whether

the court punishes someone, the court is in essence delegating a part of its coercive power to the information gatherer. You would not want to pay the information gatherer based on whether the defendant is convicted.

Arresting and gathering information are often best done close to the scene of a crime, soon after its occurrence. Arrests should be based on information. The interaction between these functions implies that it will often be helpful to have the same person carry out both tasks, or at least people with a close organizational relationship—not two separate firms.

A third function works itself into the mix, too. Because of the externalities and indivisibilities of private security (I can't hire less than one security guard, if I hire any), it often makes sense for a group of neighbors to hire a security service together, or for a city government to do so on their behalf. Before New York City hired police, it hired watchmen, who walked around during the night and made noise if they spotted fires or burglars. There was no reason for these watchmen to have legal powers of coercion; even today, hundreds of thousands of private security guards do not have that. But walking or driving in areas where crimes might occur offers many good opportunities to make arrests and gather information, and having the power to make arrests and gather information for the courts makes you a ferocious and effective watchman. That's why we have police officers who are arresters, information gatherers, and watchmen all at the same time.

So along with tasks that private market-driven entities can often do well, police officers have legal coercive powers, and discretion about how to use them. In Chapter 8, in discussing regulation, I called such powers "police powers," and now we see why.

With police powers, I emphasized there, comes the danger of abuse. Abuse for police departments comes in two chief forms: corruption and police brutality, both of which are almost always in the news. In general, those who exercise police powers are both constrained by rules and subject to weak incentive systems. These rules and weak incentive systems are costly—we saw in Chapter 8, for instance, that the inability to "sell" zoning is behind the large problems it causes—but they may be better than the abuses that might prevail in their absence. Better a lazy despot than a cruel one.

B. Can More Police Reduce the Threat of Crime?

The easiest policy to think of for reducing the threat of crime is to hire more police. If the benefits from a smaller threat of crime exceed the costs of paying for more police, such a policy is a potential Pareto improvement.

But even the easy question of whether more police reduce crime doesn't have a definitive empirical answer—not to mention the more difficult question of whether any reduction in crime that police might cause would be worth the cost.

Criminologists, sociologists, and economists have conducted many studies of the relationship between police force size and the number of index crimes committed. Eck and Maguire (2000, p. 217) surveyed a large number of these studies and concluded, "Even when we examined the most rigorous studies, we could not find consistent evidence that increases in police strength produce decreases in violent crime." Some studies found that having more cops did reduce crime, but more showed the opposite or no effect at all.

Several recent studies have found a link between more police and less index crime, but two of them (Marvell and Moody 1996 and Corman and Mocan 2000) looked only at New York City, and found this relationship only for burglary. Levitt (2002) finds relationships of borderline statistical significance for aggregates of violent and property crimes. The issue is not settled.

The controversy over whether more police reduce crime resembles the controversy over whether smaller classes (that is, more teachers) improve learning. Common sense says one thing—more is better—but the data seem to be saying something else—more makes no difference. In both cases, statistical problems make finding a relationship hard, and so do the inevitable complex and offsetting reactions that human beings have to changes in their environment. (The big distinction between police officers and teachers is probably that in the right-wing view more police matter and more teachers don't, while in the left-wing view the reverse is true.)

Two statistical problems make it hard to conclude that having more police reduces crime. First, more crime causes cities to hire more police. If you look at a cross section of cities, high crime and large police forces go together—just as medical doctors are surrounded by sick people. Second, measuring the size of a police department is tough—at any moment, some officers on the roster are on sick leave or vacation or administrative suspension, and others are working overtime. Many cities have overlapping departments: in Newark, New Jersey, for instance, the Port Authority police, the New Jersey Transit police, the Rutgers University police, and the county police all patrol parts of the city, in addition to the Newark police. When you can't measure a variable well, its effects appear to be smaller than they actually are. (In the extreme, suppose you have a variable that predicts another variable very well—the number of eyes in some human population is usually a good predictor of the number of feet. But suppose you have a very poor measure of the number of eyes—the number of pairs of eyeglasses donated to the Optimist Club, for instance. Your measured variable will predict the number of feet poorly, although if you had a good measure of the number of eyes it would predict well.)

Because of these statistical problems, one should be skeptical of studies that find no relationship between police force strength and index crime. However, the theory that increased policing reduces index crime implies that only

criminals change their behavior when a police department expands. That's not the way the world works: when a police department gets bigger, many non-criminals act differently, and their reactions offset at least partially the reactions of criminals. For this reason, we have a right to be skeptical of studies that find that adding more police officers causes large reductions in crime.

Three reactions by noncriminals are relevant here. The first is the reaction by police departments. Police departments don't concentrate all their efforts on index crimes. Police officers issue parking tickets, sort matters out at traffic accidents, direct cars at busy intersections and when traffic lights are broken, and supervise parades and marches. They also pay attention to crimes without victims—not just those involving illicit drugs, but also gambling, prostitution, gun possession, loitering, and other violations of "public order." They make sure bars don't open on Sundays in violation of blue laws, and parties don't get too loud on Saturday nights. Only one person out of every six who was arrested in 1999 was charged with an index crime (Maguire and Pastore 2002, table 4.16). Bigger departments may just spend more time fighting prostitution.

Of course almost all of these other activities are valuable. Pedestrians and motorists gain when traffic flows better, and neighbors are better off if they aren't bothered by noisy parties. And controlling victimless crime has some benefits, too. Reducing index crime is not the only goal in life.

The second reason why we shouldn't be surprised at seeing no robust relationship between police strength and actual index crime is that having additional police officers around may encourage citizens to engage in riskier behavior—choosing dogs that are more cuddly and less vicious, using one lock instead of two, leaving the car windows cracked in a shadeless parking lot on a warm summer afternoon, walking alone at dusk, living for another month with an otherwise good man who doesn't hold his liquor well. Citizens may substitute police crime-fighting efforts for their own. (If this happens, it's also a good thing. The goal of police activity should be to reduce the threat of crime, not actual crime. The empirical studies are measuring the wrong variable.)

The final reason why bigger police forces may not reduce index crime is that additional officers may not be well deployed. The police brass may react to a bigger force by demanding less of themselves and of the officers they supervise; they may be more willing to play golf on Friday afternoons and less willing to engage in unpleasant confrontations with otherwise likeable officers who want to take their vacation time when it's inconvenient for the organization. Officers on the street, faced with fewer jobs to handle, may linger a little bit longer over coffee or chat a bit more with store owners they have come to know and like.

Police departments are not sleek profit-maximizing organizations, disciplined by both product and capital markets. They cannot be and should not

be, because they have discretionary coercive powers. That means they don't always use their resources optimally. Of course, to the extent that a bigger police force allows officers to lead happier and less stressful lives, and to treat their families and coworkers more humanely, greater organizational slack is also beneficial.

In summary, although the evidence should give no one confidence that expanding the average police department will automatically reduce index crime, it might have other benefits: in additional police service; in freer and less restricted lives for citizens; in better and less pressured lives for police officers. Whether these benefits, including the possible reduction in index crime, are worth the costs is likely to vary from case to case.

C. Can Policing Strategies Reduce the Threat of Crime?

What police actually do affects the amount of index crime. Activity matters, not necessarily numbers. The main reason for believing that police activity matters is that arrest rates—something that the police create—have repeatedly been found to affect index crime. More arrests, less crime. This relationship was first documented by Blumstein, Nagin, and Cohen (1978); Cameron (1988) surveys dozens of papers, almost all of which confirm the relationship; and Levitt (1995) shows that technical statistical issues do not weaken it.

Thus it's pretty certain that *some* things that the police can do work better than others at reducing index crime. Unfortunately, it's a whole lot less clear exactly what those things are. Sorting out the claims for different policing strategies didn't become any easier in the 1990s, when crime seemed to go down no matter what the police did.

1. Directed Patrolling

Some of the best claims can be made for "directed patrolling." This strategy concentrates police resources on small geographic areas, using detailed information on crimes and complaints to determine when and where to step up patrols. New York City's Compstat program, begun in 1994 and imitated in many cities since then, is probably the best known example of directed patrolling. As Eck and Maguire (2000, p. 230) describe it, "Under Compstat, police headquarters maintains statistical profiles for each precinct, including arrests, complaints, shooting incidents, and other information. Precinct commanders are expected to be vigilant about responding to shifting patterns of crime in their jurisdictions. They are held accountable through frequent debriefings at police headquarters, where they are 'grilled' about crime-reduction strategies and resource-allocation decisions." Compstat thus combines directed patrolling with greater effort by precinct commanders and those who report to them.

Although Compstat has never been rigorously evaluated, a number of di-

rected patrolling programs have been, and most have been found to work. For instance, one controlled experiment started with 110 "hot spots" in Minneapolis and randomly chose 55 of them for extra police attention. The result was a significant drop in reported crime in the hot spots chosen for intensive patrolling (Sherman and Weissbard 1995).

Even when directed patrolling isn't intentional, it can reduce crime in the areas where it happens to take place. After the bombing of a Buenos Aires Jewish center in 1994, the Argentine government assigned police protection to every Jewish and Muslim building in the country. Car thefts fell by 75 percent on the blocks where protected institutions were located (DiTella and Schargrodsky 2004).

Directed patrolling has two potential drawbacks, however. First, it's not clear that it actually reduces crime; it may only divert crime to less intensively patrolled locations. Empirical evidence on diversion is weak, but theory suggests that whatever diversion might occur will be substantially less than complete, meaning some amount of crime will actually be stopped. Hot spots are hot because they're excellent places to commit crime; making criminals move forces them to use less excellent locations. Worse locations mean crime is less lucrative (or more arduous or less pleasant), and so to some degree, however small, people are discouraged from engaging in it. Disruption itself also may be bad for criminal business. In addition, diversion will be less than complete if some of the people whose criminal activities would otherwise be diverted to other locations go to jail or prison, and no one else steps forward to take their place.

The second drawback to directed patrolling is more serious. Actual crime is the wrong thing for police to be focusing on, and the more intently they focus on it, the more likely they are to be diverted from the real goal, reducing the threat of crime, and from the other valuable activities that police engage in. If directed patrolling cleans out corner X, which most law-abiding people can easily avoid, but makes corners Y and Z more dangerous, and corners Y and Z are more difficult to avoid, the public will be worse off even though crime has gone down.

2. Problem-Oriented Policing

A strategy of problem-oriented policing comes closer to focusing on harms to society rather than on police record keeping. In problem-oriented policing, officers learn from citizens what crime-related difficulties are bothering them, and then try to relieve those difficulties. The police in Jersey City, for instance, in a well-studied example, found that people were worried about street fighting and robberies, and they figured out how to reduce both. Several other studies have also found that problem-oriented policing has managed to mitigate the problems it addressed (Eck and Maguire 2000, pp. 243–244).

The question an economist would ask about problem-oriented policing

is, How can the police actually find out what problems are most important to solve? Talk is cheap, and citizens are not asked how much they would bid to have a particular problem resolved. Police forces may find themselves concentrating on those problems that happen to have been described most eloquently and passionately, not the problems whose solution would be most beneficial. Perhaps, however, a little bit of poor information may be a better guide to action than no information at all.

3. Community Policing

The theory behind community policing is that police officers can fight crime better if the public gives them more assistance. This is, of course, true; I would do my job as a professor better, too, if disinterested volunteers routinely came by my office to grade papers for me. The real question is whether, on the margin, a dollar's worth of public crime-fighting effort is more effective than a dollar's worth of police crime-fighting effort. In many instances, you would think it would be—as, for instance, when all a member of the public has to do is report a tiny piece of information she picked up in the normal course of her life, or when a senior citizen who would stay home anyway most of the time keeps his eye on the street. These are the sorts of opportunities community policing has concentrated on. (See Akerlof and Yellen 1994 for a sophisticated explanation of why these opportunities might be present.)

Community policing encompasses a variety of activities, some of which have been rigorously evaluated. The results have been mixed. When all you care about is actual crime and you value citizen effort at zero cost, some community policing initiatives work sometimes and others don't. In one case, foot patrols (which encourage the community and police officers to talk to each other regularly) reduced female victimization rates in a public housing project. In the 1970s, cities with training programs in police-community relations experienced smaller increases in crime. In Oakland, California, and Birmingham, Alabama, areas where police officers made door-to-door contact with citizens had sizable declines in reported violent crimes. Home visits also seem to have worked in other cities.

On the other hand, carefully designed studies of foot patrolling found no effect on violent or property crime in Boston, Newark, and a third city, and even in the public housing project just mentioned, men were no less often victimized. Several studies of Neighborhood Watch programs show no effect on reported crime (Eck and Maguire 2000, p. 223). (David Anderson [1999, p. 624] estimates that citizens devote $655 million worth of time to block watches annually.)

Even with mixed results, many of these programs, especially foot patrols, are wildly popular with the public in high-crime neighborhoods, so it might be that they are reducing the threat of crime even if actual crime is unchanged. People may be leading better lives and taking more chances. Once

again, the evidence is meager and mixed. In Houston, storefront police stations and community-contact patrols made people feel that they were safer from personal crime. Results on public perceptions in Chicago were mixed, and no community-policing initiative appeared to reduce fear of crime in Newark.

4. Zero-Tolerance Policing

In some ways the opposite of community policing, the goal of zero-tolerance policing is to make bad people fear cops, instead of making good people love them. Zero-tolerance policing aims to reclaim the public spaces from criminals by making arrests for even minor offenses, such as approaching a vehicle in traffic to wash its windshield with a squeegee, littering, panhandling, prostitution, public intoxication, urinating in public, and vandalism. The punishment for these offenses also increases—instead getting reprimands and fines, people go to jail.

The theory behind zero-tolerance policing is sometimes called "broken windows," after the title of an influential article by Wilson and Kelling (1982). People commit crimes where they think they can get away with them, in neighborhoods where nobody cares enough or is tough enough to stop them. They don't figure out which neighborhoods these are by reading crime statistics; instead they look at how the public spaces are treated. If walls are covered with graffiti and drunks are lounging in doorways, then would-be criminals will conclude that no one cares, and they will feel safe to go about their business. Along with neglect, public disorder is another signal that serious crime is likely to go unpunished. Cracking down on minor offenses tells criminals the stakes have been raised.

Like directed patrolling, then, zero-tolerance policing is essentially aimed at diverting crime from a particular neighborhood. To translate this into a goal of reducing total crime, one must argue either that the diversion is less than complete, or that most neighborhoods can be made simultaneously less attractive to criminals. Both arguments are plausible. Like directed patrolling, too, zero-tolerance policing might also work by locking up the people who would otherwise commit serious crimes. For instance, it was found that many of the people picked up for fare beating in New York City (jumping subway turnstiles without paying) were wanted on more serious charges. Another theory that might support zero-tolerance policing focuses on the private efforts at protection and security that people take if they think the police won't protect them. Elijah Anderson argues that the failure of the police to demonstrate that they can and will protect law-abiding residents of poor minority neighborhoods has led many of those residents to adopt private vengeance and reputation as their means of protection, and to withhold cooperation from the police, since cooperation makes them less safe, not more safe. "Possession of respect—and the credible threat of vengeance—is highly valued for shielding

ordinary persons from the interpersonal violence of the street . . . The code of the street emerges where the influence of the police ends and personal responsibility for one's safety is felt to begin . . . In service to this ethic, repeated displays of 'nerve' or 'heart' build and reinforce a credible reputation for vengeance that works to deter aggression and disrespect" (E. Anderson 1999, p. 10).

Zero-tolerance policing can be a signal to these communities that the police will protect ordinary residents, and that those residents need not develop reputations for private vengeance but can cooperate with the police fruitfully and without fear of retribution. Zero-tolerance becomes a variety of community policing.

Some evidence exists on the effectiveness of zero-tolerance policing, but not much. Crime of all sorts fell dramatically when zero-tolerance policing was introduced in New York City in 1994, but many other initiatives (like Compstat) were introduced at the same time, and other cities (like Boston and San Diego) that employed totally different methods were just as successful in reducing crime. Corman and Mocan (2002), in a study of New York crime rates between 1974 and 1999, find that larger numbers of misdemeanor arrests reduce robbery and motor vehicle theft, and thus provide some support for a policy of zero tolerance. In two older studies, Reiss (1985) found a small decrease in robbery and rape in sections of Oakland where police stepped up arrests and citations for misdemeanors, while Sherman (1990) found no significant change in street robberies in Washington, D.C. But police aggressiveness did make citizens feel safer, in Sherman's data.

Just as no one knows the exact benefits of zero-tolerance policing, no one knows the costs, either. Some costs can be calculated: the amount of time cops must spend processing public-order arrests; the costs of food and medical care for misdemeanants detained in jail; and, especially, the value of the time that misdemeanants lose because of their arrest. Other costs are more uncertain, but could be large. Will citizens—minority group members in particular—come to see the police more and more as an occupying, harassing force and withhold cooperation from them? Will some people—young black men, say—come to believe that they are likely to be stopped, humiliated, and arrested whether or not they commit a crime, and so commit more crimes because it doesn't make a difference? Will police officers come to think of themselves as being at war with communities, reduce the restraint they exercise, and abuse their power more? Will juries continue to believe police officers? We don't know the answers to any of these questions yet.

5. Domestic Violence

While most attention has been focused on strategies for reducing crime by strangers, many major changes have also been made in the policing of domestic violence, and a lot of them seem to be successful. Before the 1970s, most

police departments treated "domestic disputes" as unfortunate incidents, not serious crimes. Policies were focused on family crisis intervention, counseling, and alcohol rehabilitation. Arrests were discouraged, even when there was probable cause, for fear of splitting families. A woman couldn't get a restraining order against a violent husband unless she filed for divorce at the same time. No one could get a restraining order on an emergency basis. And the restraining orders that were granted carried minor penalties for violations and were enforced haphazardly.

Since that time, domestic violence has been "criminalized." Restraining orders are easier to get and are more strictly enforced. Police now treat domestic violence incidents as serious crimes, like other acts of violence. Many departments mandate arrests when probable cause exists, even if the victim declines to press charges. Some police departments maintain special units in which officers develop expertise in handling domestic violence cases. Repeat offenders are sometimes sentenced to mandatory treatment and supervision.

Spousal murders decreased dramatically nationwide after the first steps to criminalize domestic violence were taken—dropping from 2,174 in 1976 to 836 in 1999. Declining consumption of alcohol, easier divorces, and higher employment among married women have contributed to this trend (the latter two factors allow women to escape from bad marriages with less pain), but the new police policies have probably helped.

D. The Future of Policing

In summary, not every new, old, or popular idea about what the police should do works, either to reduce crime or to reduce the threat of crime. But some do—directed patrolling probably works, and also aggressive efforts to talk with citizens, and treating domestic violence seriously. More good strategies are likely to develop in the near future. Why? Because policing depends on information technology, which has been improving quickly in recent decades. In the 1850s the telegraph made coordinated policing of large cities possible; in the 1920s the telephone changed how the police learned about crimes; and in the 1950s the radio-dispatched patrol car revolutionized how police responded to crime and how they patrolled.

Recent technologies are having a favorable impact on security today, as well. Some are privately deployed: LoJack automobile tracking devices, better locking systems in cars, and video surveillance cameras, for instance. Credit cards, debit cards, direct deposits, and ATMs have changed how people use cash. Public agencies now use computerized information systems, like Compstat, and DNA testing. One strong point of these technologies is that they seem to restrain the behavior of the police as much as behavior by the public; they have not created Big Brother. DNA testing has freed many innocent people convicted of crimes they didn't commit, and videotaping of traffic stops promises to improve the behavior of both police officers and motorists.

Many new technologies introduced into policing will fail; most new dot-com businesses fail, too. But some will succeed. In Chapter 2 I discussed one way in which information technology might be making city life more attractive rather than less (by making face-to-face interaction easier to arrange); the promise it offers in controlling crime (and controlling those who are supposed to control crime) is another.

IV. Incarceration

In almost all developed countries today, people who commit serious crimes are punished by incarceration. A few countries (most notably the United States and China) use capital punishment for a few very serious crimes, and sometimes for less serious crimes fines are levied or community service is required. (Another possible exception to this generalization for the United States is Megan's Law, under which sex offenders against children are branded for life and ostracized.) Mainly, though, criminals who get caught get locked up.

In the middle of 2000, about 2 million people in the United States were incarcerated: about 1.2 million in state prisons, 600,000 in local jails, and 100,000 each in federal prisons and in juvenile detention facilities. Most people convicted of index crimes go to state prisons. Local jails mainly hold people who are awaiting trial (including those awaiting trial for index crimes) or serving short sentences for minor crimes. Federal prisons hold those who violate federal laws. Most federal prisoners today are being held on drug charges.

The number of adults incarcerated in the United States grew quickly between 1980 and 2000—from about a half million (Maguire and Pastore 2002, table 6.1) to 1.9 million. Drug laws were an important part of this increase, but they don't tell the whole story. The number of people incarcerated for drug law violations was small in 1980 (about 6 percent of state prisoners and 25 percent of federal), and it grew quickly (to about 21 percent of state prisoners in 1998, and 57 percent of federal in 2000). Still, the major growth has been in the population incarcerated for violent crimes: in state prisons, it was four times as big in 2000 as it was in 1980. The number of people incarcerated for property crimes roughly doubled over this same time period.

Incarceration is an expensive endeavor. Government operating expenditures for corrections amounted to about $57 billion in the United States in 2001 (about the same as expenditures for police), and the cost of lost production due to incarceration was about $40 billion on top of that (using David Anderson's [1999] cost per prisoner-year and an updated number of prisoners).

These calculations of forgone wages, big as they are, may substantially understate the cost of imprisonment to prisoners. Beyond their lost income,

prisoners bear three additional costs. First, prison (though not necessarily jail) carries a stigma in the labor market. Freeman (1992) and Nagin and Waldfogel (1998, 1993) show that having been in prison reduces a man's wages and makes him more likely to be unemployed. Second, families of prisoners lose their support. Many of the children of women prisoners end up in foster care. Third, being in prison—being guarded, forced to comply with orders, separated from friends, and deprived of privacy, dignity, and heterosexual intimacy—is a whole lot worse than having a job where you don't get paid. Even without loss of income, most of us would be willing to pay a lot of money to avoid a year in prison.

Incarceration is a very expensive way to treat the problem of crime. But crime, we have seen, is a very expensive problem, and so expensive treatments may well be justified. That is the question we'll examine in this section.

A. A Short History of Punishment

Like police departments and indoor plumbing, prison is an invention that followed the Industrial Revolution. Before 1800, criminals were held in jail for a short period of time before trial and then after trial until their sentence could be carried out. Even famous "prisons" like the Bastille were essentially what we would now call jails. Sentences were usually simple and swiftly carried out: banishments (Australia and the state of Georgia were originally founded as penal colonies), brandings, floggings, lashings, and fines. Execution, though, was often the punishment of choice: capital offenses in New York City shortly after the Revolutionary War included housebreaking and malicious mischief (Burrows and Wallace 1999, p. 366). (Exceptions to these practices were political prisoners, who were often de facto hostages, and debtors, who were also hostages to their creditors.)

Sentences were motivated by two concerns: retribution and deterrence. **Retribution** is the idea that those who do wrong deserve to suffer and ought to suffer, no matter what the consequences of that suffering are. (For the proponents of retribution, the hardships of prison life may count as a benefit, or perhaps as nothing, but definitely not as a cost.) **Deterrence** is the idea that punishing those who do wrong makes doing wrong less attractive, and so encourages people to live better; they will obey the laws because they fear what will happen to them if they don't. Deterrence is concerned with the consequences of punishment; retribution is not. For these two purposes, fines, torture, and occasional executions work well. They're cheap for the government and awful for the criminals. Why spend more money on inflicting pain when you can spend less and inflict the same amount of pain?

The modern prison was invented when a new purpose became important—**rehabilitation.** Reformers in the early nineteenth century believed that holding prisoners in solitary confinement would make them "perceive the

wickedness and folly" of their ways, experience "bitter pangs of remorse," and prepare themselves for "future amendment" (Burrows and Wallace, p. 367, quoting Thomas Eddy). Eastern State Penitentiary in Philadelphia, which opened in 1829, was the first prison to pursue this new model. There each prisoner worked separately at a trade in his cell or an adjoining yard, and he never saw anyone except guards and the occasional visitor.

Once the idea of locking up criminals for long periods of time took hold, people realized that those criminals were not committing any crimes, and a new rationale for imprisonment was added—**incapacitation.** Criminals in prison can't commit crimes. The more criminals there are locked up, the fewer the criminals there are on the street, committing crimes.

Thus, four separate justifications for punishment have evolved over time: retribution, deterrence, rehabilitation, and incapacitation. The latter three all justify punishment by its consequences—by how it affects future crime. Only retribution justifies punishment based on past crime—which is what the legal system bases punishment on. With deterrence, past crime matters because potential offenders should think that criminals are being punished for their crimes; with rehabilitation and incapacitation, past crime is only a predictor, one among many, of future crime.

B. Why Incarcerate?

These four rationales make prison a plausible way to handle criminals, but they don't imply that it's the best way. We've already seen that fines, torture, and occasional executions are a much more cost-effective way of securing retribution and exercising deterrence. In this section, I'll argue that prisons and courts are lousy at rehabilitation and incapacitation, too. Then we will have to figure why we have prisons.

To see the weakness of prisons and courts as a means of rehabilitation and incapacitation, think about how these goals should be pursued. First, past crime should not be the chief determinant of who gets treated. Instead people should be selected to receive treatment on the basis of a great deal of detailed personal information—age, race, ethnicity, school record, parental responsibility, home atmosphere, religious orientation, psychological profile—as well as criminal record. Ideally, they would be identified and selected before they ever commit a crime; waiting until someone is robbed or murdered seems immoral and stupid.

Moreover, people who are incapacitated should be allowed to pursue their goals and their livelihood any way they can, so long as they don't come into contact with potential victims. Otherwise they are deprived needlessly of their noncriminal opportunities and the rest of the world is deprived of what they can produce. Thus efficient incapacitation is achieved in colonies or on islands—otherwise normal environments, except nobody can leave. It's not

achieved in prisons. Similarly efficient rehabilitation isn't realized in prisons either; people being rehabilitated should be otherwise free to pursue their life and livelihood.

The four goals of incarceration are not in conflict with each other. Retribution and deterrence can be served by inexpensive tortures and executions of those who are actually guilty of crimes; rehabilitation and incapacitation, by colonies and outposts for those who are likely to commit crimes. The two or three systems can be in place simultaneously.

But no modern developed country operates like this. Why are there prisons, if every goal that prisons are supposed to accomplish can be accomplished better and more cheaply some other way? The best answer harks back to the explanation of why police departments shouldn't be efficient. Those who exercise discretionary coercive power shouldn't be too highly motivated. As Friedman (1999) has argued, because prisons are costly both to the state and the prisoner, it's to no one's advantage to lock people up without good cause.

A government that imprisoned people for long periods of time just because it disliked them would find itself greatly impoverished. A state that tried to extort money from large numbers of people by threatening them with imprisonment would find its bluff called. A malicious person who wanted to extort money from his neighbor by threatening to send him to prison with false testimony would find himself confronting a government that had a strong motive to poke holes in his story, just to save itself money.

Because it is so stupid and expensive, prison gives the government a big stake in preserving the liberty of citizens. A short story by Larry Niven titled "The Jigsaw Man" (1971), brings this point home clearly. In the story, set at some point in the not-so-distant and not-so-fictionalized future, laws permit transplantation of organs from people who are executed. Now their deaths are no longer a waste. Organs are still in short supply, though, and the political system responds efficiently. The story is about a prisoner awaiting trial and execution. At the end of the story, we learn the capital charge for which the story's hero is being tried: "The state will prove that the said Warren Lewis Knowles did, in the space of two years, willfully drive through a total of six red traffic lights. During that same period the same Warren Knowles exceeded local speed limits no less than ten times, once by as much as fifteen miles an hour."[1]

C. Does Incarceration Reduce the Threat of Crime?

Prisons, for better or worse, are the institutions modern developed countries use for people who are convicted of crimes. Three of the four rationales for

1. I am grateful to Seth Weissman and to David Friedman (1999) for independently bringing this tale to my attention.

imprisonment justify imprisonment on the basis that it reduces future crime, and so social scientists have tried to figure out just how it does affect crime.

Almost all studies conclude that more prisoners mean less crime. Typically, the better studies find that a 10 percent increase in the number of prisoners in state or federal prisons causes a decrease of between 1 percent and 4 percent in the number of violent or property crimes (for a review, see Spelman 2000).

Asking how prison population affects crime, however, is not a question with direct policy relevance. Nobody decides how many people will be in prison. Prison population—like crime—is a result, not by itself a cause. It depends on how many crimes are committed, how many criminals are arrested and convicted, how long their sentences are, and when paroles are approved. Treating the size of prison population as a cause, as many of these studies do, has some serious logical problems.

Hence the results of these studies need to be interpreted carefully. They do not say that rounding up a thousand randomly selected grandmothers and sticking them in prison for a decade will cut crime significantly. Instead the results say that whatever usually causes a 10 percent increase in prison population will usually cut crime 1 percent to 4 percent.

The studies that find an effect for prison population also do not indicate whether that effect comes from incapacitation or deterrence (little evidence supports any rehabilitative effect of current prisons). The source has important policy implications. For instance, a "three strikes and you're in" policy—the third felony conviction results in life imprisonment—looks much more attractive if prisons work primarily through deterrence than if they work primarily through incapacitation. "Three strikes" will fill prisons with old men, many of whom will have severe medical problems, and since people over forty in general, and sick old men in particular, don't commit a lot of crimes, a lot of future resources will go to preventing few future crimes, if incapacitation is the main way prisons work. But if deterrence is the main way prisons work, the threat of a long prison sentence will deter many crimes now.

Ehrlich (1981), Levitt (1995), and Kessler and Levitt (1998) are the most sophisticated studies that have tried to figure out which effect is stronger. They all conclude that deterrence is more important than incapacitation, particularly for property crimes and aggravated assault. For robbery and murder, the results are weak or ambiguous, while incapacitation seems more important than deterrence for rape.

These results do not suffer the weakness of treating prison population as a cause, because they work from variables that are much closer to actual policies. Levitt's (1995) study is about the deterrent and incapacitating effects of arrest rates; Kessler and Levitt look at sentence enhancements. Notice that if deterrence effects are strong enough, a policy of longer sentences can theoretically reduce prison population by reducing crime. This paradox once again shows the folly of treating prison population as a cause, not a result.

D. Why Not Just Count the Crimes Averted?

The incapacitation effect that economists like Ehrlich and Levitt try to compute is different from the naive estimate that most people would first think of, and the difference is an important one. Economists try to figure out how many fewer crimes would actually occur if a particular criminal disappeared from the street, not (the naive view) how many crimes that one criminal would commit. To some extent, other criminals and potential criminals make up for the crimes a particular criminal would commit. This is called the **replacement effect.**

The replacement effect reminds us of one of the basic insights of economics: to understand the effects of a policy, you need to examine what happens when people fully adjust their behavior to it; you can't just look at the first-round effects and the immediate people involved. Just as the burden of a tax is not necessarily borne by the person whose name appears on a check, locking a person up doesn't necessarily erase the deeds that person would have done, and change nothing else.

For instance, if you were to lock up most of the employees of a particular fast-food restaurant, that restaurant wouldn't go out of business or even change its operations much. The reason is obvious: fast-food workers are easily replaced because many people have about the same talents, training, and temperament that the restaurant's current employees have. On the other hand, locking up a worker whose talents are hard to replicate—for example, an interior decorator whose taste hundreds of customers have come to trust—will force a firm to raise prices, lower quality, shrink, or disappear. The more replaceable a worker is, the greater the replacement effect when he or she is incapacitated.

The application to index crimes committed by enterprises that look like firms is obvious. A car-theft ring that loses the member who hot-wires vehicles will replace him with little difficulty, since the skills involved are easy to learn; its operations will be only slightly diminished. Losing a less easily replaced component—like its chop shop—would have a bigger impact on its operations. The same is true for drug-selling businesses.

What about crimes committed by lone operators? The conclusion is no different. To see why, it's helpful to use a graphic apparatus similar to that developed by Ehrlich (1981, 1996) and Cook (1986), called the "market for crime." They use regular supply-and-demand diagrams to analyze the interaction between offenders and victims.

First, think about the amount of a crime, such as burglary, committed in a particular area. The more burglaries, the greater the scale of precautions potential victims take—installing alarm systems, putting valuables in safe deposit boxes, making frequent trips to the ATM rather than holding lots of cash, staying home in the evening. (Bartel 1975, for instance, shows that com-

Figure 16.1 Victim
response schedule.
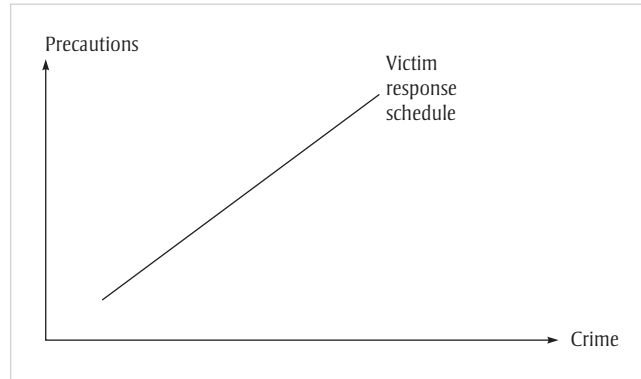
Figure 16.2 Of-
fender response
schedule.
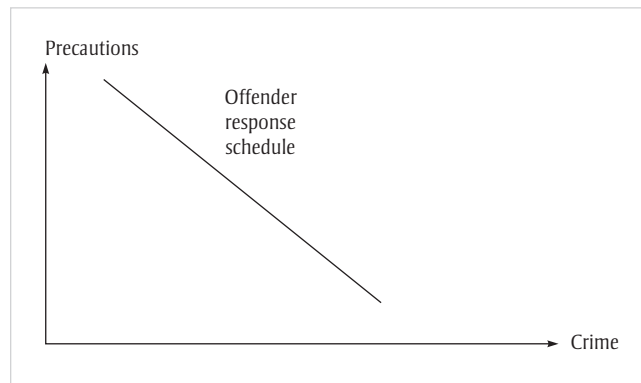
panies in neighborhoods with higher crime rates are more likely to hire secu-
rity guards.)

Figure 16.1 shows roughly how the amount of crime is related to the
scale of precautions. Call this the victim response schedule.

On the other hand, the greater the scale of precautions, the fewer the
burglaries. More precautions make burglary a more dangerous, less rewarding,
more arduous undertaking. Potential burglars turn their attention to relatively
more attractive pursuits, and actual burglars cut back on the number of bur-
glaries they commit. For offenders, then, the scale of precautions is related to
amount of crime as shown in Figure 16.2. Call this the offender response
schedule.

Over time, a community reaches a state where the offender and victim
responses are compatible with each other. That is, the scale of precautions that
victims take is just the scale that makes offenders commit the amount of
crime that induces victims to take that level of precautions. This is where the
victim and offender response schedules intersect, as in Figure 16.3. Call this
combination of crime and precautions the equilibrium.

Figure 16.3 Equilibrium crime and precautions.
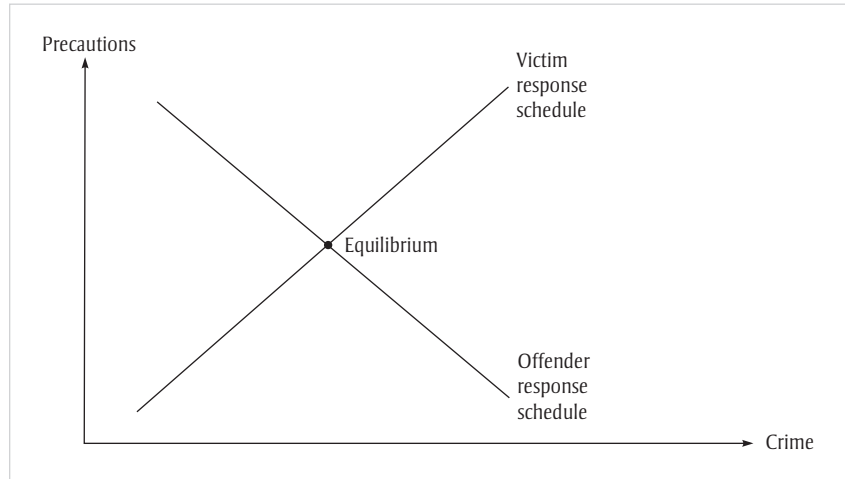
The effect of incapacitation depends primarily on what the offender response schedule looks like. If a lot of people, many more than the equilibrium number of burglars, have about the same talents and motivation for burglary, then small changes in precautions will in some small range call forth large changes in the number of burglaries. (A similar result occurs if individual burglaries do not become markedly more difficult for burglars to commit as they commit more of them.) The equilibrium number of burglaries will be somewhere in this small range. Incapacitating a few more burglars won't have much impact on the number of burglaries or the scale of precautions.

In terms of our diagram, large numbers of similar potential burglars imply a flat offender response schedule, as in Figure 16.4. Incapacitation removes a fixed number of burglaries at each level of precaution, and so shifts the offender response schedule inward. An inward shift of a flat schedule doesn't cause much to happen. A change in deterrence shifts the offender response schedule down, and so would be quite effective under these circumstances.

However, if there's a big difference between burglars and the rest of the world, and burglars aren't much dissuaded by precautions, then incapacitating burglars will have a big effect on equilibrium. Burglars in this case are hard to replace. Whether incapacitation translates into a decrease in crime or a decrease in precautions depends on the victim response schedule. But since potential victims will be better off if they're either taking fewer precautions or suffering fewer burglaries, the shape of the victim response schedule doesn't matter in assessing the social value of incapacitation. Figure 16.5 illustrates this case.

In general, if many people are fairly similar to the marginal person committing a particular crime, or criminals can increase their efforts fairly easily, then net incapacitation effects are not going to be big. People who go to jail

Figure 16.4 Effect of incapacitation with a flat offender response schedule.

Precautions
Victim response schedule
Old
New
Offender response schedule
Crime
Change in crime
Crimes that incarcerated people would have committed
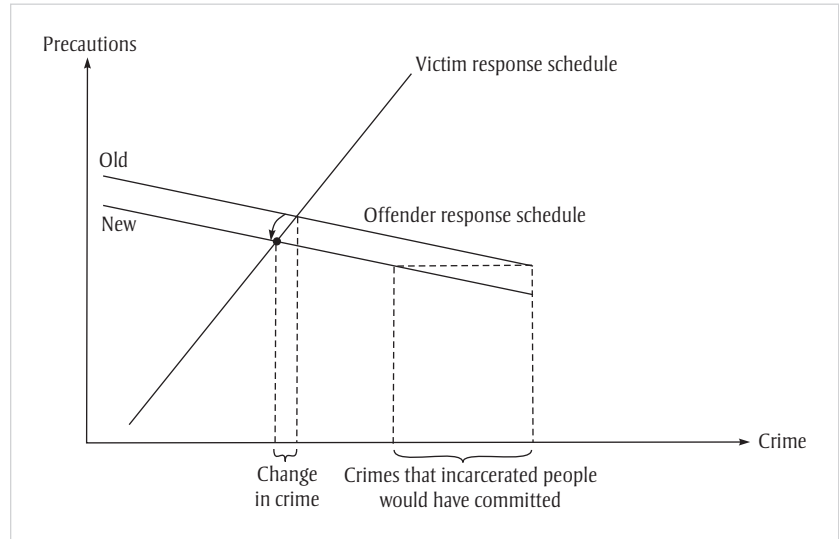
Figure 16.5 Effect of incapacitation with a steep offender response schedule.

Precautions
Old
Victim response schedule
New
Change in crime
Offender response schedule
Crime
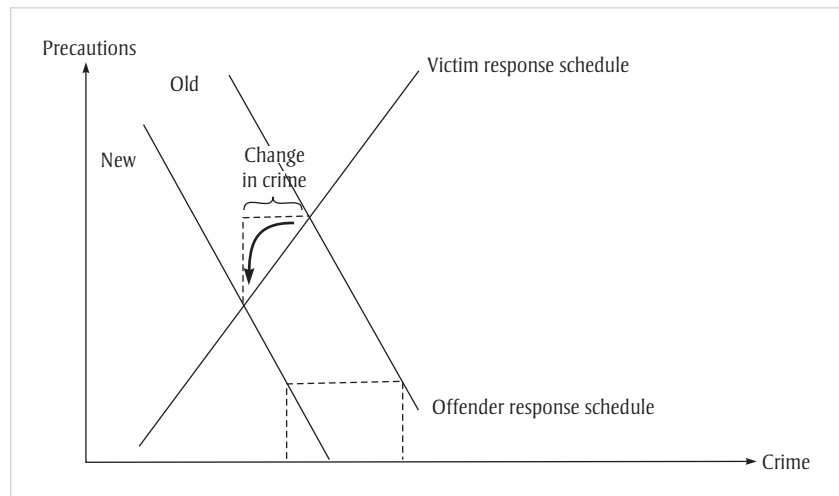
will be replaced. Where you would expect this to matter most is with crimes involving strangers and crimes for which significant skills and talents are not required. Replacement effects shouldn't be big for crimes involving intimates—like rape, frequently—or crimes taking idiosyncratic skills—like con games. But robberies and property crimes are likely to have significant replacement effects—which accords with Levitt's conclusions. Incapacitation in these cases may have little effect on actual crime or, much more important, on the threat of crime. (For a different view of incapacitation's effect, see Box 16A.)

> **Box 16A**
>
> ## Is Crime Contagious?
>
> ≡ An alternative story about how criminals interact might argue that criminal interaction increases the effect of incapacitation, rather than reducing it. Glaeser, Sacerdote, and Scheinkman (1996) develop models in which criminality is like a communicable disease. People who have it transmit it to others who are close to them. Thus if you remove a criminal from the community, he won't infect others, and crime will go down by more than the number of crimes he alone would have committed.
>
> But imprisonment as we know it doesn't remove a criminal from all communities, it just shifts him to a different one—a jail, prison, or juvenile detention facility. In these communities criminality may spread as much as in the communities their members originally came from. Empirically, Levitt (1995) finds the smallest incapacitation effects for just about the same crimes for which Glaeser, Sacerdote, and Scheinkman find the greatest social-interaction effects. This is consistent with the idea that jails, prisons, and juvenile detention facilities spread criminality rather than contain it.

E. Summary

Policies that expand prison population are one way of reducing the threat of crime, partly because they deter crime and partly because they incapacitate criminals, although the benefits of incapacitation may be exaggerated. It's expensive, though. The tough question is whether resources devoted to imprisonment do more to reduce the threat of crime than resources devoted, say, to policing or other crime-reduction activities—not just whether the current resources are worth it.

V. Better Alternatives to Crime

Police and prisons reduce the threat of crime by making criminals and potential criminals worse off—through incapacitation, by taking away the opportunity to commit crime (generally along with the opportunity for thousands of others of life's more innocent pleasures), and through deterrence, by making crime a less attractive opportunity. Another approach is to make doing something else so much more attractive that potential criminals voluntarily give up crime and are better off by doing so. If everybody is having so much fun doing something legal that they forget about committing crimes, the threat of crime disappears.

By definition, potential criminals and those who care about potential

criminals would be better off with such positive policies than with the nega-
tive policy of prisons and police (or "bricks and sticks," as it's sometimes more
poetically called). But the well-being of potential criminals (even though all of
us are potential criminals) is not the sole criterion on which policies should be
judged. It matters, but so do the well-being of taxpayers and well-being of
people threatened by crime. Once again, the question is which of these poli-
cies are potential Pareto improvements.

In this section, we'll distinguish between two types of positive programs.
Some make large segments of the population better off; others target a few
specific individuals. This distinction is important because they have different
displacement effects, and because they raise different ethical issues.

A. Prosperity

Generally, poorer people are more likely to commit index crimes than richer
people, and people are less inclined to commit most index crimes in good
macroeconomic times than they are in bad times. In 1993, 12 percent of male
high school dropouts age twenty-five to thirty-four were incarcerated, as op-
posed to 2.9 percent of the general male population of that age (Freeman
1996). Using compulsory attendance laws to check the direction of causality,
Lochner and Moretti (2004) find that completing high school reduces the
probability that a young adult will commit a crime, with the biggest effects
seen for African Americans.

Like prison, prosperity has both deterrence and incapacitation effects.
The deterrence story is that the more money you're making per year, the more
money you lose from a year in prison and the more severe the stigma penalty
of imprisonment is likely to be. The incapacitation effect is that the more time
you spend working and enjoying the money you've earned, the less time you
have available to engage in crime. To the extent that prosperity is general, re-
placement effects disappear as well—if everyone has less time, the marginal
potential criminal who would take your place needs a greater incentive to
commit crime, too. In this sense, incapacitation by prosperity should be more
effective than incapacitation by prison.

On the other hand, a lot of crimes, especially violent crimes, don't take a
lot of time to plan and commit. Even with property crimes, Grogger (2000)
found that almost all young men who had income from illegal activities also
worked, and on average worked as much as young men who didn't have in-
come from property crime. So prosperity, while it incapacitates everybody,
doesn't incapacitate anybody much.

Empirically, finding an effect of prosperity on actual crime has not al-
ways been easy. That's because even though rising wages make potential crimi-
nals less eager to commit crimes, in many cases (with robbery and property
crimes in particular) rising incomes of victims make them juicier targets.

Most recent studies, however, have found that prosperity reduces actual crime. Grogger (1998, 2000) for instance, finds that a 10 percent increase in the wages that young men receive causes a 10 percent decrease in the number of property crimes they commit (or admit to committing). Freeman (1996) summarizes a number of studies of how prosperity affects crime, and argues that the fall in the real wages that poorly educated men received during the 1970s and 1980s was a major contributor to increases in certain crimes during those decades.

By focusing on actual crimes, though, these studies understate the benefits of prosperity. To see why, suppose that only the wages of potential criminals increased, and potential targets became no richer and hence no more attractive. Then the threat of crime to potential victims would fall, and potential victims would be better off. Now increase the wealth of potential victims. Increasing their wealth (holding the income of potential criminals constant) makes them better off, even though it may make them more likely to be crime victims. If they feared the increased probability of victimization more than they appreciated the increased wealth, they could give the money to charity (or burn it). Thus potential victims are better off even if actual crime goes up. What's relevant is that the threat of crime would fall if they gave all their new-found wealth to charity.

What is the policy relevance of prosperity as a crime-fighting tool? No one is against prosperity; it brings many benefits beyond crime reduction. However, sometimes decisions have to be made about policies and programs that promote prosperity (or prosperity for certain segments of the population) but have other drawbacks. Expansionary macroeconomic policy, for instance, may induce inflation or contribute to a balance-of-payments deficit; an earned-income tax credit may reduce government revenue or force an increase in other taxes. For these decisions, the analysis in this section provides another reason for leaning toward prosperity.

The relationship between crime (property crime, at least) and wages also somewhat diminishes the long-run attraction of widespread imprisonment as an antidote to crime. Imprisonment reduces the wages of prisoners after they're released and makes them more likely to be jobless; low wages and joblessness, in turn, contribute to crime. So more prisoners today may cause more crime in the future, after they're released. The magnitude of this effect is unknown. Longer sentences for a few criminals may in the long run work better than shorter sentences for many.

B. Social Programs

Many programs try to help individuals live better lives, earn higher wages, handle their emotions and relationships better. Like prosperity, these programs have many benefits—chiefly for the people they target—that have absolutely nothing to do with crime. But they may also reduce the threat of crime.

Some of these programs work almost like prosperity—for instance, job training and school retention programs. Their goal is to make their participants more prosperous, with all the implications of that transformation for the propensity to commit crime. Other programs are like selective, nonpunitive incapacitation—midnight basketball, for instance. Some programs try to change their participants' goals and attitudes—counseling, faith-based programs, and rehabilitation, for example. People who value less what crime can achieve or value more what it can't achieve (a clean conscience, for instance) are less likely to commit crime. Preschool education programs and family counseling programs combine the first and third objectives—their goal is to make their participants, when they grow up, both more prosperous and more responsible. As with any large field of endeavor, some of the many social programs that might affect crime are more successful than others.

Unlike general prosperity, which lifts most boats, social programs try to lift only the boats their participants are in, and so are much more open to replacement effects. They resemble incapacitation, not deterrence.

Donohue and Siegelman (1998) examine a number of social programs whose effects on crime have been rigorously studied, and ask whether expanding any of these programs is likely to cause the same reduction in actual crime that an expansion of the prison system would cause, but at lower cost. They conclude (p. 1) that preschool enrichment programs coupled with family interventions would work: "Given precise targeting . . . cutting spending on prisons and using the savings to fund intensive pre-school education would reduce crime." A number of other programs, particularly the more intense and expensive programs, like Job Corps, come close to meeting this standard, while most of the cheaper programs, like regular job training and the type of Head Start program that operated in the 1980s, do not.

The standard that Donohue and Siegelman require social programs to reach, however, is too high.[2] Crime reduction is the only benefit of social programs that they count. In looking at potential Pareto improvements, you need to consider all benefits, not just those in the realm in which you're working (benefits outside that realm can be counted as negative costs). Thus expansions of some other social programs, not just intensive preschool intervention, may also be potential Pareto improvements.

Some of the implications of this conclusion appear to be troubling, however, or counter to some of our notions of fairness and equity. Consider gender: men commit almost all index crimes. In 1999, 74 percent of those arrested for index crimes were men (Maguire and Pastore 2002, table 4.8); leaving aside theft, the proportion was well over 80 percent. Placing a boy in an intensive preschool program or a young man in the Job Corps would therefore bring far greater expected benefits in crime reduction than placing a girl

---

2. This is not intended as a criticism of their work, since stacking the deck against the conclusion you eventually reach is a valid and effective rhetorical device.

or a young woman. Some programs that are potential Pareto improvements when all their participants are men are not potential Pareto improvements when some of their participants are women. Should we deny women places in these programs? Similarly for race: some programs that are potential Pareto improvements with all minority participants will not be if a substantial proportion of their participants are white.

Even without using suspect classifications like gender and race to exclude potential participants, something still seems discomforting in the way participants would be selected for crime-reducing social programs. Since nothing anyone does can reduce the probability that someone who will not commit a crime will commit one, social programs that figure out the most effective ways to exclude such people will generally produce the greatest social benefits. Social programs by definition help their participants, or try to, and so social programs aiming for potential Pareto improvements through crime prevention will help people who are likely to commit crimes, and ignore those who aren't. It sounds almost like extortion: "Help young males, because if you don't they'll rob you."

Remember, though, that no strong normative principles support creating potential Pareto improvements in themselves. Good ethical arguments support making actual Pareto improvements, and finding potential Pareto improvements is useful because they can point the way to actual Pareto improvements.

What are the actual Pareto improvements associated with problematic social programs? To be concrete, let's consider an extremely problematic social program: participants gain something from it, but what they gain is less than what it costs taxpayers to run the program. Suppose, also, that the program reduces crime among boys but not among girls (since women commit few crimes anyway, and you can't reduce crime below zero), and that the reduction in crime is big enough that operating the program for boys only is a potential Pareto improvement, while operating it for girls is not.

How can you transform operating the program for boys only into an actual Pareto improvement? By charging for it, and using the proceeds to reduce the cost that taxpayers bear. For instance, suppose you charge enough that taxpayers are no worse off after the program starts than they were before: they pay for the program, but the cost is offset by the reduced threat of crime they enjoy and the money raised from charging the male participants. The participants are better off than they would be without the program, too, even after they pay the fee for participation. (We know this because we have assumed that a male-only program is a potential Pareto improvement.) They still participate voluntarily. Taxpayers and participants are better off; nonparticipants are no worse off.

Thus charging for social programs can turn potential Pareto improvements into actual Pareto improvements. So should we charge for them? That

depends. When the class of people that benefits directly from the social program is not otherwise particularly deserving of redistribution—the class of males, for instance, in this example—charging for participation seems an eminently fair thing to do. But when the group is one that should be helped for some other reason—a minority group that has suffered maltreatment, for instance, or people who are poor because of misfortune—then one way of helping them would be, in effect, to waive the fee for participation. There are many good reasons to help people or to redistribute wealth (these are largely beyond the scope of this book), but fighting crime is not among them.

C. Abortion

One other way to reduce the threat of crime is to prevent the people who would be criminals from being born. This is the ultimate incapacitation strategy. Donohue and Levitt (2001) argue that when abortion was legalized in the early 1970s, many of the births that were averted were those of children who would have been unwanted. There's some evidence that unwanted children are more likely to grow up to be criminals than wanted children, and the hypothesis is plausible, too. Fewer unwanted children in the 1970s should have led to less crime in the 1990s, and Donohue and Levitt provide some empirical evidence that that's what happened. Joyce (2001), however, disputes this conclusion on many technical issues with an alternative empirical study.

VI. Conclusion

Index crime is a big problem, but it's not out of control. People are worse off in many ways because of the threat of crime. One of the ways they're worse off is that cities are smaller and less dense. The specter of crime also haunts race relations. Ways of reducing crime are very valuable.

We are learning more and more about which ways work. Some aggressive policing tactics work, and so do some social programs. (A lot of both fail, too, but that's a general condition of life.) Since the 1970s and the 1980s, the general trend for index crime has been downward, although not without interruptions. Information technology will probably help the forces fighting crime more than it will help criminals, at least for the traditional index crimes where propinquity still matters a lot.

Thinking about cities is impossible without thinking about crime, but thinking about crime in the United States today is also impossible without thinking about guns, alcohol, illicit drugs, and the public efforts to control or eradicate their use. These are the subjects of the next chapter.

Questions

1. Assess the argument that the federal government should subsidize terrorism insurance for major office buildings and stadiums because one building's precautions make another building a more likely target. What is the externality involved here? What is the theory of terrorist motivation? Of terrorist information? Why shouldn't the government tax precautions instead of subsidizing them? Should the government subsidize other forms of crime insurance as well? Should there be distinctions among the different kinds of antiterrorism measures a building's owners could take?

2. Given the idea of a "market for crime," for each of the innovations listed below indicate whether it shifts the victim response schedule or the offender response schedule, or both, and how it changes equilibrium crime and precautions:

   a. Higher probability of arrest.
   b. Cheaper, more effective burglar alarms.
   c. More publicity about threats of crime or terrorism.
   d. A reduction in moral scruples against community crime.
   e. More dead-end streets and more confusing street patterns.
   f. More potential criminals in the population.

3. Should off-duty police officers be permitted to engage in part-time security work in the community that they police? Should they be permitted to wear regular uniforms during this time? If they cannot do security work, should they be permitted to do nonsecurity work? Should their part-time jobs be subject to approval by police command?

4. You are advising your local police department about its policy on burglar alarms. Among the options that are being considered are:

   a. Allowing burglar alarms and responding to them as a high priority.
   b. Allowing burglar alarms and responding to them only occasionally.
   c. Allowing burglar alarms and responding to them only if the owners have never said anything unkind about the mayor or the police commissioner.
   d. Allowing only secret burglar alarms.
   e. Allowing only publicly advertised burglar alarms.
   f. Taxing burglar alarms and following one of the above response strategies.
   g. Setting up a number of different classes of burglar alarms, responding to each differently, and taxing each differently.
   h. Making burglar alarms illegal.
   i. Requiring a license to have a burglar alarm, fixing the number of licenses, and allowing licenses to be bought and sold.
   j. Anything else you can think of, including permutations of the above choices.

   What considerations should go into choosing a policy? What policy would you recommend? Why?

5. John Smith and Jane Doe like each other and would be happy to live together. Unfortunately, sometimes when Jane has too much to drink, she loses her temper and assaults John.

   John would be willing to pay $100 for each day he can live with Jane. John is also fond of Jane, even if they're not living together. If he can do something at no cost to himself that keeps Jane from being harmed, he will do it. (Hey, I just said they *like* each other; they're not passionately in love.) John is willing to pay $5,000 (including medical bills) to avoid one of Jane's violent assaults. Both John and Jane are risk neutral—that's part of the attraction. (That means that if confronted with choices between different gambles involving money, they always prefer the gamble with the highest expected payoff.)

   Jane likes John and would be willing to pay $100 a day to live with him. (But she would not be willing to give up any of her own money to prevent his loss of money.) When she's had a bad day at her job teaching urban economics, however, she is willing to pay $15 for a bottle of whiskey. Whiskey costs $10 a bottle. The probability that she will have a bad day teaching is always 3 percent. Whenever she drinks a bottle of whiskey, she assaults John. She knows this.

   Jane and John start off each morning with a clean slate. They have no memory of anything that happened in the past (although they know the probabilities of future events). Each morning looks the same to them. They decide each morning whether to live together that day.

   a. Suppose they can't exchange money in the morning or make binding promises. There is no tort law. Will they live together?

   b. Suppose the government places a tax of $6 a bottle on whiskey. How much revenue does the government collect (from this couple)? Does this tax cause a Pareto improvement? Explain why or why not.

   c. Suppose there is no tax on whiskey but there is a police department that responds to reports of domestic violence (Jane and John have thin walls and nosy neighbors) and punishes the assailant. Thus Jane knows when she decides whether to buy a bottle of whiskey that when she beats John up the police will come and fine her $10,000. (If she buys whiskey when she's not living with John, she feels violent but there's no one to pick on.) Relative to the situation in part (a), does the police policy cause a Pareto improvement?

   d. A new police chief is appointed, and she decides that victims should have more rights. She institutes new procedures for domestic violence. Instead of fining Jane on the spot, the police will issue a summons that must be paid the next morning when courts are open. The courts will also ask John whether he wants to press charges. If he doesn't press charges, Jane doesn't owe anything. What will happen in court? Jane and John both perfectly anticipate what will happen if they go to court the next morn-

ing. Will they live together? Relative to the old policy of the original police chief (question c), does the new victims' rights policy make John better off or worse off? Does it make Jane better off or worse off?

e. A third police chief takes over and decides to stop responding to domestic violence complaints altogether. Instead, he sets up an institution called "marriage," in which domestic partners can enter into long-term contracts with each other, exchange money, and compensate each other. But they can exchange money only at the beginning of the day (like a dowry paid at the start of a marriage), because promises between partners are still unenforceable. Will Jane and John take advantage of this new institution and live together? Relative to question a, is marriage a Pareto improvement? With marriage, how much domestic violence takes place? Relative to whiskey taxes (question b) or strict enforcement (question c), who is better off with marriage? Who is worse off? How much domestic violence occurs?

References

Akerlof, George, and Janet Yellen. 1994. "Gang Behavior, Law Enforcement, and Community Values." In H. Aaron, T. Mann, and T. Taylor, eds., *Values and Public Policy*. Washington, DC: Brookings Institution.

Anderson, David A. 1999. "The Aggregate Burden of Crime." *Journal of Law and Economics* 42:611–642.

Anderson, Elijah. 1999. *Code of the Street: Decency, Violence, and the Moral Life of the Inner City*. New York: W. W. Norton.

Ayres, Ian, and Steven D. Levitt. 1998. "Positive Externalities from Unobservable Victim Precaution: An Empirical Analysis of LoJack." *Quarterly Journal of Economics* 113(1):43–77.

Bartel, Ann. 1975. "An Analysis of Firm Demand for Protection against Crime." *Journal of Legal Studies* 4:443–478.

Blumstein, Alfred, Daniel Nagin, and Jacqueline Cohen, eds. 1978. *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*. Washington, DC: National Academy of Sciences.

Burrows, Edwin G. and Mike Wallace. 1999. *Gotham: A History of New York City to 1898*. New York: Oxford University Press.

Cameron, Samuel. 1988. "The Economics of Crime Deterrence: A Survey of Theory and Evidence." *Kyklos* 41:301–323.

Cook, Philip. 1986. "The Demand and Supply of Criminal Opportunities." *Crime and Justice* 7:1–27.

Corman, Hope, and Naci Mocan. 2000. "A Time-Series Analysis of Crime, Deterrence, and Drug Abuse in New York City." *American Economic Review* 90(3):584–604.

———. 2002. "Carrots, Sticks, and Broken Windows." National Bureau of Economic Research Working Paper 9061.

Cullen, Julie Berry, and Steven D. Levitt. 1999. "Crime, Urban Flight, and the Consequences for Cities." *Review of Economics and Statistics* 81:159–169.

DiTella, Rafael, and Ernesto Schargrodsky. 2004. "Do Police Reduce Crime? Estimates

Using the Allocation of Police Forces after a Terrorist Attack." *American Economic Review* 94(1):115–133.

Donohue, John J. III, and Steven D. Levitt. 2001. "The Impact of Legalized Abortion on Crime." *Quarterly Journal of Economics* 116:379–420.

Donohue, John J. III, and Peter Siegelman. 1998. "Allocating Resources among Prisons and Social Programs in the Battle against Crime." *Journal of Legal Studies* 27:1–43.

Eck, John E., and Edward R. Maguire. 2000. "Have Changes in Policing Reduced Violent Crime?" In Alfred Blumstein and Joel Wallman, eds., *The Crime Drop in America,* pp. 207–265. New York: Cambridge University Press.

Ehrlich, Isaac. 1981. "On the Usefulness of Controlling Individuals: An Economic Analysis of Rehabilitation, Incapacitation, and Deterrence." *American Economic Review* 71:307–322.

———. 1996. "Crime, Punishment, and the Market for Offenses." *Journal of Economic Perspectives* 10:43–68.

Federal Bureau of Investigation (FBI). 2001. "Crime in the United States: 2001." Accessed at *www.fbi.gov/ucr/01cius.htm.*

Freeman, Richard B. 1992. "Crime and the Employment of Disadvantaged Youth." In George Peterson and Wayne Vroman, eds., *Urban Labor Markets and Job Opportunities,* pp. 201–237. Washington, DC: Urban Institute.

———. 1996. "Why Do So Many Young American Men Commit Crimes and What Might We Do About It?" *Journal of Economic Perspectives* 10:25–42.

Friedman, David. 1999. "Why Not Hang Them All: The Virtues of Inefficient Punishment." *Journal of Political Economy* 107:S259–S269.

Glaeser, Edward L., and Bruce Sacerdote. 1999. "Why Is There More Crime in Cities?" *Journal of Political Economy* 107:S225–S258.

Glaeser, Edward L., Bruce Sacerdote, and Jose Scheinkman. 1996. "Crime and Social Interactions." *Quarterly Journal of Economics* 111:507–548.

Grogger, Jeffrey. 1998. "Market Wages and Youth Crime." *Journal of Labor Economics* 16:759–791.

———. 2000. "An Economic Model of Recent Trends in Violence." In Alfred Blumstein and Joel Wallner, eds., *The Crime Drop in America,* pp. 266–287. New York: Cambridge University Press.

Grubb, W. Norton. 1982. "The Flight to the Suburbs of Population and Employment, 1960–70." *Journal of Urban Economics* 11:348–367.

Joyce, Ted. 2001. "Did Legalized Abortion Lower Crime?" National Bureau of Economic Research Working Paper 8319.

Kessler, Daniel, and Steven D. Levitt. 1998. "Using Sentence Enhancements to Distinguish between Deterrence and Incapacitation." National Bureau of Economics Working Paper 6484.

Levitt, Steven D. 1995. "Why Do Increased Arrest Rates Appear to Reduce Crime: Deterrence, Incapacitation, or Measurement Error?" National Bureau of Economics Working Paper 5268.

———. 2002. "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime: Reply." *American Economic Review* 92(4):1244–1250.

Lochner, Lance, and Enrico Moretti. 2004. "The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports." *American Economic Review* 94(1):155–189.

Maguire, Kathleen, and Ann L. Pastore, eds. 2002. *Sourcebook of Criminal Justice Statistics* (online). Accessed at *www.albany.edu/sourcebook.*

Marvell, Thomas, and Carlisle Moody. 1996. "Specification Problems, Police Levels, and Crime Rates." *Criminology* 34(4):609–646.

Nagin, Daniel, and Joel Waldfogel. 1993. "The Effect of Conviction on Income through the Life Cycle." National Bureau of Economics Working Paper 4551.

———. 1995. "The Effects of Criminality and Conviction on the Labor Market Status of Young British Offenders." *International Review of Law and Economics* 15(1):109–126.

Niven, Larry. 1971. "The Jigsaw Man." In *All the Myriad Ways.* New York: Ballantine.

Reiss, Albert J. Jr. 1985. *Policing a City's Central District: The Oakland Story.* National Institute of Justice Research Report (April). Washington, DC: Government Printing Office.

Sampson, Robert J., and Jacqueline Cohen. 1988. "Deterrent Effects of Police on Crime: A Replication and Theoretical Extension." *Law and Society Review* 22:163–189.

Sampson, Robert J., and John D. Wooldredge. 1986. "Evidence That High Crime Rates Encourage Migration Away from Central Cities." *Sociology and Social Research* 90:310–314.

Sherman, Lawrence W. 1990. "Police Crackdowns: Initial and Residual Deterrence." In Michael Tonry and Norval Morris, eds., *Crime and Justice: A Review of Research,* vol. 12, pp. 1–48. Chicago: University of Chicago Press.

Sherman, Lawrence, and David Weissbard. 1995. "General Deterrent Effects of Police Patrol in Crime 'Hot Spots': A Randomized, Controlled Trial." *Justice Quarterly* 12:625–648.

Spelman, William. 2000. "The Limited Importance of Prison Expansion." In Alfred Blumstein and Joel Wallman, eds., *The Crime Drop in America,* pp. 97–129. New York: Cambridge University Press.

Wilson, James Q., and George I. Kelling. 1982. "Broken Windows: Police and Neighborhood Safety." *Atlantic Monthly* 249(March):29–38.

Wilson, Michael. 2004. "Kelly Reports a Sharp Decrease in Bank Robberies in the City." *New York Times,* March 26, p. B2.

U.S. Bureau of the Census. 2003. *Statistical Abstract 2003.* Washington, DC: Government Printing Office.

Chapter 17

Drugs, Guns, and Alcohol

$\mathcal{I}$*n cities, some people* use drugs, guns, and alcohol. Other people get upset about their use or possession. In late nineteenth-century Newark, for instance, one of the biggest political controversies was about drinking on Sundays: the Germans wanted to relax in beer gardens after church, but the Yankees thought alcohol profaned the Sabbath. Bringing a lot of people together in a small area forces the issue of how, if at all, these substances should be used and regulated.

One goal in regulating them has been to reduce crime. That's why I've grouped these substances together and separated them from other externality issues like traffic congestion and wastewater. Criminal justice agencies use vast resources in regulating these substances: in 1999, about 2.4 million arrests were made for violating drug, gun, or alcohol laws, almost twice as many as were made for index crimes. More prisoners are being held on drug and weapons charges than for property crimes (Maguire and Pastore 2002, tables 4.6, 6.38). You can't think seriously about crime in cities without thinking about how these substances are regulated.

I've also grouped drugs, guns, and alcohol together because governments often focus on them as an indirect way of regulating activities that might be regulated directly. Accidents and robberies, for instance, can be observed and regulated directly, but we also have laws about carrying concealed weapons and driving under the influence of alcohol, even when those activities cause no harm. Concentrating on the instrument rather than the harm has some real practical advantages, but also creates problems when technologies are changing rapidly.

The next section is a brief introduction to drugs, guns, and alcohol and the benefits they bring to their users. If nobody derived any benefits from

---

To get the most out of this chapter, you should be familiar with these concepts: externalities, incapacitation, Pigouvian taxes, and second-best. You can find these terms in the Glossary.

them, nobody would use them and we wouldn't need to talk about them. Sections II and III look at the social costs that untrammeled use of these substances might impose. If no one objected to untrammeled use, again we wouldn't have anything to say. My emphasis here, as in the popular literature, is on the connection between these substances and crime. Finally, section IV examines in more detail a few of the regulatory strategies available for handling the problems associated with these substances.

I. Private Benefits

People use drugs, guns, and alcohol for many different reasons. Drugs act on the blood supply to the brain, producing alterations of mood and perception. Ethyl alcohol, or ethanol, is a depressant; it slows down the brain and the nervous system. Many people enjoy the sensations it creates when the bloodstream bears it to the brain. It makes them relax and feel good, often slightly exhilarated; it makes them less concerned with minor irritations; it makes them forget, at least temporarily; it allows them to do or say things they normally wouldn't. In a high enough blood concentration, ethanol induces sleep.

Because it makes people more relaxed and less inhibited, alcohol is often used as a social lubricant—people get to know each other and become friends more easily than they otherwise would. To the extent that cities are built on social interaction, alcohol can help cities function better—no sense living in Manhattan if you can never meet anyone. People who drink moderately (and possibly also people who drink more heavily) earn more, on average, than abstainers, thus giving some credence to the notion of alcohol as a social lubricant. Barbiturates have similar effects.

Marijuana (or cannabis) is a mild sedative that produces a mild intoxication. It seems to increase the extent to which a person is open to external influences. Like ethanol, it has a loosening effect that can promote relaxation, sociability, talkativeness, hilarity, or introspective reflection. Thus social circles are built around marijuana, just as they are built around alcohol.

Heroin is a depressant that causes a euphoric rush. Used by doctors in the nineteenth century as a tranquilizer and a painkiller, it relieves anxieties, tensions, and general bad feelings. It doesn't relieve pain per se; it relieves fear and anxiety about pain. People who take heroin become drowsy and their limbs feel heavy.

By contrast, cocaine is a stimulant. It speeds up central nervous system activity rather than slowing it down. Freud used cocaine quite a bit between 1884 and 1887, and he reported that it turned his bad moods into cheerfulness, giving him the feeling of having dined well, "so that there is nothing one

need bother about," but without robbing him of energy for exercise or work. Cocaine combats depression and gives people more vitality and capacity for work. It decreases hunger, sharpens reactions, and gives people a feeling of great strength and increased mental capacity. Until 1903, it was one of the chief ingredients of Coca-Cola.

Amphetamines, including MDMA, or Ecstasy, and methamphetamine, are also stimulants; so is Ritalin when it's used by adults.

Guns allow people to cause (or threaten) severe physical damage, without exerting great strength or skill, and without coming extremely close to the target. Guns are also interesting in their own right. Guns are used for hunting, and for protection against both animals and humans. At least several hundred thousand times a year, people use guns to ward off attackers, usually just by brandishing them. Competitive shooting is a popular sport, and gun collecting is a popular hobby. Some people use guns for committing crimes, too.

People use drugs and guns for their jobs, too, not just for recreation. Businesses buy about 20 percent of the alcohol sold in the United States. Many jobs require that people carry guns. Stimulants, like caffeine, nicotine, amphetamines, and cocaine, help people stay awake longer and concentrate harder. Cocaine, in fact, was invented by a German army doctor in 1844 to help soldiers endure fatigue. The U.S. Air Force regularly prescribes amphetamines for its fighter pilots, to help them stay alert.

These substances also promote health sometimes. The protection that guns afford is obviously good for peoples' health. People who drink moderately (two or three of glasses of wine a day) are significantly less likely to suffer chronic heart disease than either abstainers or people who drink heavily. Moderate alcohol use may also reduce the threat of strokes. For several thousand years, cannabis has been used as an herbal remedy; it appears in Culpeper's *Complete Herbal* of 1653, for instance. Marijuana is a pain reliever, an appetite stimulator, and an antiemetic that's helpful in conjunction with chemotherapy. Heroin is a powerful antiemetic. Relieving pain and promoting sleep, which all the depressants do, are also generally good ways to promote health.

Thus alcohol, guns, and illicit drugs all provide their users with some wonderful benefits. Each of these substances also has drawbacks for its users, but that's not surprising: everything else that people do has a downside for the people who do it, too. If you go swimming, you could drown, and you definitely can't read a book while doing it. Eating chocolate tastes great, but it can make you put on weight. Using paper for writing on is convenient, but paper clutters up the house, it can catch fire, and you can get paper cuts from it. When people do these things, our presumption is that they take the downside into consideration; by itself, the downside is not an argument for regulation. The next section will present what are the arguments for regulation.

II. Crime

Alcohol, guns, and drugs are all regulated, and endless controversies swirl about the wisdom of tightening or loosening their regulations. In this section and the next, we'll look at several different kinds of reasons behind the regulations.

Because individuals have many good reasons to acquire these substances, we'll be looking for reasons to interfere in an otherwise untrammeled market. The problems that will concern us are those that would occur if the substances were unregulated, not those that actually occur now. The difference is significant. For instance, a great deal of violence now occurs among people involved in the business of illicit drugs, because violence is the only method they have for settling disputes. If drugs were legal, people in the business could resolve their disputes with lawyers and judges instead. The violence that accompanies drug selling is a result of regulation (prohibition, in this case), and so we won't pay attention to it here. The question for this section is what would go wrong if these substances were treated like cabbages or bookshelves.

A. Tools for Criminals

Criminals sometimes use guns to commit crimes, particularly murder. The majority of homicides are committed with guns, as well as 22 percent of robberies and 23 percent of aggravated assaults (U.S. Bureau of Justice Statistics, table 66). Holding everything else constant, having access to guns at home makes juveniles more likely to commit robbery, burglary, and theft (Mocan and Tekin 2003). Criminals also use alcohol and Rohypnol, the so-called date-rape drug, to reduce the defenses women exert against sexual attacks. Criminals use alcohol, barbiturates, and stimulants like cocaine and amphetamines to steel themselves to commit crimes, to make themselves feel stronger, more courageous, less squeamish morally or physically—and to give themselves excuses, too.

The harder it is for criminals to obtain these substances, then, the harder it is for them to commit crimes and the fewer crimes they'll try to commit. The threat of crime will fall. This argument is logically flawless, although few data are available to test the strength of these effects. Anything that makes crimes harder to commit reduces the threat of crime; hardly any crimes would be committed if everyone had to tie his left and right shoelaces together at all times, or if the atmosphere contained only 2 percent oxygen instead of 20 percent. The question here is whether the reduction in the threat of crime that the regulation of these substances causes is large enough to justify the losses that the regulation imposes on the consumers of these substances. The reduction in the threat of crime is difficult to measure because potential criminals can react to regulation many different ways. If guns are harder to obtain, for

instance, they can exert extra effort and get them anyway; substitute another weapon, like a knife or a box cutter; or substitute an easier victim, like an elderly woman, for a tougher one, like a gas station.

Making these substances more difficult to obtain only for those likely to commit crimes is an obvious solution to this problem: you might reduce crime a lot without forcing reductions in the beneficial uses of these substances. But this approach raises a new dilemma. If you define "criminals" narrowly, you won't cut crime much, because of the replacement effect or because you may miss some actual criminals, but if you define "criminals" broadly, you'll lose many beneficial uses. With guns, moreover, even restricting their purchase to people highly unlikely to use them for criminal purposes is not totally effective, since guns can be stolen. About a half million guns a year are reported stolen, and the actual total is probably considerably more.

B. Money as a Motive

For drugs and, to a lesser extent, for alcohol, regulation may reduce the motivation to commit crime by reducing the amount of money that users need in order to support their consumption. This is called the **economic compulsion** argument for regulation. In the current regulatory regime, some drug users commit a great deal of (nondrug) crime to get money to support their drug purchases. Making it difficult to get drugs or alcohol may hold down the amount of crime users commit for this reason.

Such a conclusion, however, rests on several premises, none of which is logically necessary or has strong empirical support. We can think of regulation (or prohibition) as raising the price to the user of drugs or alcohol. The question is what effect drug and alcohol prices have on the amount of income acquired illegitimately.

The chain of argument for regulation has four links (I'll use "drugs" instead of "drugs and alcohol" to save time and space):

1. Lower drug prices induce consumers to buy more drugs.
2. The increase in the quantity of drugs consumed is big enough that the total amount spent on drugs goes up, even though the price has gone down.
3. Because they spend more money on drugs, consumers must either spend less money on other goods or make more money, or both. For some reason, the entire increase in drug spending doesn't come from reduced spending on other goods.
4. To make more money, consumers rob or steal more, at least in part, rather than engage in more legitimate work.

Notice that if any of these four links fails, the whole chain fails—and regulation won't reduce crime.

Link 1 is almost surely true—demand curves slope down. You might think that such bourgeois concerns as price would make little difference in the realm of irrationality, adventure, abuse, and addiction, but there is abundant empirical evidence that price does matter. To take the most extreme example: Sarbaum, Polachek, and Spears (1998) describe a series of experiments with ethanol-addicted rats. To get more ethanol, the rats had to pay a "price" in terms of reduced availability of nonethanol food. Higher ethanol prices made the rats consume less ethanol.

Link 2 is more questionable. How responsive is the volume of drugs that people buy to the price they have to pay? For alcohol, where the data are best, most studies indicate that the quantity purchased is probably not responsive enough to make link 2 hold. Cook and Moore (2000), for instance, cite findings that a 10 percent decrease in the price of beer causes a 3.5 percent increase in the quantity consumed, a 10 percent decrease in the price of wine causes a 6.8 percent increase in quantity purchased, and a 10 percent decrease in the price of spirits causes a 9.8 percent increase in quantity consumed. Many other studies have similar results. For beer and wine, then, lower prices result in less money spent (for beer, for instance, 10-percent lower prices lead to a decrease of about 6.5 percent—10 percent minus 3.5 percent—in the amount spent), while for spirits, lower prices seem to have approximately no effect on the total amount spent.

Studies of opium in Indonesia during the Dutch colonial period (Van Ours 1995), heroin in Detroit in the early 1970s (Silverman and Spruill 1977), and marijuana among students at the University of California, Los Angeles, at the same time (Nisbit and Vakil 1972) found decreasing drug prices decreased the total amount spent on them, either strongly (heroin) or weakly (marijuana or opium). Some more recent studies (Saffer and Chaloupka 1995) of household survey data on several drugs come to the opposite conclusion, but they reach direct conclusions only on whether people use illicit drugs, not how much they use them.

In looking at the total response by all users and potential users, however, we may be looking at too broad a group. Links 3 and 4 show why. Link 3 says we are interested only in people who will reduce the total amount they spend on other goods by less than the increase in the amount they spend on drugs. Most likely, then, we are interested in heavy users and substance-dependent people, who are originally spending only a bare minimum on other goods and so are not likely to cut back on that spending much. Link 4 says we are interested mainly in people whose legitimate wages are rather low, since most people make more money per hour in legitimate work than they could in crime, once legal and incarceration costs are figured in.

Thus we should be primarily interested in the response to price changes by poor, substance-dependent users. A stock broker who snorts cocaine once a

month may or may not increase his expenditures on cocaine in response to a price change, but whatever he does, he's unlikely to change the rate at which he sticks up gas stations as a result. Links 3 and 4 imply that poor, substance-dependent users are the only ones whose criminal activity is likely to be affected by the amount they spend on drugs or alcohol, and so when we ask about link 2, these are the people we should concentrate on.

Does raising the price of drugs or alcohol induce poor, substance-dependent users to spend less on them? Your first instinct is to say, "Of course not; don't be silly. These unfortunate wretches are not in control of themselves. They will spend whatever it takes to maintain their habit." The alcoholic rat experiment, however, should be a warning against coming to a conclusion like this too swiftly.

There are two reasons why the quantity of drugs that substance-dependent users buy might be more responsive to price than your first instinct tells you it is. First and most important is the fact that, over a period of several years, the number of substance-dependent people can change. Higher prices make people less willing to experiment with drugs or alcohol, less eager to escalate their dosages, more reluctant to head down any road that might trap them in dependence, and more amenable to seeking treatment and sticking with it. Second, as the rat experiment suggested, prices do affect alcoholics and probably addicts, too. This has been shown in controlled experiments with fixed populations of alcoholics, and in the strong responsiveness of cirrhosis death rates to state taxes on alcohol (for a review, see Cook and Moore 1993, pp. 1642–1643, 1649–1650). (Cirrhosis is a liver disease usually brought on by heavy, long-term consumption of alcohol.) Chaloupka and Wechsler (1995) find that a 10 percent decrease in price creates 1.5 percent more binge drinkers in college.

Nevertheless, even with these two caveats, the first instinct on this question is probably, on net, the right one. Most of the evidence indicates that although consumption by substance-dependent users does respond to price, it probably doesn't respond strongly enough to make Link 2 hold. Most studies find that total consumption doesn't respond strongly enough for quantity changes to offset prices changes, and there isn't any evidence that consumption by the substance-dependent is more responsive than total consumption.

Thus for most users Links 3 and 4 don't hold, and while Links 3 and 4 may hold for substance-dependent users, Link 2 probably doesn't. So regulating drugs and alcohol probably doesn't reduce crime by reducing the amount users rob or steal to support their consumption.

Some evidence supports this conclusion. Silverman and Spruill (1977) found that higher heroin prices in Detroit in the early 1970s increased property crime but had no effect on nonproperty crime. Brown and Silverman (1980) found a similar result for New York City.

C. Making People Evil

Another way that these substances might increase crime is by unintentionally changing the personalities of people who use them, making those people more likely to commit crimes. (Intentional personality changes we've already considered under the rubric of "tools for criminals.")

Specific substances act different ways, and long-term effects are often different from short-term effects. In the short run, stimulants like cocaine and amphetamines may directly promote aggression, although violence may occur not because users are aggressive but because the drugs affect how they interpret social signals. On the other hand, some of the psychotherapeutics were explicitly formulated to reduce aggression under some circumstances, and so may make users less likely to commit violent crimes. (Ritalin reduces aggression in children but acts as a stimulant with adults, and so probably increases aggression among adults.) Heroin, OxyContin, and some of the other strong depressants may also cut crime through their pharmacological action. People who are sleeping or nodding out don't commit crimes. In the short run, marijuana is also a depressant and might reduce violence, but evidence on this point is mixed and meager.

Alcohol is a depressant, but there is some evidence that it causes violence through pharmacological changes on the central nervous system, even though it may sometimes suppress violence. As Cook and Moore describe it (2000, p. 1652):

> Under the influence of alcohol, a parent may be provoked to strike an irritating child, a college student may forcefully insist on having sex with his date, friends may escalate an argument into a bloody fight, . . . soccer fans may riot in response to an unsatisfactory game . . . Drinking may change the objective consequences of violence, since alcohol acts as an anesthetic as well as an excuse. It may . . . mak[e] people myopic and [narrow] their repertoire of responses to a tense situation. It may also cause self-management problems.

Long-term use of alcohol, marijuana, and amphetamines may also alter the nervous system in a way that makes people more likely to engage in violence. Long-term cocaine use can lead to paranoia or psychosis that can be accompanied by violence. (For more detail on the pharmacology of violence, see Fagan 1993 and Miczek et al. 1994.)

Guns can change the nature of disputes and lead to violence in two different ways. First, disputes that would otherwise be resolved by fisticuffs or shouting may be resolved by shooting if one or both of the angry, emotional parties has a gun. An assault can become a murder. A substantial portion of the rise in the murder rate in the late 1980s can be attributed through this mechanism to greater gun ownership by juveniles.

Second, guns can increase the number of disputes that come to a boil. The mechanism is described by Donohue and Levitt (1998). Start with the observation that fights, like wars, are almost always caused by mistakes. The losing party, if he knew he was going to lose, would capitulate immediately, agree to whatever the final resolution would eventually be, and save both himself and the winner from the exertion, pain, expense, and danger that fighting entails. Not fighting is Pareto superior to fighting, and so if people know who will win a fight, that fight will never happen.

When the only way to resolve a dispute is with pure, unaided physical force—punching, wrestling, kicking, and gouging—most of the time the loser of a prospective fight is obvious before it starts: it's the guy who's smaller, weaker, less physically fit. Small elderly men don't beat up pro football players. That's why those fights never happen.

But small elderly men can shoot pro football players. The pistol is an equalizer that "renders mere physical strength of no account, and enables the weak and delicate to successfully resist the attacks of the strong and brutal," as an 1875 gun industry pamphlet put it (cited in Bellesiles 2001, p. 28). With guns, the winner of a fight is not so easy to predict, and so the incentive to concede before the fight starts is much weaker. The availability of guns can cause disputes that would otherwise be settled peacefully (although not happily) to escalate into fighting. The disputants would both be better off if neither had a gun.

D. Making People Victims

The availability of these substances can affect actual crime by changing the behavior of victims (and potential victims), as well as the behavior of criminals (and potential criminals). Alcohol and heroin provide simple examples of this effect: sleeping drunks and nodding junkies are easy robbery targets. Voluntary use of date-rape drugs is similar; greater vulnerability can lead to more rapes.

Guns have the opposite effect. It's a bad idea to rob or assault or burglarize people with guns, and so having a gun reduces vulnerability. If guns are harder to obtain, more people will be vulnerable; if alcohol and heroin are harder to obtain, fewer people will be vulnerable. Everything else being equal, vulnerability will affect actual crime.

By itself, though, increased or decreased individual vulnerability is no reason to regulate these substances. Remember that the threat of crime is what matters for economic analyses, not actual crime. All sorts of unregulated activity affects vulnerability to crime: going outside makes you more vulnerable to robbery and burglary, concentrating on your work makes you more vulnerable to theft, going on dates makes you more vulnerable to date rape. The usual presumption is that when you engage in an activity, you take into con-

sideration the harm that might befall you; the benefits you get from doing it are the net of the dangers it exposes you to.

Regulation gets to be an issue when your degree of vulnerability affects other people, not just yourself. If robbers attack drunks rather than sober citizens, then sober citizens are better off if more people are drunk; the government should encourage intoxication. If the visible display of weaponry diverts burglars to other houses, the government should discourage visible displays. This is an issue we visited in Chapter 16.

The most controversial claim about how one person's protective measures affect other people's safety is that concealed weapons make everyone safer. This claim has been put forth strongly by John Lott (1998). Lott's argument is that concealed weapons work the way we saw that LoJack worked in Chapter 16. Criminals don't want to encounter a victim who is carrying a concealed weapon. They can't tell who's carrying a concealed weapon (if they could tell, we wouldn't say it was concealed); all they know is the rough probability that a potential victim has one. The higher this probability, the less attractive crime is and the smaller the threat every law-abiding person will confront. Since criminals estimate this probability from their experience and the experience of those they learn about on the news or through the grapevine, the more people carry concealed weapons, the higher they estimate this probability to be. Carrying a concealed weapon, then, creates a positive externality: it makes criminals more afraid of people in general, and so reduces the amount of crime they try to commit.

Notice that Lott's argument applies to cocaine and amphetamines, as well as concealed guns. To the extent that criminals don't want to attack someone high on stimulants and can't tell who is, the more people there are high on stimulants, the safer everyone is.

Although Lott's argument specifies a simple response by criminals to a higher probability of meeting a concealed weapon (they reduce the amount of crime they try to commit), there is an alternative. They might also react by escalation: increasing their armaments, becoming more aggressive, planning to shoot before they get shot at. In this case, unarmed law-abiding citizens would be made worse off by an increase in concealed weapons: they would be less likely to be attacked, but if they were, the outcome would be more horrible.

E. Incapacitation

Strict regulation of guns, alcohol, and drugs may reduce crime not because anything is bad about these substances, but simply because they give the criminal justice system a pretext for locking up large numbers of people who might otherwise be committing crimes. Suppose the police know roughly what sort of things a lot of people who commit crimes do (wear backwards baseball caps, or listen to hip-hop music, or carry marijuana), but they find it

difficult to link specific individual crimes with specific individual criminals. Then if they lock up a lot of people who wear backwards baseball caps and listen to hip-hop, "because" they carry marijuana, they will incapacitate a lot of criminals. They don't even have to incarcerate people who smoke marijuana but are otherwise "not the criminal type." Drug laws, or laws against carrying concealed firearms, may thus be an effective tool for policing, provided replacement effects are not too big.

Some evidence supports this reasoning. Kuziemko and Levitt (2001, abstract) find that "incarcerating drug offenders is . . . almost as effective in reducing violent and property crime as locking up other types of offenders."

It's effective, but costly. A policy of incarcerating drug users because they might commit index crimes has four kinds of costs. First are the direct costs to taxpayers of running prisons; second are the direct costs to prisoners of being in prison. We saw in Chapter 16 that these costs are likely to be large. Third is the cost of reduced deterrence: if people think they are going to be incarcerated whether or not they commit index crimes, the rational response is to commit index crimes whenever there is a personal advantage to doing so. The fourth possible kind of cost is cynicism and an accompanying reluctance on the part of some of the public to cooperate with a criminal justice system that they perceive as hypocritically punishing people for crimes they *might* commit, not anything they actually did. Perhaps fear of such cynicism is the reason why this perfectly plausible argument for drug and gun laws is so rarely set forth in explicit terms.

F. Empirical Evidence

We have identified a series of reasons why drugs, guns, and alcohol might affect the threat of crime, one way or the other. For some substances, though, the arguments conflict: heroin makes you drowsy but may give you an incentive to commit property crime; guns may get stolen and become a tool for criminals, but concealed weapons may deter criminals. To resolve these conflicts in theory requires data and empirical analysis.

Unfortunately, for the substances that generate the most heated debate, guns and illicit drugs, empirical analysis is difficult and the results at this time are not conclusive. Only for alcohol is there a consensus among researchers. The consensus is that alcohol is a major cause of violent crime.

Cook and Moore (1993), for instance, find that lower state beer taxes and greater per capita consumption of alcohol lead to more rape and robbery, and possibly more aggravated assault. There's no effect on murder. Chaloupka and Saffer's results (1992) are about the same as Cook and Moore's, but they find an effect on murder, and none on assault. Markowitz (2000b), in contrast, using slightly different techniques and data, finds that lower beer taxes increase assaults but have no effect on rapes and robberies. In international compari-

sons, Markowitz (2000a) finds lower alcohol prices leading to more robberies, rapes, and assaults. For more specific populations, Grossman and Markowitz conclude that lower alcohol excise taxes increase violence toward children (1999b) and violence and rowdyism among college students (1999a). Alcohol consumption also causes teens to get into more fights (Markowitz 2000c). Wasserman, Varnik, and Eklund (1994) document how the restrictive alcohol policies in the early years of perestroika (1984–1988) in the Soviet Union led to a massive drop in the violent death rate.

Comparable evidence about illicit drugs is scanty. Price and consumption information is obviously much harder to come by. Model (1992) finds that states that decriminalized marijuana had more property crimes and fewer violent crimes. In contrast, Markowitz (2000b) finds that decriminalizing led to more assaults and robberies. Chaloupka and Saffer (1992) also conclude that decriminalizing increases violent crime. Decriminalization may not be a good indicator of use, and Pacula and Kilmer (2003) find that marijuana use causes more property crime but no more violent crime.

For cocaine, Markowitz (2000b) found lower prices leading to more assaults and robberies, but saw no effect on fighting by teens (2000c). Corman and Mocan (1996) looked at variation in crime in New York City over the 1980s and concluded that more drug usage (probably cocaine) led to more robberies, burglaries, and motor vehicle thefts. De Simone (2001) also finds that lower cocaine prices led to more murders, rapes, and robberies. Mocan and Tekin (2003) also find that cocaine use leads to more crime. Cocaine thus appears to be associated with crime, particularly robbery, but we should be cautious in interpreting this result. Some of the crime associated with cocaine may arise from its regulation, not its use. Cocaine dealers, for instance, are attractive robbery targets because they have lots of cash, valuable goods, and no recourse to the police. The question we are asking in this section is whether, starting from no regulation, modest increases in regulation would cut crime. Markowitz and Corman and Mocan ask whether, starting from a highly regulated state, modest tightenings of prohibition would cut crime. The answers to these two questions do not have to be the same. Still, both pharmacology and empirical evidence point toward a tentative conclusion that cocaine can cause crime.

I'm not aware of any studies linking heroin and violent crime; nor do I know of any studies of crime of any kind focused on the newer drugs (amphetamines, barbiturates, hallucinogens). Pharmacology matters in this relationship, and so one would expect varied results.

Finally, for guns, Lott (1998) concentrates on "non-discretionary right-to-carry laws." These are state laws that mandate that a gun permit for a concealed weapon be issued to anyone who requests it and meets a few simple criteria—usually a clean criminal record, no demonstrable mental illness, and adulthood. He finds that these laws cut murder, rape, and aggravated assault

rates, have no discernible effect on robbery and burglary, and increase theft and motor vehicle theft. These laws are most successful in reducing crimes against women.

Notice that this result does not imply that concealed weapons, even for adults with clean criminal records and good mental health, should be unregulated. The case for regulating guns is that they impose external costs. Some of the crime reductions that Lott found are reductions experienced by gun owners themselves. This is part of the private benefit of guns, and is already part of the demand for guns. Lott doesn't show that you are less likely to be a crime victim if I buy a gun. Instead, he shows that, for certain violent crimes, the reduction in my probability of being a victim is greater than the increase, if any, in your probability. Lott's results argue against gun prohibition, not against gun regulation. On the question of regulation, they permit no conclusion one way or the other.

Lott's methods, moreover, have been seriously questioned, and his results are no longer generally accepted within the economics community. Duggan (2001) outlines two major problems with Lott's work. The first is a statistical problem dealing with the way to calculate confidence in the results and the way to time changes in nondiscretionary right-to-carry laws. Making the corrections Duggan identifies leads to the conclusion that these laws have no discernible effect on violent or property crime, and that in fact they're slightly more likely to increase almost all varieties of crime than to decrease them. He shows that nondiscretionary right-to-carry laws did not either increase gun ownership or make gun owners more willing to carry their weapons outside the house (if they had, crime would have gone down more in counties with more gun owners, and it didn't). Since the laws didn't make people do anything that would reduce crime, it's not surprising that crime didn't go down when they were enacted.

Second, Duggan shows that, in contrast to Lott's findings, more guns result in more murders and practically no other changes in crime. Rises in gun ownership precede rises in murders, so the relationship is likely to be causal; it doesn't appear to be a situation in which more murders are associated with more guns because murders cause fearful people to arm themselves. Even leaving aside murder, Duggan's results point strongly toward regulation: if gun owners are safer because they have guns, nonowners must be less safe.

While Lott and Duggan focus on the effects of gun availability among people who qualify under the nondiscretionary laws, police research has shown that stricter enforcement of gun permit laws also reduces murder, and possibly other violent crimes. Controlled experiments in Kansas City and Indianapolis during the 1990s showed that when the police blanketed small geographic areas and stopped and searched people for illegal guns, homicides and gun crimes decreased drastically, even though very few guns were actually confiscated (Eck and Maguire 2000, pp. 235–236). Simply the enhanced police

presence or expanded incapacitation, however, may have produced this effect; it didn't necessarily have anything to do with guns.

In summary, then, empirical evidence as it now stands supports strongly regulating alcohol for purposes of controlling crime. Weaker but still fairly compelling evidence supports some kinds of regulation of guns and cocaine. The evidence on marijuana is mixed. No evidence supports regulating heroin or other illicit drugs on the basis of crime effects. But the incapacitation effects of drug prohibition and gun control may also reduce crime.

III. Other Social Costs

A. Accidents

Guns, drugs, and alcohol are implicated in many accidents now, and if none of them was regulated, they would almost certainly be implicated in more. In a world of double-strict liability where no one was judgment-proof, these accidents would be no cause for government interference. You would buy a gun or a can of beer if and only if the pleasures and protection you expected it to give you were worth more to you than the expected harms it might do to you or to other people. But as we saw in Chapter 3, the actual system for dealing with motor vehicle accidents is very far from this ideal: most people are judgment-proof with respect to all but the most trivial accidents, insurance is not well calibrated to risk, lives are valued too cheaply, and no state uses double-strict liability. As a result, drivers have too little incentive to exercise care in all ways, including staying sober. Much the same is true for the personal liability laws under which gun accidents are adjudicated.

But do these substances really make accidents more likely? And if they do, would the excessive damage be greater if they were unregulated? As with crime, the best available evidence is about alcohol, and it indicates that alcohol-induced accidents are a serious problem and would be a worse problem if alcohol were unregulated. About 16,000 people died in 1999 in automobile crashes in which at least one driver had a substantial blood alcohol concentration; this was about the same as the number of people who died at the hands of murderers (U.S. Bureau of the Census 2002, table 1099). Some of these accidents would have happened anyway (almost all of the 42,000 people who died in traffic accidents were wearing underwear, but we don't attribute their deaths to underwear). More important are the results of a large body of research that shows repeatedly that higher beer prices reduce traffic fatalities generally (higher underwear prices wouldn't), and that higher minimum drinking ages reduce traffic fatalities among youths. (This literature is summarized in Cook and Moore 2000, pp. 1647–1649.) For instance, the results in Saffer and Grossman (1987) imply that eliminating state and federal taxes on

beer in the late 1970s would have raised traffic fatalities among those age fifteen to twenty-four by about 15 percent.

Gun accidents are not nearly so common as alcohol-related traffic accidents, under current regulations. There are about 1,000 deaths a year from gun accidents and about 130,000 injuries (Cook et al. 1999). (In contrast, more than 3 million people a year are injured in car accidents, a substantial fraction of which are alcohol-related.) Most of the gun deaths and injuries would not have occurred if no guns were available, but little is known about the relationship between accidents and regulation. Lott (1998, pp. 110–112) finds that nondiscretionary right-to-carry laws cause small increases in accidental shooting deaths, and so it is not implausible that, in an unregulated state, there would be many more deaths and injuries.

Even less is known about accidents connected with illicit drugs. Currently their numbers are few, in part because prohibition reduces drug use generally. To guess what would happen in a regulated state, the best we have to rely on is pharmacology.

Since barbiturates are similar pharmacologically to alcohol, we would expect unregulated barbiturates to cause accidents about the same way that unregulated alcohol would. Marijuana would also be similar to alcohol, although possibly slightly safer, since marijuana doesn't make drivers overestimate their abilities. Heroin and other opiates are probably the safest drugs of all, since users are rarely capable of driving while under their influence.

Stimulants like cocaine and amphetamines are more problematic. Police and highway authorities have long encouraged drivers to use stimulants—caffeine, in particular—as a safety measure. Drowsy drivers are a great danger, and stimulants may also improve reflexes and coordination. But stimulants may also push users over the line from awake and alert to aggressive and menacing. No one knows the net effect, especially when you have to adjust for other aspects of people's driving styles and practices, and conducting a large-scale experiment wouldn't be a good idea.

B. Noise and Rowdy Behavior

Some people in some settings react to alcohol by becoming noisy and rowdy. Alcohol is also a diuretic and sometimes induces vomiting. As a result, places where a lot of alcohol is consumed are often unpleasant. Even a single seriously inebriated person on a street corner or a park bench can be disconcerting and upsetting for passers-by.

Similar effects may be present for illicit drugs in some combinations, but prohibition makes it difficult to know for sure. Heroin and marijuana may induce somewhat more sedate behavior than alcohol, but secondhand smoke from marijuana is an environmental hazard, and the appearance in public of

people unconscious from heroin can be disconcerting. Cocaine and amphetamines probably make people noisy.

These externalities can probably be mitigated cheaply, however. Zoning and thick walls may well be sufficient in most cases.

C. Medical Costs

Aside from accidents, alcohol and illicit drugs also contribute to other health problems for their users. If everyone paid the full (marginal) cost of health care as they actually received it, then medical costs would be no reason for regulation; they would just be part of what people considered when they made the choice to use these substances. But most people don't pay all of their medical bills this way; insurance companies, charity, and governments pay a major portion of medical costs in this country, particularly for people with chronic diseases.

With alcohol, the major chronic medical problems arise from heavy use, including cirrhosis, heart disease, and some forms of cancer. Generally, moderate drinkers are healthier than abstainers, but this relationship may arise because abstainers include a portion of formerly heavy drinkers who had to quit for health reasons.

The chronic health effects of illicit drug use without prohibition aren't well known. They would probably be of the same order of magnitude as those of alcohol, however, and be concentrated among heavy users.

Many other substances we ingest—red meat and chocolate, for instance—also contribute to medical costs, but are not regulated.

D. Externalities within Families

Many of the costs of unregulated use of guns, alcohol, and illicit drugs would fall on the families of heavy users, particularly children. We have already seen that alcohol fans the flames of domestic violence (and cocaine would plausibly do so, too), and family members are especially likely to be victims of gun accidents. In this section we'll look at several other ways these substances might affect family members, children in particular.

Damage to children can start before they're born. Heavy alcohol use during pregnancy can cause fetal alcohol syndrome (FAS), which involves mental retardation, growth retardation, hyperactivity, organ abnormalities, and unusual head shape. FAS is one of the leading causes of mental retardation in Western countries (NIAAA, 2001). No link between moderate drinking and fetal damage has yet been established, but doctors generally advise pregnant women to abstain totally as a prudential matter.

Illicit drugs also cause in utero harm, but the extent to which they do so is unknown. Marijuana causes short gestations and low birth weight, both of

which are associated with a host of complications and developmental problems. Since marijuana is generally smoked, the mechanism seems to be similar to the mechanism through which the smoking of tobacco causes these same difficulties—a reduction in the blood's oxygen-carrying capacity (House of Commons, 2000).

In the 1980s, cocaine had been linked to a large number of serious perinatal problems, but subsequent research has shown that a substantial number of these concerns were unfounded. For instance, the wave of "crack babies" that concerned journalists at the time never materialized. Still, cocaine poses real problems: low birth weight (Kaestner, Joyce, and Wehbeh 1996), longer hospital stays at birth (Joyce et al. 1993), and fairly subtle problems in attention and alertness and in motor skills development (Zickler 1999).

After birth, children may also be less well supervised if their parents use alcohol or illicit drugs. Chatterji and Markowitz (2000) show that children of drug- or alcohol-using mothers are more likely to develop behavioral problems.

A major worry about the effect of alcohol and illicit drugs on families is that in the absence of regulation, many parents would spend all their money on personal inebriation and neglect their children. No direct empirical studies have been done on this question. Still, economic theory can shed some light on it.

There is a legitimate concern here. Parents can spend their money either on goods and services that primarily benefit their children (food and clothing for kids, toys, trips to the zoo, Barney videos) or on goods and services that primarily benefit themselves (drugs and alcohol, fancy restaurant dinners, golf club memberships, good books, Jacuzzis, comfortable chairs, bird-watching equipment). As I argued in Chapter 10 on education, parents don't get the full benefit of a dollar they spend on kids' goods (because some of them are somewhat selfish and because kids can't cut Coasian deals), but they get the full benefit of a dollar they spend on parents' goods. If the government were doing nothing else to deal with this problem, regulations that made parents' goods more expensive or harder to obtain would therefore be a potential Pareto improvement.

While this argument for regulation is a legitimate one, it's also weak, because it is easy to think of substantially less costly ways to increase expenditures on kids' goods—subsidizing them directly, for instance. Drugs and alcohol now represent a small portion of what adults spend on recreation, and would continue to represent a small fraction (probably even smaller) in a regime without regulation (see Box 17A). Making drugs and alcohol hard to get would probably cause only small increases in the amounts spent on kids' goods, because parents could substitute other parents' goods (just as new goods like the Internet, CNN and ESPN, and Club Med, which are great fun

for adults, have taken a share of adults' expenditures without causing massive child neglect). Regulating only specific adult goods, on the other hand, causes deadweight loss in allocating expenditures among adult goods.

Making drugs and alcohol harder to get would also affect adults who are not responsible for raising kids (the majority of adults).[1]

E. Productivity

The belief that drugs and alcohol (though not guns) will harm productivity is one of the oldest and most popular reasons for regulating them. In 1932, the argument for Prohibition, for instance, was phrased in these terms (Warburton 1932, pp. 195–196, cited in Cook and Moore 2000): "Prohibition, if it actually resulted in the cessation of use of alcoholic beverages, might be expected to affect the efficiency of industry in several ways . . . [A] decrease in the consumption of alcohol . . . is accompanied by greater skill at work . . . The elimination of drinking bouts should tend to eliminate absenteeism, especially on Monday, and irregularity in reporting at work. The impossibility of drinking to excess should result in less sickness and absence on account thereof, and in a longer average working life."

Although this argument contains a small germ of truth (which we'll explore later), it's mainly wrong. There are two problems. First, the factual premise is probably flawed. Various drugs have long been used to enhance productivity, cocaine and amphetamines especially. The American workday is built around the stimulative properties of caffeine, another drug. I have already noted the usefulness of alcohol in the workplace. Empirical evidence on the effect of alcohol use on wages is quite mixed, and evidence for other, illicit drugs doesn't exist.[2]

The second problem is philosophical. Productivity has no value in itself. If a person faces the right prices for her labor, and chooses to work less than eternally or chooses not to be a Stakhanovite, that's fine—that's what economists want to happen. Maximum output is the goal only of a slave economy. Televisions, sofas, the outdoors, books, sex, kids, sports—all of these reduce productivity even more than drugs and alcohol do, because they take time that people could be using to sell insurance policies or mine coal. An economy is doing well to the extent that it allows people to achieve their goals—to

1. About 69 percent of adults in 1999 were not responsible for raising kids. The population over age eighteen was 199.7 million (U.S. Bureau of the Census 2000, table 53). In that year 25 million married couples (table 67) and 1.5 million unmarried couples lived with children (table 57); that implies 53 million adults. In addition, 9.5 million children lived in households with a male or female householder and no spouse present (table 67), but this double-counts one adult in each unmarried couple. Thus about 61 million adults lived with kids.
2. Forty states ban welfare receipt by people who have been convicted of drug crimes; none bans welfare receipt by people convicted of murder. This is an interesting example of the power of the productivity argument.

Box 17A

What Do Adults Spend for Fun?

≡ The Office of National Drug Control Policy estimates that Americans spent about $65 billion on illicit drugs in 1998, primarily cocaine ($39 billion) and heroin ($12 billion). Reuter (2001) argues that $50 billion is a more plausible figure. Kleiman (1992) estimates that drug prices would fall by 80 percent if drugs were totally unregulated, and the elasticity estimates in Saffer and Chaloupka (1995), some of the highest around, imply that these price decreases would cut expenditures on heroin by 47 percent and on cocaine by 60 percent. Retail sales of liquor were $99 billion in 1996, but employers picked up the tab on about 20 percent of sales (Cook and Moore 2000). Removing regulation would probably cut expenditures by 15 to 20 percent. Thus in the absence of regulation, consumers would probably spend $100 billion to $150 billion a year on alcohol and illicit drugs. People who have child-rearing responsibilities probably spend less on alcohol and drugs than other similarly situated adults, and a sizable portion of this $100 billion to $150 billion would be spent by teenagers.

For comparison, here are just a few of the other activities on which adults spent their money around 1998.

	$ spent (billions)	Source
Full-service restaurants (not fast food)	$111.8	1275
Mailed charitable contributions	68.2	1289
Tobacco	57.3	1243
Fishing[a]	34.6	1275
Women's clothing store sales	34.8	427, 428
Magazines and newspapers	31.9	418
Lawn and garden	30.2	425
Wildlife watching	29.2	428
Jewelry store sales	21.6	1275
Beauty and barber shop sales	20.9	1301
Books and maps[b]	19.6	421
Hunting[a]	19.4	427, 428
Boating	19.1	428
State lotteries, net	14.1	518
Sound recordings[a]	11.5	420, 423
Athletic footwear[a]	9.7	426, 427
Total	$533.9	

Source: U.S. Bureau of the Census 2002, table numbers listed.
a. Excludes estimated proportion of participation by those under 18.
b. Excludes juvenile trade books; elementary, high school, and college textbooks.

watch television or fall in love, write poetry or get high—not to the extent that it thwarts them by forcing them to do somebody else's work. Falling in love is the goal of coal mining, not a diversion from it.

On the other hand, although the general argument about productivity is totally backward, a specific argument about income taxes has some validity. Income taxes depend, in part, on how long and how hard you work. Because they keep you from realizing the full value of your work, they distort the choice between working in the market and pursuing your own goals outside the market, where your efforts won't be taxed. One way of correcting this problem—of leveling the playing field between labor and leisure—is to tax commodities that people use in connection with their leisure time ("complements to leisure" is the technical term). This is a second-best approach to the problem. For example, suppose that the only thing that you want to do on your days off is go to NBA basketball games. If you pay 30 percent of the money you make as income taxes, and nothing for the time you take off, your choice between working and taking time off will be distorted in favor of time off. But if you have to pay a high enough tax on NBA tickets, your choice won't be distorted. Adding a tax on NBA tickets and lowering the rate of income tax appropriately can be a Pareto improvement: you're better off because you're not arranging your life around taxes anymore, and the government is collecting the same revenue.

If alcohol and illicit drugs are complements to leisure—they're something to do on your days off—taxing or regulating them would be similar to taxing NBA tickets. It would be one way of offsetting the distortions the income tax causes. But it would not be a good way of doing so. The reason is similar to the reason why taxing or regulating these substances would not be a good way of redressing the problem of underspending on children. Alcohol and illicit drugs are just one small way people spend their leisure time (one of the most important other ways, for instance, is raising children). Asking these goods to bear the burden of offsetting the distortions the income tax causes is like trying to eliminate mosquitoes from a region by selecting two ponds out of a thousand and filling them in, rather than spraying every one. It's not likely to work.

F. Self-control

Neoclassical economics mainly presumes that people act in their own self-interest, and so when somebody does something, it indicates that the benefits they derive from doing it are greater than the costs they bear. With drugs and guns, this presumption is open to doubt, however, so people might be better off if the government sometimes prevented them from doing what they wanted to do at a particular moment.

Why would people act against their own best interests? A number of different reasons have been advanced. One reason concerns how our brains are wired (Bernheim and Rangel 2002). Humans don't think seriously about every decision they make. If you're walking through the woods and see a snake in your path, you don't calculate the probability of being bitten, review the provisions of your medical insurance, and think about how you can reschedule your appointments if you're hospitalized. You just jump out of the way. Your nervous system makes this automatic decision, not the parts of your brain that handle detailed calculations.

Most of the time your automatic reactions are in your best interests. If your ancestors didn't have good reactions—fearing snakes and loving food, for instance—they would have perished and you would not be here to imitate them. But sometimes the automatic reactions make you do things that are not in your best interest. This occurs most frequently when you encounter stimuli that were not present with your ancestors on the African savannah tens of thousands of years ago—things, for instance, like drugs and guns. On these occasions you would be better off if somebody stopped you from doing what you wanted to do.

Sometimes automatic reactions develop with repetition. I look to the left first when I start to cross a street; I don't think about it (in fact, I had to draw myself a diagram to describe what I do). Addiction is another example of an automatic reaction that develops with repetition. Substances like nicotine, caffeine, ethanol, and heroin often cause physical changes in users' brains that make them crave more use (just as the sight of a snake makes most of us jump), even when they can calculate that such use is not good for them.

What does the possibility of the automatic-reaction system running amok mean for policy? Decisions made wrongly don't lead to benefits for their makers. An automatic reaction that is counter to the decision a person would make in a cool, contemplative, and calculating mood surely doesn't produce consumer surplus in the way an ordinary decision would. Government activities designed to restrain people from making bad decisions might lead to potential Pareto improvements.

But not every instance of an automatic reaction or even an addiction should be countered by government prohibition. Most automatic reactions are beneficial, and even addictions are harmful only to the extent that the underlying activity is harmful. If the chief consequence of current use is future use, and there's nothing bad about current use, there's nothing bad about future use either. Caffeine, for instance, is powerfully addictive, but since it's close to innocuous we don't regulate or prohibit it. To build a case for regulating something, you need to point to real harms (which I have done already); addiction by itself is not a harm.

G. Suicide

Averting suicide is probably a good idea, even if suicide appears to be a voluntary act. At the very least, suicide creates an externality—the grief and guilt that surviving relatives and friends feel. Many suicides are also mistakes—they result from judgments impaired by temporary stresses and chemical imbalances, either induced or natural. Some libertarians might not accept these arguments for government policies to avert suicide, but the case for such policies is reasonably strong.

Empirical evidence indicates links between suicide and both alcohol and guns. Little is known about how illicit drugs affect suicide. In the United States, the blood of suicide victims is often found to contain a high percentage of alcohol. High correlations have been found between alcohol consumption and suicide (based on suicides committed a short time after drinking) in Hungary (Skog and Elekes 1993) and in the Soviet Union (Wasserman, Varnik, and Eklund 1994). In the early years of perestroika, 1985 to 1988, Soviet Union president Mikhail Gorbachev initiated many restrictions on alcohol, and the male suicide rate fell by more than a third.

Markowitz and colleagues (2002), in a careful quantitative study, found that alcohol and illicit drugs caused suicidal thoughts and attempts among American college students (it wasn't just that suicidal students tended to drink). Chatterji and others (2003) concluded that binge drinking did not cause suicide attempts, but that clinical alcohol-use disorders did cause suicide attempts by young women.

While alcohol affects peoples' willingness to commit suicide, guns affect their ability to do so. Guns are a cheap, easy, and effective way to commit suicide; a majority of Americans who kill themselves use guns to do it. People who attempt suicide with a gun are more likely to be successful than those who attempt it by other common means (Sloan et al. 1990; Kellerman et al. 1992). People who decide to commit suicide at one moment may change their mind if they can't find a way of doing so or if they try and fail. The more easily available guns are—just lying around the house, for instance—the more likely it is that someone who wants to commit suicide will try and will succeed.

Would the people who used guns to kill themselves have found some other way to do it if the guns had not been around? Some evidence suggests that the answer is no: suicidal desires in many cases seem to be fleeting, and to go away if not acted on quickly. When Britain switched to using a less lethal form of gas for kitchen ovens, suicides by other means did not increase. Seiden (1978) investigated what had happened to people whose attempts to jump from the Golden Gate Bridge had been foiled. After more than two decades, over 90 percent of them were either still living or had died a natural death.

No rigorous studies have yet established a causal relationship between

guns and suicide (good data on gun ownership is a serious stumbling block here), but since gun suicides are more numerous than homicides (of all types), it's possible that the gains from some form of gun regulation could be significant.

IV. Regulation

Getting run over by a drunk driver is bad, but no worse than getting run over by a sober one. Better a thousand drunks on the road with no accidents than one sober driver who hits a pedestrian. Similarly, murders committed with pickaxes are just as bad as murders with guns, loud parties with Coca-Cola are just as bad as loud parties with marijuana or alcohol, and parents who neglect their children so they can spend their money on antique teacups are just as bad as those who neglect their children so they can buy crack. In an ideal regulatory system, therefore, each pair of equally bad actions would be treated the same way.

We will begin our analysis of how to regulate alcohol, guns, and drugs by looking at how such an ideal system would work. This ideal system is not the system we actually have, nor is it one that's likely or even possible to be set up, but it's best to understand what sorts of outcomes are desirable before tackling the problems of how to achieve them. Next we will look at the difficulties that would prevent such an ideal system from being implemented. Finally, we will address technology, because in this area we find the greatest difference between an ideal system and what we have now, and the least serious impediments to creating an ideal system.

A. Ideal Regulation

Ideal regulation is simple: you pay for the harm you cause, not for the weapon you use or the substance you ingest. Then (if you don't make any mistakes) you'll take all those actions for which the expected benefits to you are greater than the expected harm they cause to other people. Nothing new here.

Some of the benefits of an ideal regulatory system, relative to what we have now, are obvious. The most obvious are the benefits consumers can get from drugs, alcohol, and guns—the benefits I described in section I. Other benefits are more subtle. For instance, for the drugs that are now illicit, quality and consumer protection would improve. For ordinary, licit goods, quality is maintained by governmental inspection, by consumers' right to sue, and by vendors' concern for reputation. For illicit drugs, the first two mechanisms are missing, and the third, always weak to begin with, is impaired by police efforts to prevent sellers from establishing themselves in normal fashion. If people were punished for what they did, not what they ingested, the quality of illicit

drugs would improve, and become more consistent and reliable. This would reduce medical problems.

Another subtle benefit from an ideal regulatory system would come from a reduction in the violence that now accompanies the drug-selling business. Ordinary businesspeople selling a lawful product can rely on courts and lawyers to settle disputes; they can use checks and accept credit cards; and when they are robbed, they can go to the police. Businesspeople who sell drugs can do none of these things. Thus they must cultivate an image of toughness, hire hit men instead of lawyers, and handle large amounts of cash, which makes them attractive targets for robbers. Goldstein and colleagues (1989) found that almost all "drug-related" murders in New York City were related to the business of drugs, not the use of drugs. Miron (1999) argues that almost all the variation in U.S. murder rates over the twentieth century is due to changes in the population's age structure and changes in the vigor with which the government has pursued its wars against drugs and alcohol—the more vigorous the pursuit, the more difficult it becomes for sellers to act like normal businesspeople, and the more vicious the business must become.

Health would also improve under an ideal regulatory scheme, from a reduction in cases of HIV and hepatitis. Hypodermic needles shared by users who inject drugs are a major means of transmission of both these diseases. People inject heroin and other illicit drugs because the price for the drugs is so high and their purity is so low that more normal means of ingestion don't provide an adequate high. People share needles because most states, in an effort to discourage illegal drug use, have banned the sale of needles except by prescription, and as a result, needles are expensive.

Legal drug-selling businesses would also be far less disruptive to neighborhoods than illicit ones. Other businesses that sell small, valuable recreational items—jewels and cameras, for instance—are considerably less of a neighborhood nuisance. Because they are legal, they can insure their inventory and use visible security devices. They also don't employ children and give them guns. Legal drug-selling businesses would be governed by zoning laws and could enter into enforceable contracts with their neighbors.

In the long run, probably the major benefit that would come from an ideal regulatory regime is that it would encourage technological innovation and channel it in socially desirable directions. This holds true for all the substances discussed in this chapter. Drugs, guns, and alcohol are all technological products dating from the nineteenth century or before (from biblical times, essentially, for ethanol and marijuana) that have stagnated in the past century, while we have seen amazing advances in other closely related technologies—in psychotropic drugs, for instance, and military weapons. If users were fully responsible for the harm that they caused when using these substances, they would have strong incentives to seek technologies that produced less harm,

and researchers would have strong incentives to invent those technologies. They would also have patent protection.

B. Problems with Ideal Regulation

Unfortunately, an ideal regulatory regime is probably impossible to implement. While some problems might be fixed easily, other problems are larger and more fundamental. Even the fundamental problems, however, may well be susceptible to mitigation in ways more socially constructive than those being used now.

The simplest problem with an ideal regulatory system is that it doesn't produce the incapacitation benefits that drug prohibition and strict gun-control laws produce. There is no pretext for arresting and incarcerating hundreds of thousands of people who cannot otherwise be convicted of committing a crime, but who are, police and prosecutors believe (maybe correctly), highly likely to commit index crimes in the near future. This problem—if it is a problem—can be straightforwardly solved by publicly acknowledging the benefits of such incapacitation and proceeding with it, but compensating the people who are shipped away. No one should suffer just for being a member of a group that is likely to commit crimes. Of course, if compensated incapacitation is too expensive for governments, relative to the benefits in decreased threat of crime that it produces, then uncompensated incapacitation is too expensive, too (the losses that would be compensated are real losses that enter into a sound cost-benefit analysis, whether they are compensated or not).

The second problem is harm from mistakes. The ideal system relies on people to act in their own best interest, and sometimes they don't. One response to this problem is to deny that it's a problem by arguing for the traditional position: that in the long run, at least, however people act is the best approximation that they can make to acting in their own best interest, at least so far as other people can tell.

Another response is to encourage people to look for devices that make it hard for them to make mistakes—to help them substitute their own better judgment in a more calculating and contemplative mood for their automatic reactions in the heat of the moment. Most of us use alarm clocks to help override our automatic reaction of sleeping later than we want to, and doctors use anesthetics so patients don't scream and recoil when painful procedures are being performed. Some people keep potato chips and candy out of their house to keep from succumbing to temptation, and others join Weight Watchers, Smokenders, and Alcoholics Anonymous. Chemicals that block the effects or reduce the enjoyment of nicotine, alcohol, and heroin are also available.

These mechanisms let people be paternalistic with themselves. They block only those decisions that the individuals themselves think are mistakes

when they are able to be calculating and contemplative. Governments should probably actively promote these mechanisms.

The most general problem with ideal regulation is that the right penalties may be too high because the damage done is too great. The "right" penalty, for instance, for a driver (drunk or not) who kills someone is probably about $5 million, assuming that all such drivers are apprehended. If only half of the drivers are apprehended, then the penalty should be about $10 million. But such huge penalties are not going to be imposed—people are judgment-proof, and juries are not going to award huge damages just because malefactors are usually difficult to detect, or sentence defendants to long prison sentences. Even if such long sentences were handed down, they would still be costly to governments.

This problem can be mitigated by shifting the responsibility to parties who can't escape it. This is essentially the strategy followed fairly well with legal prescription drugs, where pharmacists, physicians, and pharmaceutical companies take responsibility for the harms that might arise from mistakes. The strategy has worked fairly poorly with cars, where automobile insurance companies take responsibility for the harms that drivers cause. The difference with prescription drugs is that pharmacists and physicians have much more control over the drug taking of their patients than insurance companies have over the driving of their insured clients. In neither case does the strategy work perfectly, but it allows a modicum of protection for future selves without destroying the opportunity for current selves to enjoy cars and prescription drugs. The liability that bartenders assume when they serve inebriated patrons is another example of how this mitigation process can operate. (Notice that prohibition works in the opposite direction by making it impossible for a third party to take responsibility for anything.)

With recreational and occupational drugs, liability could be shifted to manufacturers, for instance, although they usually have little control over how, or how much, their products are used. Guns could be insured like cars. Psychoactive drugs (including ethanol) could be handled through prescriptions, by pharmacists and physicians or by people in special new occupations that don't have any other medical responsibilities. Or perhaps establishments like bars or hotels could be set up as venues for drug taking (leave your car keys and credit cards at the front desk), thus reducing the nuisance externalities involved. This would also be a way of finding third parties that could be held responsible for the harms drug users cause.

But there's no guarantee that we will always be able to find or manufacture appropriate third parties for every conceivable weapon and every conceivable drug. When we cannot, clearly other measures are in order: Pigouvian taxes, selective prohibitions (but with compensation to the class of people harmed), general limitations on use, and subsidies for substitutes. The types

of analysis we did in Chapters 8 and 9 can be applied fairly straightforwardly here.

C. Technology

The major shortcoming of drug regulation in the United States is that nobody knows what's permitted and what's not. Of course everybody knows that a few specific drugs are totally prohibited, and a few are available only with a prescription. The Drug Enforcement Administration (DEA) maintains five lists of drugs—called schedules—and has different rules for each of the lists. The drugs on Schedule I—including cocaine, heroin, marijuana, MDMA, other methamphetamines, psilocybin (the hallucinogenic chemical in mushrooms), PCP, and peyote—are totally banned. Schedule II drugs, like methadone, opium, morphine, Percodan, and Ritalin, are available only under rigidly controlled circumstances. Drugs on the three higher-numbered schedules are treated more loosely but are still subject to numerous restrictions.

But the vast majority of existing and potential drugs are not on any schedule. MDMA, for instance, although it was invented in 1912, was not placed on a schedule until 1985; during most of the intervening time the DEA paid it no attention. None of the drugs that are under development or have yet to be invented—which are probably most of the drugs we'll be using twenty years from now—are on any schedule.

The rules under which drugs are assigned to schedules are therefore far more important in the long run than the specifics of what's on what schedule now. Assigning drugs to schedules is the DEA's job, and the criteria it works with are established by statute. The problem is that these criteria are at best vague and at worst wrong.

For listing on Schedule I, the strictly prohibited drugs, the two most important criteria are "potential for abuse" and "no currently accepted medical use." The medical-use criterion is clear enough on its face, although probably not good public policy. Everything has "potential for abuse"—I could try to stick chairs into my ears—and so the DEA has tried to clarify what this means. It has done this by stating that a drug has potential for abuse if any one of a large number of specified conditions holds, or if some unspecified conditions hold. Three of the most important conditions that would indicate potential for abuse are (U.S. Drug Enforcement Administration 2001, p. 2):

1. There is evidence that individuals are taking the drug or other substance in amounts sufficient to create a hazard to their health or to the safety of other individuals or to their community; or
2. There is significant diversion of the drug . . . from legitimate channels; or
3. Individuals are taking the drug . . . on their own initiative rather than

> on the basis of medical advice from a practitioner licensed by law to administer such drugs.

The most favorable interpretation of these conditions is that something has potential for abuse if a lot of people like it a lot and it has some possible bad consequences. This reason for banning substances is neither clear nor good public policy.

Alcohol regulation, in contrast, is clear. But it's very restrictive: there's no possibility of replacing ethanol with something better and then being allowed to market the product.

Together, this ethanol requirement and the DEA's vagueness are strong disincentives to investment in the research needed to produce healthier, safer, less physically addictive alternatives to alcohol and the illicit drugs that are currently popular. People crash their cars, abuse their spouses, rob gas stations, and neglect their families quite frequently, even when they subsist on marmalade and lettuce. If all it would take to get your product banned is a few people who do some of these things after ingesting it, or a few people who like it so much they change how they live, then you would be extremely foolish to invest in the research needed to develop it. Recreational and occupational drugs are not illegal per se; they just have to be perfect (or unpopular) or they will be banned. And you can never prove a drug is perfect.

This lack of research on alternatives to alcohol and illicit drugs appears all the more astonishing and tragic when you think about how cities approached the externality problems they faced in the nineteenth century. We saw in Chapter 9 that one of the most frequent and successful government responses was to subsidize the provision of a less noxious substitute for a hard-to-police activity that caused negative externalities. Garbage collection was an example: it's hard to detect and punish people who dump their trash by the side of the road, and so densely populated cities provide free trash collection as a substitute. One of the strongest lessons of the twentieth century is that drug and alcohol use is hard to police, and so we would expect governments to be *subsidizing* alternatives, not effectively discouraging them.

And the substitutes would not have to be chemical. People use alcohol and illicit drugs to change their mood and to achieve altered states of consciousness. Chemicals aren't the only way to do this—physical exercise, spiritual meditation, sex, diet, recreation, bathing, massage, travel, art, entertainment, rituals, and games are alternatives; skiing, fast driving, and other sports also do the job. Children enjoy the mood-altering and consciousness-altering experiences that cradles, merry-go-rounds, swings, and teeter-totters offer (Brecher 1972, chapter 66). The same reasoning that brought us free garbage pickup—a great public health policy—should also bring us subsidized roller coasters, video games, bungee jumping, and super-scary movies.

Something like this policy is evident in the current mild subsidies for

drug and alcohol rehabilitation, and in the even more tepid support for methadone maintenance programs to treat heroin addiction. But these subsidies are targeted to people who have become substance dependent, and thus they act as an incentive, however mild, to substance dependence. It is as if you had to be convicted of illegal dumping before the city would pick up your garbage. And neither methadone nor rehab is much fun.

Even if an all-out subsidy program to produce thrilling roller coasters and wonderful recreational drugs is not in order—perhaps because the income tax is enough of a disincentive to work already—there doesn't seem to be any reason to discourage the search for better chemical means of mood altering and consciousness changing. Even though the nonchemical proposals sound alluring, the most likely substitutes for chemicals are other chemicals.

Here's where the vague, contradictory, and generally impossible-to-meet DEA standards take their toll. The hard question in drug and alcohol policy is this: How much harm can a drug whose benefits are only recreational and occupational do before it's placed on Schedule I or Schedule II? What kind of harm? What proof of harm must be offered? Remember, we spend most of our income on goods for which the benefits are only recreational and occupational (cars and almost all houses, for instance) and which also cause significant harms. So the answer should not be zero harm. With good ancillary policies, like Pigouvian taxation and third-party responsibility, the answer can be quite a lot of harm. But what is most important is that there be a clear, enforceable answer, so that research can proceed toward known objectives. Otherwise we'll be stuck with lousy drugs like ethanol and marijuana forever.

Could a similar policy work with guns, too, to guide technological change toward finding less dangerous ways of meeting users' goals? Probably, but to a lesser extent. Because guns are legal, gun makers have not faced the same regulatory constraints on improving their products that drug makers have; the main problem has been that they and their customers have inadequate incentives to reduce external harms. Moreover, guns, unlike drugs, are long-lived assets; a gun you buy today will last you for many decades. That means that the effects of technological progress will show up much more slowly.

Still, technology can probably play an important role. One way we already subsidize substitutes for guns is through free provision of police services. Nonlethal methods of personal protection—devices that stun or paralyze or cause hives or diarrhea or mark their victims with indelible ink or make them smell like ammonia—are also substitutes for guns that should be subsidized, following the garbage-pickup argument. Alternatives to traditional hunting should also be sought and subsidized. Perhaps hunters could try to implant tiny, personalized radio transmitters into animals instead of killing them; then hunters would have proof of their skill and could even compete with one another.

V. Conclusion

Guns, alcohol, and drugs aren't really all that different from cars, water, garbage, houses, and the other goods we've focused on elsewhere in this book. They bring many benefits to their users, and they can also cause great harm. When people live in close proximity to each other, ways have to be found to extract as much benefit as possible while reducing the amount of harm. This is as true for cars as it is for cocaine.

Of course, if the same people who regulate alcohol and drugs regulated cars, they would have banned all cars except Model T Fords produced in 1938. Then they would have required that any new vehicle either be an ambulance or meet vague but incredibly high standards of passenger safety and pollution control so tough that only carriages pulled by constipated horses could qualify. And they would be opposed by a group of legalization advocates incessantly singing the praises of 1923 Studebakers and insisting they should be allowed back on the roads because they were better than Model T Fords.

On the other hand, if cars were treated like guns, we would be witnessing debates between those who thought traffic lights violated their constitutional rights and impeded citizens who were trying to evade or capture criminals, and their opponents, who wanted to restrict driving to ambulances, police cars, and funeral processions.

Fortunately, cities have not approached car regulation in this fashion—or water or sewers, garbage or housing. Mistakes have been made, to be sure, but the debate has not degenerated into competing nostalgias, and the door has been left open to technological progress. The history of cities really does have lessons to teach.

Questions

1. The safety of roller coasters, especially the newer generation of them, has recently been the subject of controversy. Since 1979, more than a dozen people have sustained serious brain injuries on roller coasters, and one elderly patient has died. One medical study, however, found that even the most spectacular roller coasters exerted no greater G-force on the brain than "plopping into an easy chair," far less than the force needed to tear brain tissue or cause swelling. Critics of this study note that its conclusions apply only to normal, healthy individuals, and that it says nothing about roller coaster safety for people with preexisting medical conditions.

Should roller coasters be permitted to operate? Why or why not? How should they be regulated? What liability rules should apply?

2. Suppose you are working for a congressional committee rewriting the laws on recreational and occupational drugs. Outline the conditions that a recreational or occupational drug should meet before its sale is permitted. Why are these the right conditions? What standards of proof should be used to determine whether or not the drug meets these conditions? What restrictions on its sale should apply? What standards of liability should be connected with recreational and occupational drugs?

3. "Buy and bust" is the traditional tactic that police departments use to arrest drug dealers—an undercover police officer, posing as a user, buys drugs and then signals for the dealer's arrest. An alternative tactic, much less used, is "sell and bust"—an undercover police officer, posing as a dealer, sells drugs and then signals for the buyer's arrest. Compare the two tactics. Who bears the burden of the enforcement activity? Which tactic is more effective in reducing drug use? Be explicit in your model of the markets for drugs and drug dealers.

4. In Chapter 8 I discussed the conditions under which rules worked better than taxes in regulating an activity with external costs. Drug consumption and gun ownership are both activities with external costs. Use the criteria in Chapter 8 to determine whether taxes or rules are the better instrument for regulating each of them.

5. Mandatory car insurance has often been suggested as a model for how guns could be regulated. We saw in Chapter 3, however, that car insurance as currently constituted leaves much to be desired. If mandatory insurance were to be required for guns, how should it operate? What would be the advantages of a policy like this? The disadvantages?

References

Bellesiles, Michael A. 2001. "The Regulation of Firearms: An Historical Overview." Mimeo, Department of History, Emory University.

Bernheim, B. Douglas, and Antonio Rangel. 2002. "Addiction and Cue-Conditioned Cognitive Processes." National Bureau of Economic Research Working Paper 9329.

Brecher, Edward M. 1972. "The Consumers Union Report on Licit and Illicit Drugs." Available at *www.druglibrary.org/schaffer/library/studies/cu/CU66.html*.

Brown, G., and L. P. Silverman. 1980. "The Retail Price of Heroin: Estimation and Applications." In Irving Leveson, ed., *Quantitative Explorations in Drug Abuse Policy.* New York: Spectrum Publications.

Chaloupka, Frank J., and Henry Saffer. 1992. "Alcohol, Illegal Drugs, Public Policy, and Crime." Paper presented at annual meeting of the Western Economic Association, San Francisco, July.

Chaloupka, Frank, and Henry Wechsler. 1995. "The Impact of Price, Availability, and

Alcohol Control Policies on Binge Drinking in College." National Bureau of Economic Research Working Paper 5319.

Chatterji, Pinka, and Sara Markowitz. 2000. "The Impact of Maternal Alcohol and Illicit Drug Use on Children's Behavioral Problems: Evidence from the Children of the National Longitudinal Surveys." National Bureau of Economic Research Working Paper 7692.

Chatterji, Pinka, et al. 2003. "Alcohol Abuse and Suicide Attempts among Youth: Correlation or Causation?" National Bureau of Economic Research Working Paper 9638.

Cook, Philip J., and Michael J. Moore. 1993. "Economic Perspectives on Alcohol-Related Violence." In S. E. Martin, ed., *Alcohol-Related Violence: Interdisciplinary Perspectives and Research Directions,* NIH Publication 93-3496, pp. 193–212. Rockville, MD: National Institute on Alcohol Abuse and Alcoholism.

———. 2000. "Alcohol." In Anthony J. Culyer and Joseph P. Newhouse, eds., *Handbook of Health Economics,* vol. 1B, pp. 1629–1673. Amsterdam: Elsevier.

Cook, Philip J., et al. 1999. "The Medical Costs of Gunshot Wounds." *Journal of the American Medical Association* 282(5):585–591.

Corman, Hope, and H. Naci Mocan. 1996. "A Time-Series Analysis of Crime and Drug Use in New York City." National Bureau of Economic Research Working Paper 5463.

DeSimone, Jeffrey. 2001. "The Effect of Cocaine Prices on Crime." *Economic Inquiry* 39(4):627–643.

Donohue, John, and Steven D. Levitt. 1998. "Guns, Violence, and the Efficiency of Illegal Markets." *American Economic Review Papers and Proceedings* 88:463–467.

Duggan, Mark. 2001. "More Guns, More Crime." *Journal of Political Economy* 109:1086–1114.

Eck, John E., and Edward R. Maguire. 2000. "Have Changes in Policing Reduced Violent Crime?" In Alfred Blumstein and Joel Wallman, eds., *The Crime Drop in America,* pp. 207–265. New York: Cambridge University Press.

Fagan, Jeffrey. 1993. "Interactions among Drugs, Alcohol, and Violence." *Health Affairs* 12, no. 4(Winter):65–79.

Goldstein, Paul J., et al. 1989. "Crack and Homicide in New York City." *Contemporary Drug Problems* 16:651–687.

Grossman, Michael, and Sara Markowitz. 1999a. "Alcohol Regulation and Violence on College Campuses." National Bureau of Economic Research Working Paper 7129. Also published in Grossman and Hsieh 2001.

———. 1999b. "Alcohol Regulation and Violence towards Children." National Bureau of Economic Research Working Paper 6359.

Grossman, Michael, and Chee-Ruey Hsieh. 2001. *Economic Analysis of Substance Use and Abuse: The Experience of Developed Countries and Lessons for Developing Countries.* Cheltenham, UK, and Northampton, MA: Edward Elgar.

House of Commons. 2000. "Cannabis." Research Paper 00/74.

Joyce, Theodore, Andrew D. Racine, and Naci Mocan. 1993. "The Consequences and Costs of Maternal Substance Abuse in New York City." National Bureau of Economic Research Working Paper 3987.

Kaestner, Robert, Theodore Joyce, and Hassan Wehbeh. 1996. "The Effect of Maternal Drug Use on Birth Weight: Measurement Error in Binary Variables." National Bureau of Economic Research Working Paper 5434.

Kellerman, Arthur L., et al. 1992. "Suicide in the Home in Relationship to Gun Ownership." *New England Journal of Medicine* 327(August 13):467–472.

Kleiman, Mark. 1992. *Against Excess: Drug Policy for Results.* Champaign, IL: Basic Books.

Kuziemko, Ilyana, and Steven D. Levitt. 2001. "An Empirical Analysis of Imprisoning Drug Offenders." National Bureau of Economic Research Working Paper 8489.

Lott, John R. 1998. *More Guns, Less Crime: Understanding Crime and Gun Control Laws.* Chicago: University of Chicago Press.

Maguire, Kathleen, and Ann L. Pastore, eds. 2002. *Sourcebook of Criminal Justice Statistics* (online). Accessed at *www.albany.edu/sourcebook.*

Markowitz, Sara. 2000a. "Criminal Violence and Alcohol Beverage Control: Evidence from an International Sample." National Bureau of Economic Research Working Paper 7481. Also published in Grossman and Hsieh 2001.

———. 2000b. "An Economic Analysis of Alcohol, Drugs, and Violent Crime in the National Crime Victimization Survey." National Bureau of Economic Research Working Paper 7982.

———. 2000c. "The Role of Alcohol and Drug Consumption in Determining Physical Fights and Weapon Carrying among Teenagers." National Bureau of Economic Research Working Paper 7500.

Markowitz, Sara, et al. 2002. "Substance Abuse and Suicidal Behavior among Adults." National Bureau of Economic Research Working Paper 8810.

Miczek, K. A., et al. 1994. "Alcohol, Drugs of Abuse, Aggression, and Violence." In Albert J. Reiss and Jerry A. Roth, eds., *Understanding and Preventing Violence,* vol. 3. Washington, DC: National Academy Press.

Miron, Jeffrey A. 1999. "Violence and the U.S. Prohibition of Drugs and Alcohol." National Bureau of Economic Research Working Paper 6950.

Mocan, H. Naci, and Erdal Tekin. 2003. "Guns, Drugs, and Juvenile Crime: Evidence from a Panel of Siblings and Twins." National Bureau of Economic Research Working Paper 9824.

Model, K. 1992. "Crime, Violence, and Drug Policy." Unpublished paper, Harvard University; cited in Saffer and Chaloupka, 1995.

National Institute on Alcohol Abuse and Alcoholism (NIAAA). 2001. "Fetal Alcohol Syndrome." Alcohol Alert #13. Available at *www.niaaa.gov.*

Nisbet, C., and F. Vakil. 1972. "Some Estimates of the Price and Expenditure Elasticities of Demand for Marijuana among UCLA Students." *Review of Economics and Statistics* 54:473–475.

Pacula, Rosalie Liccardo, and Beau Kilmer. 2003. "Marijuana and Crime: Is There a Connection beyond Prohibition?" National Bureau of Economic Research Working Paper 10046.

Reuter, Peter. 2001. "The Mismeasurement of Illegal Drug Markets: The Implications of Its Irrelevance." in S. Pozo, ed., *The Underground Economy.* Kalamazoo, MI: W. E. Upjohn Institute.

Saffer, Henry, and Frank Chaloupka. 1995. "The Demand for Illicit Drugs." National Bureau of Economic Research Working Paper 5238.

Saffer, Henry, and Michael Grossman. 1987. "Beer Taxes, the Legal Drinking Age, and Youth Motor Vehicle Fatalities." *Journal of Legal Studies* 16:351–374.

Sarbaum, Jeffrey K., Solomon W. Polachek, and Norman E. Spears. 1998. "The Effects of Price Changes on Alcohol Consumption in Alcohol-Experienced Rats." National Bureau of Economic Research Working Paper 6443. Also published in Frank Chaloupka et al., eds., *The Economic Analysis of Substance Use and Abuse.* Chicago: University of Chicago Press, 1999.

Seiden, Richard H. 1978. "Where Are They Now? A Follow-up Study of Suicide At-

tempters from the Golden Gate Bridge." *Suicide and Life Threatening Behavior* 8(4). Also accessed at *http://members.tripod.com/cfasggb/Suicide/Where_Are_They_Now.htm.*

Silverman, L. P., and N. L. Spruill. 1977. "Urban Crime and the Price of Heroin." *Journal of Urban Economics* 4(1):80–103.

Skog, O.-J., and Z. Elekes. 1993. "Alcohol and the 1950–1990 Hungarian Suicide Trend: Is there a Causal Connection?" *Acta Sociologica* 36:33–46.

Sloan, John Henry, et al. 1990. "Firearm Regulations and Rates of Suicide: A Comparison of Two Metropolitan Areas." *The New England Journal of Medicine* 6:369–373.

U.S. Bureau of the Census. 2000. *Statistical Abstract 2000.* Washington, DC: Government Printing Office.

———. 2002. *Statistical Abstract 2002.* Washington, DC: Government Printing Office.

U.S. Bureau of Justice Statistics. 2005. "National Criminal Victimization Survey." Accessed on February 21, 2005, at *www.ojp.usdoj.gov/bjs/abstract/cvusst.htm.*

U.S. Drug Enforcement Administration. 2001. "Formal Scheduling." Accessed on August 7, 2001, at *www.usdoj.gov/dea/concern/abuse/chap1/control/formal.htm.*

Van Ours, J. 1995. "The Price Elasticity of Hard Drugs: The Case of Opium in the Dutch East Indies, 1923–1938." *Journal of Political Economy* 103:261–279.

Warburton, Clark. 1932. *The Economic Results of Prohibition.* New York: Columbia University Press.

Wasserman, D., A. Varnik, and G. Eklund. 1994. "Male Suicides and Alcohol Consumption in the Former USSR." *Acta Psychiatrica Scandinavica* 89:306–313.

Zickler, Patrick. 1999. "NIDA Studies Clarify Developmental Effects of Prenatal Cocaine Exposure." In *NIDA Notes* 14. Rockville, MD: National Institute on Drug Abuse. Available at *www.nida.gov.*

Urban Economic Development

$\mathcal{U}$*rban development is* a popular notion that seems to promise something for everyone. Who can quarrel with the idea that richer is better? Prosperity reduces crime, increases education, boosts minority achievement, pays for better houses and a cleaner environment, and cuts poverty. Because of increasing returns to scale, more jobs mean new opportunities for specialization, for escaping monopoly, for finding better matches, for building bigger and better common facilities, and these costs can be spread over a wider base. Focusing development on the central parts of older cities promises to curb sprawl without sacrificing the benefits of growth—and at the same time to improve jobs for poor people and minorities.

Development makes cities great. New York is a great city because it built the Erie Canal and it dominates finance, law, advertising, and entertainment—not because it prices water right or zones efficiently. Concentrating, as I have done, on the mechanics of city life seems to miss the point—to pay attention to the windshield wipers while ignoring the car's engine. The gains that a wise development policy can bring a city and its people seem huge compared with the gains that might be reaped from the other types of policies I've discussed. A rising tide lifts all but the leakiest of boats.

In this chapter we will examine urban development policies. We'll engage in two levels of analysis. In the first part of the chapter, we'll adopt the point of view of an official or a landowner in a particular city, and ask how various policies affect people and prices in that city, holding policies in other cities fixed. In the second part, our point of view will be national or global. What happens when many cities pursue economic development strategies, optimal or otherwise? Do they cancel each other out, or does some residual benefit—or harm—bubble out of the mixture? What policies can national or state gov-

---

To get the most out of this chapter, you should be familiar with these concepts: average cost, closed-city case, deadweight loss, externalities, increasing returns to scale, localization economies, marginal cost, open-city case, Pareto optimality, potential Pareto improvement, and urbanization economies. You can find these terms in the Glossary.

ernments adopt to make economic development competition among cities more beneficial to the public at large?

The phrase "economic development" refers to several different concepts, and so at the outset I want to clarify which of these concepts we'll be concerned with. Sometimes economic development means the study of poor countries—what makes them poor, and what policies can be used to make them richer. Why textile manufacturing never took hold in India and whether China should reduce foreign direct investment are questions addressed by economic development when it's conceived in this way. This version of economic development won't be our subject. It's very important, but doesn't belong in this book.

Instead we will be concerned with another usage, one more common among people who work in city governments. In this context, economic development refers to activities and policies designed to increase the amount of market activity taking place in a particular city—bringing in jobs, for instance. Economic development in this sense examines such questions as how much cities should be willing to pay to subsidize sports stadiums, and whether slum areas should be knocked down and replaced with expensive shops.

Economic development (poor-countries version) and economic development (U.S.-cities version) resemble each other. Both are concerned with places and generally ignore the possibility that a lot of the people living in those places might be much better off if they left and went elsewhere (this is a lot more feasible with U.S. cities than it is with big countries like Indonesia or Brazil). The U.S.-cities version, however, often seeks to encourage immigration from outside, while the poor-countries version doesn't.

Some of the tactics used are also the same. The Beijing Olympics in 2008 might (or might not) contribute to China's economic development (poor-countries version), just as the 1996 Olympics might (or might not) have contributed to Atlanta's economic development (U.S.-cities version). But poor countries also consider many tactics that U.S. cities don't: enhancing agricultural productivity, for instance, building hydroelectric plants, or imposing tariffs or exchange controls. U.S. cities can't do these things and so never discuss them. Thailand has a monetary policy; New York City doesn't.

Poor countries also consider such measures as strengthening the rule of law, developing market institutions, reducing corruption and nepotism, and increasing government accountability. Some scholars even make a distinction between growth—how fast incomes increase—and development—how well the social and political underpinnings of economic growth are being put in place. Almost none of these issues figure in discussions of economic development in the U.S.-cities version, although there's no reason why they shouldn't. One shouldn't presume that all poor countries have lousy cultures and all U.S. cities have great cultures.

More fundamentally, the difference between the two topics is huge: U.S.

cities are trying to take some of the most productive, tiny pieces of real estate in the world and make them more valuable; developing countries are trying to take huge numbers of impoverished people and improve their living standards. Since a reasonably fertile imagination can usually find a few commonalities between any two randomly chosen topics, it's not surprising that a few analogies can be drawn—the most important being the shared idea that wise and careful governance helps. Nevertheless, economic development (poor-countries version) is different from economic development (U.S.-cities version), the subject of this chapter.

I. Economic Development in One City

A. Why?

1. Does Anyone Benefit?

Why might it be good to have more market activity in a city than would otherwise take place through the normal workings of a modern (developed) economy? Practically any of the problems of markets, governments, or families that I have discussed in this book can lead to too little activity in a city—high taxes or bad sewers, inadequate buses or poor policing, excessively strict zoning or dysfunctional schools. When such a problem occurs, the response is obvious: fix the problem. In this sense, this entire book has been about economic development.

Some problems, though, can't be fixed, or their solution would require political revolutions that are unlikely to be seen in the immediate future. Property taxes, for instance, instead of land taxes, will almost certainly be used to finance much of local government for as long into the future as we can anticipate, and enterprises with increasing returns to scale, from candy stores to electricity transmission companies, are likely to price their products closer to average than marginal cost. Under these circumstances, any increases in market activity often have widespread external benefits.

A new office building that wouldn't otherwise have opened in town brings with it workers who buy soda in candy stores they wouldn't otherwise have patronized. The difference between the price they pay for the soda and its marginal cost is an external benefit. So are the property taxes the building pays. If the office building pays average cost for its electricity or for transmission of natural gas, another external benefit is generated. Notice that these external benefits arise not from any particular activity that goes on in the office building, but from activity in general.

New activity also brings with it the urbanization and localization economies I talked about in Chapter 2. Most of these benefits are external: knowledge and most innovation, for instance, and the lower prices that come about through specialization, more efficient queuing, and less monopoly power. Ur-

banization benefits come from any activity and affect any other activity, while localization affects only those firms in the same general industry.

Jobs can have external benefits, too. If more employment means fewer people are receiving welfare or unemployment insurance, the reduction in government expenditure is an external benefit. The extent (if any) to which employees are paid more than the minimum amount that would make them willing to do the work is also an external benefit; this may be substantial if the workers would otherwise have been idle.[1]

Jobs may also reduce the threat of some types of crime. Raphael and Winter-Ebmer (2001), for instance, find that positive economic-development events, like winning defense contracts, cut a state's property crime, and negative events, like increases in the price of gas, increase property crime. Higher unemployment for men also increases rape.

However, the external benefits of more market activity in a city, while they are often substantial, can sometimes be overestimated. First, remember that, in most cases, only part of a secondary expenditure counts as a benefit, not all of it. The entire expenditure on soda, for instance, is not an external benefit, just the difference between price and marginal cost. In particular, jobs are a benefit only to the extent that they pay more than their workers could make elsewhere or more than the value of their workers' leisure, whichever is greater. Paul Courant, in a famous essay on economic development (1994, p. 869), gives a good example to show why jobs should not be considered a benefit: "[Suppose] a locality decided to engineer a gold rush on land that had no gold. The local government could go out and buy the gold on the open market, and incur the cost of transporting it and burying it. The gold would then surely be worth digging up, but in the end all that would be available would be the value of the gold. The costs of transporting, burying, and digging would be lost forever."

Second, activity causes negative externalities as well as positive: pollution and congestion, for instance, storm-water runoff and accidents. Police, fire suppression, and trash pickup services are usually priced below marginal cost, as well as many mass transit systems at peak hours; the difference between marginal cost and price for these services is a negative externality.

Finally, increases in the price of land are often counted among the benefits of development. As I emphasized in Chapter 6, this procedure isn't right. The amount of land being used at any particular location doesn't change, and potential Pareto improvements occur only when the amount of something be-

---

1. The marginal cost of producing soda is a real cost to the world; it represents resources that would otherwise have been put to some other valuable use. Similarly, the value of what a worker could have produced working another job or staying home minding the kids is also a real cost. Soda sold and jobs generated thus count as economic benefits only to the extent that they're overpriced or overpaid, and no more.

ing done changes. Whatever landowners gain from higher prices, renters lose; nothing else happens.

2. Who Benefits?

The answer to this question about economic development in the open-city model is easy: only landowners. This answer usually doesn't please economic development advocates. However, the standard alternative to the open-city model, the closed-city model, is entirely inappropriate for analyzing development, since it presumes no development.

Intermediate models—models in which better opportunities in one city trigger migration, but not fast enough migration to offset those opportunities immediately—give intermediate answers. People other than landowners gain for a while, but those gains are progressively transferred to landowners over time, and they eventually disappear. In no model, of course, do renters gain. Retired people who don't own real property, for instance, are almost certainly big losers from economic development in any model.

Simple supply-and-demand diagrams can illustrate these issues. First, consider the labor market. Suppose an economic development victory (or a stroke of luck) raises demand for labor in a particular city. A car manufacturing plant opens up, or a university in town gets a large federal grant for ceramic research, or a drug that a local company has patented turns out to cure arthritis. In Figure 18.1, this moves the labor demand schedule up from D to D^*: at any wage, companies in this town want to hire more workers.

Whether workers gain from this demand shift depends on the elasticity of supply. In the short run, say a year or two, not many workers can migrate to this city, and so supply is fairly inelastic, as in Figure 18.1a. Wages go up, employment doesn't go up much, and workers—the workers originally living in the city—are better off. This is essentially the closed-city model.

But over time, more and more workers will migrate if wages in this city are higher than wages elsewhere. Thus in the long run, the supply curve of labor may be almost flat, as shown in figure 18.1b. So employment goes up but real wages stay the same, and hometown workers are no better off.

On the other hand, in the land market supply is inelastic in the long run and in the short run. Figure 18.2 illustrates. There just isn't any more land within ten miles of the new plant. So when demand for land shifts up from d to d^* because people are attracted by the job opportunities (or because they learn that this city is a very nice place to live), the price of land goes up but the quantity has to stay the same. Owners of land gain; consumers of land lose.

Who reaps the benefits of development, then, is ultimately an empirical question about the elasticity of labor supply. Many studies have been done, and on some points there is consensus. In the long run (five to ten years), growth raises housing prices considerably, and in the short run, it raises wages and reduces unemployment. The increase in housing prices is large: a 1 per-

Real wage

D^*

Supply in short run

D

Quantity of
labor

Real wage

D^*

D

Supply in long run

Quantity of
labor

cent increase in the demand for labor in a metropolitan area (because of a
manufacturer's new defense contracts, say, or some huge new factories built by
foreign firms) raises housing prices by between half of 1 percent and 1 per-
cent. Bartik (1991) surveys a large number of studies on which this consensus
is based.

The major point of disagreement is on whether economic development
has any effect at all on unemployment or wages in the long run—or whether
migration eliminates all the short-run gains that workers make and transfers

Figure 18.2 Land
market response to
economic develop-
ment.
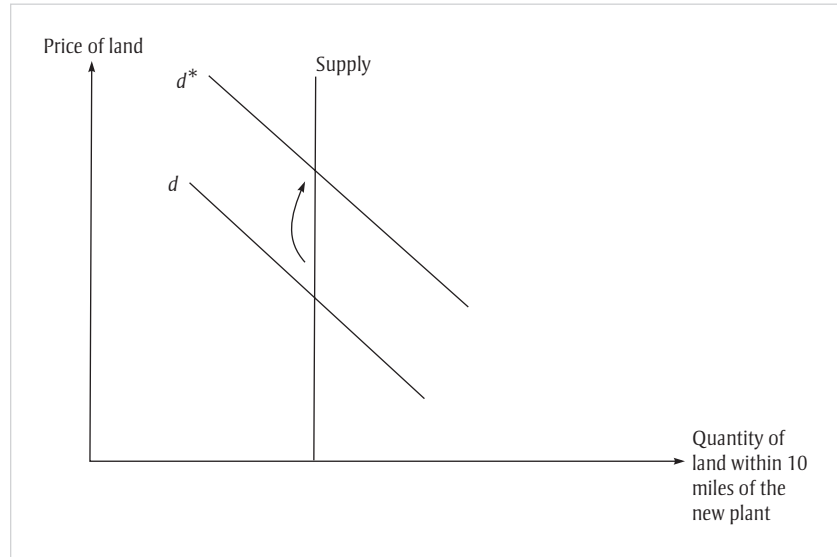
them to landowners. Blanchard and Katz (1992) look at state data between
1950 and 1990, and conclude that after five to seven years, economic develop-
ment has no effect on unemployment, and after about ten years it has no ef-
fect on wages. New defense contracts or new foreign plants are offset by new
migration, which is triggered by low unemployment. Migration eliminates the
low unemployment that spurred it.

Bartik (1991) claims that economic development can have permanent ef-
fects on unemployment and wages, although the effects he finds are quite
small. Using data for twenty-five large metropolitan areas between the years
1972 and 1986, he concludes (p. 95): "Suppose some economic development
policy creates 100 net new jobs for a metropolitan area . . . [I]n the long run, 6
or 7 of the 100 jobs will go to local residents who would otherwise by unem-
ployed, and 16 will go to local residents who would otherwise be out of the la-
bor force. The other 77 or 78 jobs go to in-migrants." He also finds a small
long-run effect of economic development on wages, although it's not statisti-
cally significant.

Bartik reconciles his results with the open-city model by arguing that the
one-year effect of economic development permanently changes who the local
residents are. People emerge from a temporary boom with different skills and
attitudes that stay with them permanently, he says (p. 76): "Due to faster
growth, in the short run some persons in the area will obtain jobs who other-
wise would be unemployed. In addition, some will move up to better jobs. The
short-run experiences of these persons change their values, skills, self-con-
fidence, and reputation." This is called the "hysteresis effect"—when tempo-
rary events have later, permanent repercussions.

Aside from the disconfirming results of Blanchard and Katz (and several other studies), there are two more reasons to be skeptical of Bartik's conclusion. First, Bartik's analysis implies that states with higher average employment growth should have lower average unemployment. Blanchard and Katz look for such a relationship and do not find one. Second, on a theoretical level, if employment during a boom has the long-run benefits Bartik claims for it, then workers should be willing to accept somewhat lower wages in return for those long-run benefits. So what appear to be long-run benefits, if they exist, would really just be repayment for an investment made in the temporary boom.

Clearly, then, migration offsets in less than a decade most (according to Bartik), if not all (Blanchard and Katz), of the benefits workers might gain from economic development. Is it really plausible that people are really so willing to pack up, move their families, and abandon their loved ones, just to add a couple of bucks to their paychecks or to find a job a little more quickly? Marston (1985) finds that migration flows are probably large enough to do what's claimed. During a four-year period, he found that 13.9 percent of metropolitan population moves between areas; this doesn't count immigrants from abroad. Since economic development policies that raise employment by even 1 percent a year are rare (if they exist at all), small changes in normal migration are more than big enough to offset almost all economic development policies. There are lots of people, all the time, who want to leave their home city, and many who wouldn't particularly mind doing so—young adults setting out on their own, people recovering from bad relationships or traumatic incidents, recent arrivals, people who hate the weather where they live. Once again, when numbers count, introspection is a poor guide.

Finally, notice that all of this empirical work deals with metropolitan areas or states—big areas where leaving is fairly costly. They do not focus on the effect that jobs in a particular neighborhood or a particular political jurisdiction have on the workers who live there. Since "migration" to small areas is easy—you just change your commute or you move to a place that doesn't preclude your seeing old friends or relatives frequently—the expectation is that short-run benefits to workers will be much smaller, adjustment will be much quicker, and the long-run benefits, if any, much smaller. Bartik finds some evidence to support this idea, although some of the evidence in the spatial mismatch literature (see Chapter 11) seems to contradict it.

However, even if most of the benefits of development accrue to landowners, there are still many reasons to undertake development activities that are potential Pareto improvements. Landowners are people, too, and many of them are quite ordinary people. Landowners' gains, moreover, are potentially taxable, and can be used, at least in the short run, to help people who aren't landowners. But the arguments in favor of development activities that aren't potential Pareto improvements—that they're a good way of helping poor people, for instance—are fairly weak.

3. Are Minorities or Minority Neighborhoods Different?

A variety of theoretical arguments can be made for the proposition that minority workers retain a larger proportion of the benefits of economic development than whites do—why, in other words, development might be a good strategy for helping minority workers even if it's a poor strategy for helping white workers. Many theoretical arguments can be made for the opposite conclusion, too. Most of the arguments on both sides are weak and speculative.

One argument in favor of this proposition is based on Bartik's idea that boom-time employment changes people. Perhaps minorities are changed more than whites, or more minorities than whites are changed. Minorities would change more, for instance, if they were more lacking in self-confidence or were more deeply afflicted with negative stereotypes that work experience could overcome. More minorities would change if normal unemployment rates were higher and fell more—which is true.

Speed of migration may also matter, and some evidence indicates that minorities, particularly African Americans, are slower to migrate than whites (Krieg 1990, Raphael and Riker 1999). If, for some reason, minorities are more likely than whites to receive benefits from development, then slow migration may allow them to benefit for a longer time, and reduce the gains to landowners. But for slow migration to make a positive difference for minorities, somehow the benefits of development have to be restricted to minorities. For instance, if development creates jobs that pay more than market-clearing wages, slow migration can help minority workers only if white workers don't take those jobs. Jobs in small businesses that have predominantly minority customers might be difficult for white workers to take and thus might fit this scenario. But it's not clear what other types of jobs would work like this.

So we need to turn to empirical evidence. Here, too, the picture is mixed. Bartik (1991, 1993), Frey and Speare (1988), Bound and Holzer (1993), and Freeman (1991) find that job-creation programs increase employment more among African Americans than among European Americans, while others find the opposite (Ihlanfeldt and Sjoquist 1991), or no difference (Ihlanfeldt 1992). Peters and Fisher (2002) find that the vast majority of jobs in urban enterprise zones go to people who live outside the zones, and that the vast majority of zone residents who are employed work outside the zones. None of these studies attempts to correct for rising rents. Little is known about Hispanics or Asian Americans. The safest conclusion is that African American workers probably gain a little more than European Americans from metropolitan economic development, but in the long run neither group gains much.

4. Are Old Central Cities Different?

No good empirical evidence is available on this question. The external benefits of development may be greater in old central cities than in newer and less densely populated parts of metropolitan areas, for several reasons. Infrastructure may be priced further above marginal cost because usage has been declin-

ing and excess capacity is present. Jobs may be reachable by mass transit, and the social cost of trips by mass transit may be less than the social cost of trips by car. Mass transit also makes jobs available to poorer people, who might otherwise be idle or receiving welfare or unemployment insurance. Densely packed areas may also be more likely to give rise to localization and urbanization economies (although the most famous recent example of localization economies, Silicon Valley in California, did not require significant population density, even by American standards).

On the other hand, roads may be narrower and more congested, in old city centers, and the dangers of pollution and accidents greater. Some of the infrastructure may be obsolete or rotting away. The social cost of rush-hour mass transit trips may be very high, and the new jobs may be farther from the residences of the people who actually get them. Coming up with a blanket rule for the net benefits of central city development is probably not possible.

There are also some reasons to think that a somewhat smaller proportion of these benefits accrue to landowners in the medium term (five to ten years). Central cities are the location of the largest concentrations of public housing and other types of subsidized housing. When this housing becomes more attractive, tenants don't face higher rents. The same applies to rent-controlled tenants, who are also more common in central cities. On the other hand, in outlying suburban areas and newer urban areas, more people own land.

5. Summary of Benefits

Development often has net external benefits, and those benefits may sometimes be greater in old central cities. The short-run benefits are also probably greater for African Americans. Many of the benefits of development accrue to landowners in the long run, which is not disastrous. Different development projects have different kinds of benefits, and there is little reason to do development projects that are not potential Pareto improvements.

B. How?

At the most basic level, any worthwhile project is good for economic development. A new fire truck in town, for instance, if the services it provides are worth more than the taxes needed to buy it and operate it, makes living or running a business nearby more attractive, and so induces new investment and migration with all the attendant externalities, both positive and negative.

But usually economic development (the U.S.-cities version) refers to a more restricted set of policies. These are policies for which a major portion of the benefits are external effects that any measures that increased market activity would have. The fire truck doesn't qualify, because most of its benefits— faster suppression of fires and concomitant lower insurance rates, greater freedom to build more flammable structures—aren't common with every policy

that increases market activity. In contrast, building a baseball stadium is usually considered an economic development project because most of the benefits cited are the same ones that any other endeavor that induced similar amounts of spending in the same area by similar people would have.

This is not a hard-and-fast distinction. Proponents of baseball stadiums could conceivably argue their case on the grounds of the consumer surplus that fans would enjoy. Such arguments, however, are heard so rarely that it seems legitimate to presume that stadium proponents think most of the benefits come from stimulation of market activity. Conversely, fire truck proponents might emphasize the stimulative effect good fire trucks have, but they usually don't. And "major," of course, is an imprecise word.

The distinction has two advantages. First, it seems to correspond roughly with how ordinary people use words. Second, it lets us concentrate in this chapter on a new set of benefits and costs that we haven't considered before, rather than rehashing material we've already gone over. But in concentrating on the benefits that arise from stimulating more market activity, we still won't forget about more traditional costs and benefits, even if they're in the background.

With this in mind, we can now look at several of the more prominent activities that are traditionally considered ways to promote economic development.

1. General Taxes

Lower taxes, everything else being equal, make a city more desirable, and so should attract more people and more business. This would trigger development externalities. Lower taxes, of course, produce other benefits as well, just like a good fire truck: less burden, and smaller deadweight loss.

There is good empirical evidence to support this assertion. Many economists have studied the effects of taxes on business location and economic development, and several others have tried to summarize the results of these studies (Bartik 1991, Phillips and Goss 1995, Wasylenko 1997). The consensus is that a 10 percent decrease in the taxes residents and businesses pay in a metropolitan area will cause an increase in most measures of market activity (jobs, investment, new start-ups, and so on) of between 1 and 5 percent in that metropolitan area, everything else being equal. Taxes have a bigger effect on location within a metropolitan area—a town that cuts taxes by 10 percent, when surrounding towns don't, will gain more than a state that cuts taxes 10 percent, but much of the new activity will just move from nearby within the metropolitan area. Such shifts in business location create much smaller external benefits than shifts from farther outside the metropolitan area.

Notice that when I say that taxes affect business location, I'm not saying that they're decisive in all or even most business decisions. No amount of tax cutting by Florida is going to induce ski resorts to move from Vermont to

Pensacola. Instead, I'm saying that there are a few business decisions in which other factors are close enough that taxes matter: a company that can't otherwise decide where to build a new plant, an existing firm that may or may not hire a part-time bookkeeper, a store owner who is trying to decide whether to retire now or stay in business for another year. Just as events that swing only a handful of voters and have no effect on the vast majority can nevertheless decide elections, so taxes can exert the kind of influence econometric studies show they have, without ever affecting the decisions of an overwhelming majority of businesspeople.

Moreover, the businesspeople whose decisions change because of taxes may not even be aware of this. The firm deciding between two locations for a new plant may just add up two long columns of costs, each of which includes taxes, and choose the location for which the sum is smaller. It has no reason to ask itself whether its conclusion would have been different if one city's taxes had been lower. The store owner deciding when to retire may have no idea how much more his customers would spend in his store if their local income taxes were lower.

For both of these reasons, then, introspection is a very poor way of assessing the effect of taxes on business location. Surveys of businesspeople's introspection are even worse, because on top of all the other weaknesses of such surveys, they exclude businesses that have shut down and those that never started.

Even though lower taxes do spur economic development, across-the-board tax cuts aren't necessarily the best way or even a good way of making a city grow. Taxes pay for public services, and in the long run, the only way you can cut taxes is to cut public services. Even the most wasteful of public service expenditures is still an expenditure and so carries with it economic development externalities. Most public service expenditures, moreover, are not wasteful: good schools, good roads, good police, and good sewers all attract firms and residents. Public services are a little harder to measure than taxes, but the consensus of empirical studies is similar to the consensus on taxes: more spending on public services has a definite, measurable, positive effect on economic development, although the effect is not terribly big (see Bartik 1991 and Fisher 1997).

We should not forget, either, that many public services produce benefits above and beyond those connected with attracting new jobs. A long-time city resident, for instance, gains from better police protection if it permits her to take walks in the park at dusk, even if her residence and spending patterns stay the same.

This argument against cutting taxes, however, is somewhat imprecise. Tax revenue is what matters for public service expenditure, while tax rates are what matter for economic development. A 10 percent cut in tax rates translates into a 10 percent cut in tax revenue only if the tax rate cut spurs no eco-

nomic development. Because we know that such a tax rate cut is likely to spur some economic development, the argument against tax (rate) cuts is not nearly so strong as it first appears. But the economic development effect of tax rate cuts is small, as we have seen, and so in general being precise about the distinction between tax rates and tax revenues is not especially important.

This general conclusion about the effect of taxes on economic development has one possible exception: business location within a metropolitan area. The studies that confined their attention to a single metropolitan area found that differences among jurisdictions in tax rates corresponded with much larger differences in economic development variables; many studies, in fact, found that 10 percent lower tax rates corresponded with more than 10 percent more jobs or investment. If this were in fact how the world worked, then jurisdictions could always collect more revenues and provide more services by cutting their tax rates.

There is some reason to be skeptical of these results, however. Causation may not be running from tax rates to development: jurisdictions that were better managed or more friendly to business or in general more attractive in unmeasured ways might have both lower taxes (either because of their good management or because of their larger tax base) and more jobs. The tax rate might just be the visible tip of a larger, unobserved—and probably more expensive—package.

We should also note that the effects reported in the general surveys are long-run effects—how much more business activity will eventually occur if taxes stay fixed at the new low rate for a long enough time. The process of adjustment is gradual, because many of the decisions in which taxes matter have to be made only occasionally. The conventional estimate is that every year about 9 percent of the gap between where business activity was at the beginning of the year and where it will eventually end up (if nothing changes) is eliminated (Bartik 1991; Helms 1985).

So local governments must lose tax revenues immediately from cutting rates, and will regain those revenues only far in the future. Even if studies show increased activity making up for lower tax rates in the long run, the government will lose money in the short run, and the long-run gains won't necessarily make up for the short-run losses. (For more on this, see Box 18A.)

Finally, most of the studies look only at a specific type of business activity—manufacturing investment, mainly, or new start-ups. Since cutting tax rates doesn't produce more land, some of the measured activity comes about by displacing other activity. Thus the measured increase in manufacturing, for example, may overstate the net increase in business activity.

Of course, the sort of cost-benefit analysis we have done throughout this book would almost always call for reducing tax rates that are so high that they net almost no money. Such tax rates cause big deadweight losses, even if the activity they dissuade carries with it no economic development benefits.

Box 18A

When Do Tax Cuts Really Pay for Themselves?

≡ Let G be the annual addition to tax revenue in the long run that a tax rate cut causes through increased economic development, and let L denote the annual loss in taxes that the tax rate cut would cause if it spurred no business expansion. The standard question is whether the gains from a tax cut will outweigh the losses:

(1) $$G > L.$$

But the real question for determining if the government can cut tax rates without cutting public services is whether the present value of gains will be greater than the present value of losses. Let r denote the rate of interest, and a denote the phase-in rate (about .09, the proportion of the difference between the current level of business activity and the ultimate level that disappears every year). The present value of losses is

$$L/r,$$

since the losses start immediately. The actual level of gains in year t is

$$G(1 - e^{-at}),$$

and so the present value of gains is

$$(G/r)(a/[a + r]).$$

Thus the present value of gains exceeds the present value of losses if and only if

$$G(a/[a + r]) > L,$$

which is a considerably more stringent condition than (1). If $r = .045$, for instance, a 10 percent cut in tax rates will increase the present value of city revenue only if it induces an increase of 15 percent or more in business activity.

Talking about jobs and business activity adds almost nothing to the standard analysis.

Interestingly, too, the standard analysis arrives at a prescribed method for creating potential Pareto improvements that's also a good prescription for economic development: land taxes. Taxes on land can't discourage business activity; what discourages business activity are taxes on improvements, or wages, or utility use, or something else that people can control. If all jurisdictions relied solely on land taxes, economic development would be indepen-

dent of tax rates. Oates and Schwab (1997) found that when towns in Pennsylvania shifted some of their property tax burden from structures to land, considerable economic development ensued.

2. Tax Abatements

The big drawback in trying to induce economic development through cuts in general tax rates—tax rates that apply to broad classes of individuals or firms—is that many of the beneficiaries don't change their behavior as a result. Another drawback is that the effects of general tax cuts are hard to perceive unless you're an economist or a statistician. With broad cuts, political leaders can't point to any particular job or building and say they created it.

Both problems suggest the same solution: give tax cuts only where they're going to make a difference. That way money isn't wasted by giving cuts to firms for doing what they would have done anyway, and political leaders can point with pride to tangible accomplishments, rather to than pages of statistical gobbledy-gook. Such sharpshooting has another advantage, too—it lets governments focus the biggest tax breaks on the activities with the biggest positive externalities.

Targeted, specific tax reductions are therefore a popular (but controversial) economic development tool. The main method of targeting tax cuts is called a **tax abatement.** In a tax abatement, the government agrees to collect lower property taxes than it ordinarily would from a property—usually a newly constructed or substantially rehabilitated building—for a specified period of time.

a. Why Tax Abatements Have to Be Administered Well   Tax abatements work well if the people administering them can discern those decisions for which taxes matter, and if they have incentives to act properly on that information. Similarly, preventive detention would be a good way to fight crime if police could tell who was going to commit crimes and when, and could be trusted to act on that information. And enforcing speed limits and drunk driving laws only against drivers who were going to have accidents would also be a significantly more efficient system, if information and incentives made it feasible. In all three cases, though, the selective programs will work poorly—in fact will often be worse than no program at all—if either the information or the incentives are not of extremely high quality.

A poorly executed attempt at selective tax reduction (or selective preventive detention or selective traffic-law enforcement) can be worse than no attempt at all, for two reasons. The first reason is the mathematics of deadweight loss: cutting a tax in half doesn't cut the deadweight loss it causes in half. Taxes are like rainstorms: an extra little bit of rain coming on top of a drizzle causes a lot less damage than the same amount during a torrent, when

many streams and sewers are already overflowing. Any loss in net revenue that a tax abatement causes has to be made up with higher taxes on other taxpayers—and greater deadweight losses.

Unless abatement recipients are chosen accurately and the size of the abatement calibrated finely, a tax abatement is likely to increase deadweight loss rather than decrease it. If the abatement is too small, the recipients won't change their behavior much and little deadweight loss will be alleviated directly. Other taxpayers, however, will have to pay higher taxes to make up for the subsidy, and since they will be starting from a high base, the presumption is that their deadweight losses will overwhelm whatever gains the recipients will realize.

If the abatement is too large, the last dollar of the abatement will come from reducing taxes much lower than those paid by other taxpayers, who will see their taxes being forced up. The reductions in deadweight loss the world gains from this last dollar of abatement are likely to be smaller than the deadweight losses associated with raising the high taxes others pay. Box 18B illustrates these outcomes.

The moral of this story is that a perfectly designed tax abatement system is better than uniform taxation, but a tax abatement system that falls short of perfect design, even by a fairly small amount, is worse than uniform taxes. Designing tax abatements is like swallowing swords: the reward is handsome if you do it well, but if you make even a small error it would be better never to have tried at all.

The other reason why a poorly conceived tax abatement system might be worse than uniform taxation goes by the name of **rent-seeking.** If firms with certain characteristics get abatements and others don't, then firms may devote resources to acquiring the characteristics that result in abatements. To the extent that abatements are just a shifting of the tax burden from one taxpayer to another, devoting resources to securing an abatement is a waste of those resources: if the abatement were awarded the same way and no resources were expended, no one would be worse off. The term *rent-seeking* refers to the expenditure of resources to seek "rent"—some recompense greater than whatever is required to make you indifferent about doing something you do.

Whether rent-seeking occurs and what form it takes depends on how abatements are awarded. Rent-seeking wouldn't occur at all if abatements were handed out randomly or on the basis of some immutable criteria. But random abatements would contradict the premise of targeting, and even if criteria couldn't be changed by potential recipients, they could be changed by the agency giving out abatements. New Jersey abatement law, for example, has been amended approximately every five to ten years since it was first adopted in 1961. Such mutability indicates that lobbying for beneficial decisions may pay off for an individual firm. Lobbying is the quintessential rent-seeking ac-

tivity, since it uses resources like the time of owners, lawyers, and publicists that might otherwise be spent in directly productive ways.

Tax abatements can induce other forms of rent-seeking, too, but these depend more closely on the actual criteria by which abatements are awarded. When we consider those criteria, we'll consider the rent-seeking behavior they might cause.

b. Are Tax Abatements Likely to Be Administered Well?   We have looked at two reasons—nonproportionate deadweight losses and rent-seeking—why poorly targeted abatements may be worse than none at all. The question thus becomes how well abatements can be targeted: how much relevant information do decision makers have?

Decision makers have several possible sources of information. First, they could ask businesspeople to tell them how much tax relief they need. The basic problem with this approach is that businesspeople have no reason to answer truthfully and every reason to exaggerate. Moreover, the number of people you would have to ask is enormous. Land, for instance, is an important cost, and to find the relevant price for a particular plot of land, you have to figure out what the next highest bidder would be willing to pay, which entails finding out who the next highest bidder is. Data on job growth show that new businesses from outside a metropolitan area are a relatively small part of job growth. That means that you can't confine your attention to outside businesses that have expressed an interest in moving in; you have to gather this information on a regular basis from existing businesses that might expand or contract, as well, and from would-be entrepreneurs thinking about starting up a business.

Second, decision makers could look at the sort of business that firms engage in and base decisions about abatements on that. For instance, Michigan permits abatements for manufacturing firms but doesn't allow them for commercial activities (commercial abatements were permitted between 1978 and 1988, however, and commercial firms in Michigan can benefit from tax-increment financing). The theory is that manufacturing firms are more footloose than commercial, and that they provide greater positive externalities in the form of semiskilled jobs paying above-market-clearing wages. Little is known, however, about how sensitive different types of businesses actually are to tax incentives or what types of externalities they usually generate; economists, after all, have barely reached a weak consensus on the effects of business as a whole.

Bartik (1991, pp. 216–247), in his summary of tax effects, cites several papers that try to break down tax sensitivity for various types of firms, but few distinct lines can be drawn. Manufacturing does appear to be more sensitive to taxes than the average business, but no clear pattern emerges in compari-

See for Yourself

≣ Suppose there are four firms: two commercial firms and two manufacturing companies. Before taxes (that is, if taxes were zero), the benefit of locating in our town is $19 for one commercial firm, and $8 each for the other commercial firm and one manufacturing firm. The benefit for the other manufacturing firm is $4. (For profit-maximizing firms, the benefit is the difference between profit at zero taxes in our town and profit at the next best alternative location.) Figure 18B.1 illustrates. Our town can distinguish between manufacturing and commercial firms, but cannot distinguish between firms within a class.

The total amount that has to be raised in taxes from the four firms is $22.50.

The town government is contemplating offering tax abatements for manufacturing firms, since they appear to be sensitive to taxation. What range of manufacturing abatements would produce less deadweight loss than uniform taxation? What range of abatements would produce more deadweight loss?

Step 1: Find uniform taxes that raise the required $22.50. First, see how many firms operate. Only if the tax is less than $4 will all four firms locate in our town, but at that rate the revenue will be less than $16. So at least one firm—the second manufacturing firm—will not operate if uniform taxes raise more than $16. If three firms locate in the town, the tax on each will be $7.50, for a revenue total of $22.50. Since the remaining three firms all have a benefit of $8 or more, this uniform tax will work.

With a tax of $7.50 for each operating firm, the second manufacturing firm will not operate in our town. The deadweight loss will be $4. Figure 18B.2 illustrates.

Step 2: Find the range of manufacturing abatements too small to change deadweight loss at all. The town decides to lower
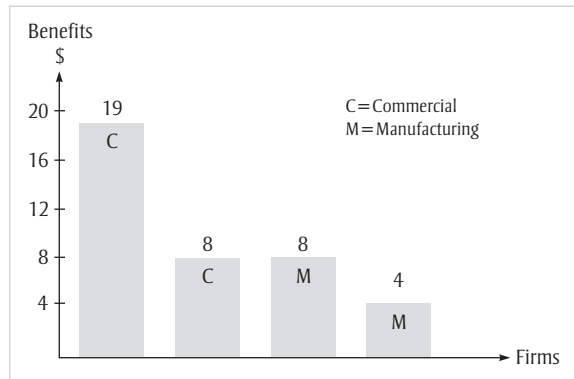
Figure 18B.1 Benefits without taxes.
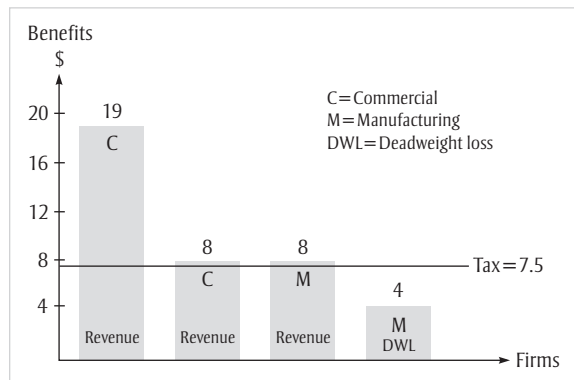
Figure 18B.2 Uniform taxes.

the tax on manufacturing firms, starting from $7.50. As the tax is gradually lowered, at some point one of two things will happen: either the second manufacturing firm will enter, or commercial taxes will become so high that the second commercial firm will leave. The second manufacturing firm will enter when the manufacturing tax falls to $4. Before that, for every dollar that the manufacturing tax falls, the tax on each commercial firm rises by $0.50 to make up for it. Thus when the manufacturing tax falls by more than a dollar—so it's below $6.50—the tax on commercial firms will be above $8, and the second commercial firm will leave.

Step 3: Summarize the results for manufacturing taxes above $4 and calculate deadweight loss. If the manufacturing tax is $6.50 or above, nothing real happens—the abatement is just a transfer from the commercial firms to the manufacturing firm. Deadweight loss remains at $4.

If the manufacturing tax is between $4 and $6.50, only one commercial firm and one manufacturing firm will operate. Deadweight loss is $12: $4 from the manufacturing firm that isn't operating, and $8 from the commercial firm that isn't operating. Abatements like this are worse than uniform taxes.

Step 4: Find out what happens when the manufacturing tax is $4 or below. In this range, both manufacturing firms will operate. The question is whether the second commercial firm will operate. If manufacturing taxes are $4, then manufacturing firms contribute $8 in taxes and the tax on commercial firms is $7.25 each (($22.50 − 8)/2). The second commercial firm operates, and deadweight loss is zero. This is how abatements are alleged to operate always.

Since the number of manufacturing firms is now the same as the number of commercial firms, in this range the commercial tax goes up at the same rate as the manufacturing tax goes down. The second commercial firm will shut down when the commercial tax goes over $8. This is $0.75 more than the commercial tax when the manufacturing tax is $4. So when the manufacturing tax falls below $3.25, the second commercial firm shuts down, and deadweight loss rises to $8—more than the deadweight loss with uniform taxes.

Figure 18B.3 summarizes deadweight loss as a function of the tax on manufacturing firms. Only when that tax is between $3.25 and $4 will total deadweight loss be less than it was under uniform taxes.
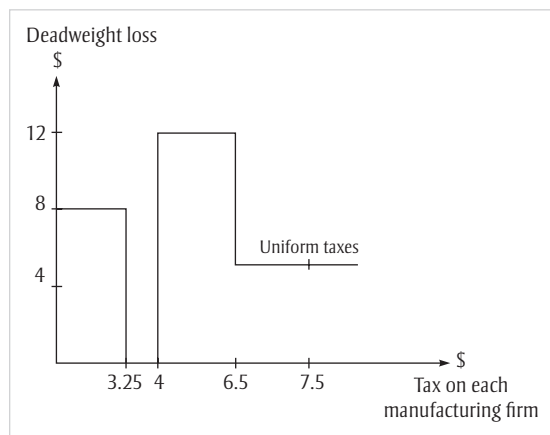
Figure 18B.3 Deadweight loss as a function of manufacturing tax.

sons of large firms with small, or high-tech firms with others. Within manufacturing, plastics, printing and publishing, and electrical components appear to be quite insensitive to taxes, but it's not clear which industries are more sensitive than average. The numerical example in Box 18B shows that dispersion within the group of firms getting tax abatements matters as much as averages, and since we know little about dispersion, it seems safe to conclude that the amount of useful information local decision makers have about the types of firms that should get abatements is small. (Even Michigan's manufacturing distinction is questionable; officials in Newark, in contrast, have argued that manufacturing plants are polluting and therefore generate negative externalities, these negative externalities make few New Jersey municipalities willing to zone them in, and thus they are neither footloose nor desirable.)

Tenure is another piece of information that decision makers could use in determining which firms should be granted abatements. In many states, abatements are limited to new construction and substantial rehabilitation. The theory is that decisions about the size and height of structures are much more easily influenced in the planning stages than they are after a building is built. The cost of adding a fifth story to an existing four-story building is much greater than the difference between the cost of a new five-story building and a new four-story building; the cost of wiring an existing building for Ethernet is more than the difference between the cost of a new building with Ethernet and one without. Thus the cost of inducing expansion and enhancement of structures is less before they're built than after. Since employment and many other aspects of market activity are often complementary to structure, this ease of influencing structure carries over to ease of influencing employment and market activity. Once a building is built, going from five employees to six can be expensive if you have to split somebody's office or reconfigure the interior walls. In the planning stage, it's a lot easier: you can design six offices in the space you had for five, or you can plan to make the building a little bigger. On average, then, inducements to firms involved in building new structures should be more effective than inducements offered to other firms.

That doesn't mean that firms in existing structures make no decisions that are sensitive to taxes. Payroll taxes affect how many workers an existing firm employs, and other taxes can affect decisions about whether to stay in business or whether to move to a new location. Buildings can also be on the margin of abandonment.

In some circumstances, however, abatements on new construction can lead to obvious deadweight losses. As an example, consider the office-building market in a declining city. Suppose the existing office buildings are all identical, and owners of the existing buildings collect enough rent to cover their operating costs and taxes, but no more. If an abatement or other subsidy leads to the construction of a new office building, rents tenants pay will have to go down if the new building is occupied, or some old buildings will have to be

abandoned—since demand curves slope down, the only way you can get more office space rented is by lowering rents. But if rents fall, all the old buildings will be abandoned. So in equilibrium, every square foot of new office space built must be offset by a square foot of old office space abandoned—or rents will fall. Since all that happens is that new office space replaces perfectly serviceable old space, no one is better off, and the new construction the abatement induced was pure waste. For more detail, see Box 18C.

By itself, the greater average malleability of new capital doesn't guarantee that a policy of abatements for new construction will pick out the right properties to abate, or that the costs of its misses will be less than the benefits of its hits. But it might happen that way—the question is ultimately empirical.

Confining abatements to new construction and substantial rehabilitation also induces two kinds of rent-seeking behavior. Rewarding new construction reduces the incentive for maintenance: the faster you let your building run down, the sooner you're in a position to fix it up (or tear it down) and get an abatement.

On the other hand, with parking lots and mainly empty spaces, new construction abatements can cause rent-seeking owners to delay construction. The reason is uncertainty: once you build, you lose forever the option of getting an abatement or the prospect of getting a more generous one. Abatement rules change frequently, and to the extent discretion is part of the process, government favorites and priorities change even more often. Even if building is profitable now, it might be more profitable if the owners waited a little while and the rules, the players, or the owners' political standing improved. In this way, tying abatements to timing can distort the timing of construction and maintenance.

The final piece of information that's sometimes used in awarding abatements is location. Many states specify that abatements can be granted only in areas labeled "blighted" or "in need of rehabilitation," or some similar phrase. The argument for location is primarily an argument about external benefits, rather than an argument about sensitivity to taxes: buildings in these areas are believed to produce greater external benefits than buildings in other places. The first new buildings are said to "jump-start" development, in a frequently used metaphor. Rauch (1993) shows that developers of private industrial parks offer early tenants discounts in order to attain a "critical mass" that will permit economies of scale to be realized. The external benefits of early entrants are great—they make the developer credible and they guarantee later entrants that they won't be alone—and so developers are willing to pay them. The location argument for abatements is that cities should act the way these private developers do.

Of course, spotting potential for agglomeration economies and matching that potential with the right locations is a difficult task—at least as difficult as spotting tax sensitivity. Successful developers are able to do it, but few people

Box 18C

When Can Abatements on New Buildings Topple Old?

≡ How could it come about that existing buildings collect enough rent to cover their operating costs and no more? This circumstance is quite likely in a city that's declining faster than people many years ago thought it would.

In the short run, the supply of office space is like this: any amount you want is available at operating cost, up to the amount of current capacity. Nothing more than current capacity is available. So the supply curve looks like Figure 18C.1.

When you build an office building, you hope that in the future demand will be sufficiently high, relative to capacity, that at equilibrium all space will be used, and rent will be greater than operating cost, as in Figure 18C.2.
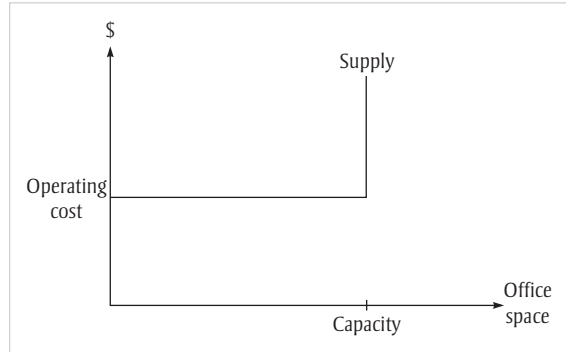
Figure 18C.1 Short-run supply curve.
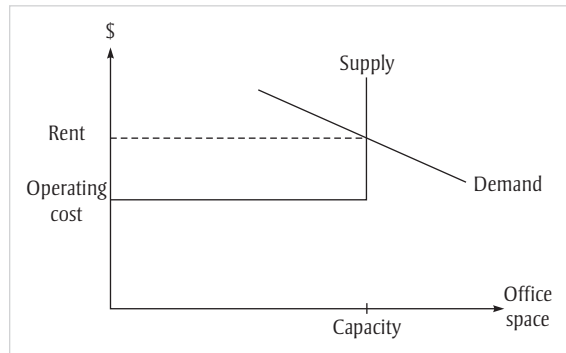
Figure 18C.2 What builders hoped would happen.

have the talent or luck to be successful developers; that's why the few who are successful are so rich.

c. Summary on Tax Abatements   So while it's possible that a regime of tax abatements could create potential Pareto improvements, information and incentives may keep them from being used effectively. They could cause more harm than good. Notice that I haven't even examined the question of whether decision makers would use good information if they had it.

Empirically, few economists have found abatement policies to do more good than harm. James White (1988) looked at small cities in upstate New
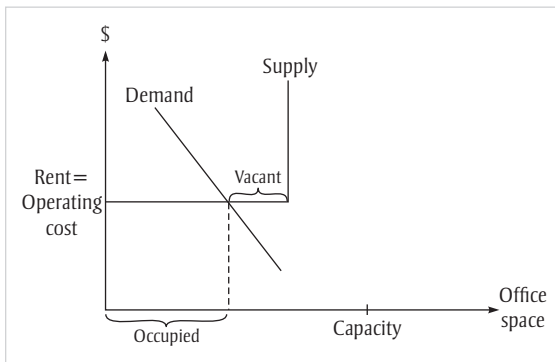
Figure 18C.3 What happened (short run).

Only if the excess of equilibrium rent over operating cost is large enough to pay for construction do builders want to go ahead with construction.

Suppose, however, that the demand curve isn't where the builders hoped it would be. Twenty years after construction, the city has declined a lot. In particular, suppose the short-run equilibrium looks like Figure 18C.3. Rents equal operating cost, and some office space is vacant.

Over time, vacant office space will be abandoned, and the city's office capacity will fall. This will keep happening until capacity falls to the level warranted by demand at operating cost, as in Figure 18C.4.

Figure 18C.4 illustrates the situation described in the text. Thus such a situation would not be unusual at all—something pretty close to it is what you would expect to see in declining cities.

Figure 18C.4 What happened after supply adjusted.

York and found that those that offered more generous abatement policies had less total capital stock and higher property tax rates: abatement displaced more investment than it stimulated. (White tried to control for the fact that towns in more desperate economic straits tended to adopt generous abatement policies.) Wassmer (1992), using Michigan data, found that commercial abatements increased commercial investment but decreased manufacturing and median home value and caused regular property tax rates to rise. Manufacturing abatements also reduced median home value, but they did encourage manufacturing investment and had no effect on regular property taxes. Anderson and Wassmer (2000, pp. 143–146) found that manufacturing abate-

ments after 1983 decreased manufacturing property value and increased local poverty and the regular property tax rate (results before 1977 were somewhat more optimistic). Commercial tax abatements worked the same way, even before 1983.

More indirectly, Anderson and Wassmer (1995) found that abatements don't work like Rauch's developer inducements to early movers. Using an abatement in one year makes a city more likely, not less likely, to use it later, and abatements grow more generous as time passes, not less generous.

Are there alternatives to tax abatements? As we've already noted, land taxes would solve cleanly all the problems that tax abatements are supposed to address. A less radical alternative is outlined in O'Flaherty (2000): selling the right to build without taxes on improvements. The idea here is to prohibit the origination of abatements on privately owned land; New Jersey law, for instance, carried this prohibition between 1961 and 1967. A property where a tax-sensitive development was being contemplated would be sold to a public agency, the taxes (except land taxes) would be stripped from it, and then it would be sold again to the ultimate developer. In the first sale the public agency would pay the value of the existing property with regular taxes; in the second sale the public agency would receive the value of the existing property with an abatement on whatever would be built there.

This difference between the value of a property with regular taxes and the value of the same property with an option to build abated is the heart of the problem with current tax abatements. It's the rent, and so it drives rent-seeking behavior. It's also one measure of the potential loss in tax revenue, and so it's a major source of the losses for other taxpayers. By trying to secure a significant portion of the efficiency gains from tax abatements for taxpayers generally, this proposed alternative attacks both problems. But it would need considerable refinement before being implemented.

3. Other Specifically Tailored Incentives

Tax abatements aren't the only way that governments have tried to provide targeted incentives. Instead of giving abatements—or in addition to giving abatements—governments sometimes construct specific infrastructure (expressway exits, for instance), or establish special free training programs, or subsidize construction of private physical plant, either directly or through industrial revenue bonds or tax increment financing authorities. Since the mechanisms are so varied, these incentive programs have generally been little studied. They present many of the same opportunities and issues as tax abatements—they would be excellent if administered by extraordinarily percipient and wonderfully altruistic beings, but they aren't.

Of these incentives, **tax increment financing** (TIF) probably deserves the most notice, both because it has been the object of some scholarly attention and because it represents an extremely opaque way of giving subsidies. In one

version, TIF centers around some physical improvement that would normally be privately constructed, because it benefits only one or a few properties. Under TIF, the government (or some governmental agency) builds this improvement, issuing bonds to pay for the construction. The improvement increases value and market activity at one or more properties, and these increases would normally result in higher tax payments from these properties (although the improvement may also result in lower tax payments from elsewhere). Under TIF, tax payments from the benefiting area are frozen where they were before the improvement was built; owners in the area pay the government no more than what they used to pay.

The fiction that the properties are still paying at normal rates, however, is maintained. The owners must take the tax revenue that otherwise would have gone to the general treasury and use it to pay off the bonds that the government issued to build the improvement. This maneuver is fiscal legerdemain. Requiring me to use $100 of my taxes for some expense like my mortgage that I would otherwise pay myself is the same thing as cutting my taxes by $100. TIF is thus just a fancy (and probably deceptive) way of giving a tax abatement.

For instance, consider an office building owner thinking about borrowing money to build a parking garage adjacent to her building for use by her tenants. This will increase the value of her property (at the very least because it will now have a garage on it), and increase her property taxes in the normal course of events.

Suppose the city government wants to encourage this construction, and can use either tax abatement or a TIF to do so. Among possible tax abatements, one of the most generous would be to exempt from property taxation the entire value of the garage and whatever enhancement to the value of the building the garage caused—at least until the garage loan was paid off. In other words, the most generous tax abatement freezes property taxes at the pre-garage level. The owner pays taxes at the old level, the city receives taxes at the old level, and the owner pays for the garage by paying back the money she borrowed.

With TIF, the city rather than the owner borrows the money to pay for the garage, and the owner's taxes go up. But the increased taxes the owner pays are used to pay off the loan. So the owner is paying the same amount she would have paid with the most generous tax abatement: the old taxes plus the amount of the loan. The city government is receiving, on net, exactly what it would have received under the generous tax abatement: the old taxes. (It receives incremental taxes from the owner but has incremental taxes in paying off the loan, and these two increments cancel each other out because they're equal.) That's why I say that TIF is just a roundabout way of giving a tax abatement.

(Not all TIF is used for facilities the developer would build privately in

the normal course of events. Sometimes TIF is used for things normally financed from general tax revenues—a street, traffic light, or sewer extension—but that the city thinks it can't afford. In this case, TIF is worse than a tax abatement, because the improvement is worth less to the developer than it costs the city to build.)

Not surprisingly, then, the empirical results on TIF are about the same as the empirical results on tax abatements. Dardia (1998) found that in California in the 1980s, small areas covered by TIF districts grew faster than areas without TIFs, but this growth was almost never fast enough to offset the revenue loss they caused. Anderson and Wassmer (2000, pp. 142–143) find the presence of an area with TIF increases commercial property values but raises property taxes. It causes a small decrease in poverty.

4. Linkages

"Linkages" is the catchphrase for efforts to tie specific economic development incentives to specific outcomes that are thought to make economic development desirable. The standard example of a linkage is a provision in a tax abatement agreement requiring the firm receiving the abatement to hire a certain number of poor people from certain neighborhoods. Linkages are often talked about but not much implemented.

Two arguments can be made for linkages. The first one, and the most frequently heard, is cynical: if the government is going to throw money away on useless tax abatements, poor people should have a place at the trough, too. The second is that linkages are needed to make sure economic development efforts accomplish the goals they are supposed to accomplish.

It's difficult to see how linkages do anything that direct subsidies to the desired ends couldn't do better. If the goal is to find good employment for poor people from a certain neighborhood (assuming housing prices won't rise fast enough to offset any benefits from good employment), then subsidizing their wages is much more straightforward and effective than subsidizing a specific firm with specific jobs in a specific location. If I were hungry, I would much prefer that you give me a few dollars to buy whatever food I wanted, rather than that you pay thousands of dollars to someone to establish a pickle farm on the condition that I be permitted to eat a few of the pickles.

5. Reducing Regulations

Many states and localities enforce regulations, especially environmental regulations, that increase the cost of doing business there for various types of firms. Relaxing some of those regulations is like cutting taxes: it encourages those firms to move in or expand, and when they do, they produce the external benefits of economic development.

Like taxes, though, environmental regulations, especially moderately in-

telligent ones, are a two-faced coin. Loosening regulations attracts some firms but it repels others. Laundries don't like to operate near incinerators, for example. Environmental problems can also harm large numbers of firms indirectly: a city with severe air pollution will have to pay high wages to attract workers to any job, and so firms of all kinds will face higher employment costs if some firms are allowed to pollute egregiously.

Empirically, there is little evidence that stricter environmental regulation discourages economic development. Many studies do show that more stringent environmental regulations have a negative impact on measures of economic development, but most find that this impact is small. Levinson (1996), for instance, estimates that a big increase in the stringency of a state's regulation—one that would move it past about ten to fifteen states on his measure—would cost the average state only about 500 start-up manufacturing jobs over a five-year period. The studies also concentrate on particular types of jobs—manufacturing, for instance—or on new plants. Thus they can't tell whether the regulations are augmenting other types of market activity because they make the environment better or because they reduce competition for land and labor from manufacturing. Tannenwald (1997) gives a comprehensive review of many of these studies.

Of course, these studies don't imply that really bad regulation couldn't stifle economic development or really brilliant regulation couldn't spur it tremendously. What they show is that variation within the range of regulation that states in this country have adopted doesn't make a big difference for economic development.

6. Urban Enterprise and Empowerment Zones

Urban enterprise and empowerment zones package together all the programs I have just discussed, and focus them on a small, heavily distressed geographic area, usually just a few neighborhoods in a larger city. Various kinds of taxes are lowered, regulations are loosened, and often some sort of linkage is required. Government authorities have a great deal of discretion in setting the zone boundaries, and sometimes have some discretion in changing them, but once the boundaries are established, benefits are available to pretty much every business that qualifies. Zones also generally don't face a budget constraint: lower tax revenues are made up for by a much bigger entity like a state or national government, not by the other taxpayers in the zone. Implicitly in the United States, urban enterprise and empowerment zones are places where minorities live, but nobody states this explicitly.

The phrase **urban enterprise zone** and the concept originated in Great Britain under Prime Minister Margaret Thatcher and migrated to the United States in the 1980s. By the early 1990s, most states had set up urban enterprise zone programs, and a federal version of the program, called "urban empower-

ment zones," was started in 1996. Over time, emphasis gradually shifted from a Thatcherite fondness for deregulation to a Clintonite fondness for linkage and social services. State programs differ widely on many important details.

The theory behind enterprise zones is that the external benefits of business activity in poor and minority neighborhoods are much greater than their external benefits in other neighborhoods. (Otherwise, dragging businesses away from the locations in which they would be most productive is just a way of creating deadweight losses.) We've already seen why this might be so—jobs might be filled by people who would otherwise be idle or less productively employed, external economies of scale might be rekindled, infrastructure might be more fully used. But in the discussion of spatial mismatch (Chapter 11), we also saw the advantages of moving people out of these neighborhoods, instead of moving jobs into them.

Since urban enterprise zones represent a packaging of the programs we've already examined, together with a restriction to a small geographical area, most of our analysis has already been done. Lower taxes should increase the amount of market activity occurring in the zones (and decrease the amount occurring outside them), with the exact form of the increase depending on the form of the tax breaks. Regulatory relaxation will help some industries and hurt others. Since the amount of land in the zones doesn't change, the simple fact that some uses will expand will force others to contract. Business expansion will raise the price of land for housing, and so raise the price of housing.

Housing prices will rise for another reason, too: if better job access (or better shopping, or more amenities) make the zones more attractive places to live, more people will want to live in them, and housing prices will be bid up until those who live in the zones are no better off than similarly skilled people who live elsewhere in the metropolitan area, or in the world. Because migration into and out of a small sliver of a city should be much faster than the migration among states and metropolitan areas that concerned, say, Bartik (1991) and Blanchard and Katz (1992), any hysteresis effects are likely to be small, and almost all benefits from the zones will probably accrue to landowners.

The empirical results on enterprise zones have been somewhat surprising: the effects have been much smaller than the studies of intrametropolitan tax differentials would lead you to expect. Some studies have found that zones spurred some kinds of business activity, but the effects were small. Papke (1993, 1994) found that inventory investments rose and unemployment fell in Indiana's urban enterprise zones, but per capita income didn't improve. Alm and Hart (1998) found mixed results in Colorado, and Engberg and Greenbaum (1997) found a small impact in moderately distressed cities nationwide. Bollinger and Ihlanfeldt (2003) found that jobs in the metropolitan Atlanta area grew faster in census tracts where an enterprise zone lowered

property taxes. On the other hand, more studies have found no positive impact whatsoever: Boarnet and Bogart (1996) for New Jersey, Dowall (1996) for California, Bondonio (1998) for five states, Engberg and Greenbaum (1997) for severely distressed cities, Greenbaum and Engberg (2004) for six states, and Peters and Fisher (2003) for thirteen states.[2] That zones seem to have little or no effect on business activity is also consistent with the finding of Greenbaum and Engberg (2002) that they have little or no effect on housing markets either.

Why haven't enterprise zones been as powerful as general reductions in taxes? One possible explanation is that many of the zones just promote churning—the replacement of old capital stock with new capital stock, as I discussed in the model of tax abatements and office buildings in a declining city. Greenbaum and Engberg (2004) note that "since many of the zone subsidies are tied to the number of new hires or the amount of new investment, new establishments will receive a much larger total subsidy than existing establishments." This explanation is consistent with their finding that enterprise zones made businesses new to the zones start up or enter and grow faster, but depressed the growth of existing businesses; the result was a wash as far as the total volume of business activity was concerned. Even when they're eligible, existing businesses may not take advantage of the tax breaks and subsidies: Dowall (1996) surveyed firms located in California zones, and found that 48 percent of them did not use the incentives available to them. This may be because incentives are marketed to new firms and firms contemplating new construction. As with tax abatements, the real goal of enterprise zones may be to make casual observers think that people's lives are getting better, rather than to make people's lives better.

7. Sports Stadiums and Arenas

Professional sports is a very small industry that receives an awful lot of attention. In 1996, average gross revenue for a team was about $75 million in football, $65 million in baseball, $55 million in basketball, and $30 million in hockey (Noll and Zimbalist 1997, p. 86). The average hospital in 1997 had $63 million in revenue (U.S. Bureau of the Census 2000, table 1303). Total revenues for all four major professional sports in 1997 were around $6 billion, less than a quarter of what Americans spent on books, half of what they spent on

---

2. More precisely, Boarnet and Bogart found that enterprise zones in New Jersey did not increase employment in the cities they were located in. On average, each enterprise zone covered about a third of its city, and Boarnet and Bogart had no data on employment within the zones. These results are compatible with three different outcomes: employment stayed the same in both the zone and the rest of the city; employment grew in the zone at the expense of the rest of the city; or employment fell in the zone and grew in the rest of the city. Greenbaum and Engberg (2004) found that zones didn't affect employment in nearby areas (or in the zone itself), and so the two studies together weakly support the first possible outcome.

sound recordings or athletic footwear (2000 Statistical Abstract, tables 421, 423, 426).

I'm writing about this particular tiny industry because many people claim that attracting and keeping major league sports franchises has an almost magical role in promoting economic development. Around $7.2 billion has been spent, directly or indirectly, on such ventures in recent years (Siegfried and Zimbalist 2000), most of it from public sources. Again, $7 billion is not a lot to invest over a decade or more—public entities spent more than $7 billion on new construction of water supply facilities in 1999 alone, and families in the South spent more than $8 billion on mobile homes (2000 Statistical Abstract, tables 1191, 1200). But much more attention has been paid to the $7 billion invested in stadiums and arenas than to the much larger amounts invested in public water supplies and southern mobile homes.

The claim that sports franchises promote economic development has two different versions, one about regions and the other about the allocation of market activity between the central city and suburbs within a region. I'll examine each in turn.

The regional claim is that stadiums (or arenas—I'll say "stadiums" when I mean both) make metropolitan areas grow faster. Getting a new franchise, in this view, causes a one-time increase in market activity and employment in a region, and this increase is sustained as long as the franchise stays. The increase comes because people spend money on tickets and related goods and services, and this spending generates jobs.

This reasoning is seriously flawed for obvious reasons. Teams can create economic development externalities within a region only to the extent that they draw money from outside the region into it—money that would not otherwise be coming in—and keep money in the region that would otherwise be going out. The local fan who goes to a ball game but would have gone bowling if the ball team were elsewhere thus makes no difference on the demand side—and that's the vast majority of fans. Neither does the visitor from out of town who would have been in the region anyway—the salesperson who would have gone to a club in the absence of a ball game, or the visiting cousin you would have taken to a restaurant. All that counts so far as bringing in money from outside the region are network broadcast revenues, the expenses of visiting teams and reporters, and those visitors to the region who would not have visited if the team hadn't been there.

On the other hand, once they receive money, from inside the region or outside of it, sports teams tend to spend more of it outside the region than regular enterprises do. Most of their revenues go to players, coaches and managers, and owners—none of whom are especially likely to live in the region year-round. Players and coaches must be on the road for half of the season and often live elsewhere in the off-season. Players save a lot of their money and invest it in worldwide markets. They also pay hefty income taxes, which

leave the local area and go to Washington. Relative to other possible recipients of money that local residents are going to spend on entertainment, professional sports teams are much more likely to send money outside the region. Bartenders, bowling alley mechanics, and waitresses don't live like professional athletes.

Thus there's no reason for thinking that a baseball team, for example, can create any economic development externalities for a region. Moreover, any public subsidy used to attract or keep the team will have to be funded from higher taxes on existing taxpayers—with the negative implications for economic development that higher taxes entail.

Serious empirical work confirms this rough, commonsense reasoning. As Siegfried and Zimbalist (2000, p. 103) summarize the literature: "Few fields of empirical economic research offer virtual unanimity of findings. Yet, independent work on the economic impact of stadiums and arenas has uniformly found that there is no statistically significant positive correlation between sports facility construction and economic development." For instance, Baade (1994) found that personal income grew between 1958 and 1987 at the same rate in thirty-six metropolitan areas that had major league sports teams as it did in twelve otherwise comparable areas that didn't. Walden (1997) concluded that having a major league sports team slowed growth, everything else being equal. Baade and Sanderson (1997) found no perceptible net increase in economic activity or employment in cities that acquired new sports teams between 1958 and 1993, and for Coates and Humphreys (1999), new stadiums and new teams actually reduce per capita income in host communities.

The empirical results are really a lot stronger than the theoretical argument about flows in and out of the region. Proponents of stadiums often claim a different route for stimulating economic development: getting a franchise makes a town a "big-league" city; it generates free publicity; it makes prospective employers feel good about the place; it promotes a favorable image. If any of these effects were large, the empirical results would have been different.

The alternative claim about sports franchises is that they are the key to revitalizing downtown business districts. Teams may have no effect on a metropolitan area, but they can redirect investment to aging downtowns and jump-start development there. Stadiums attract large crowds to the downtown. Large crowds encourage investments in restaurants, bars, and retail outlets. Recreation and retail opportunities make the area attractive for corporate offices, and for higher-income families looking for a lively downtown where they can both work and live while enjoying urban amenities.

On the other hand, stadiums take up large amounts of land—land that might be used for any number of other activities that also attract crowds. Stadiums are often idle, and so may be more of a damper on street activity than a catalyst. Moving large numbers of people into and out of a stadium at the

same time also causes congestion—much more congestion than a block of stores would generate to achieve the same number of person-hours in the downtown. (One plausible story for the recent spate of downtown stadiums is that suburban homeowners found the disamenities of stadium operation so high that their political opposition was guaranteed.)

Building stadiums also seems like a roundabout way of achieving downtown revitalization objectives. If you want restaurants, why not subsidize restaurants? If you want jobs, why not subsidize jobs? (Indeed if you want stadiums as a way of attracting crowds, why not subsidize them by paying something for every ticket sold, rather than providing capital subsidies that are the same no matter whether the teams attract crowds or not?)

At any rate, the question of whether stadiums are likely to help revitalize downtown areas is an empirical one, and the answer is probably not. Rosentraub (1997) found that population levels fell more in the downtown areas of cities with downtown stadiums than in those without, and job levels fell in both sets of cities at about the same rate. In a study with a less rigorous set of comparisons, Austrian and Rosentraub (1997) looked at Cleveland, perhaps the most famous of the urban revitalization stories, and concluded that there were about 1,000 more jobs downtown than you would expect, but that each job represented a public investment of $231,000. Since these jobs would have been located somewhere in the Cleveland metropolitan area, anyway, it's difficult to think that this sort of investment is the best way to go about revitalizing downtowns.

If stadiums produce little or no economic development benefits, does that imply that public subsidies are a bad idea? Not necessarily. Teams may provide all sorts of other external benefits: they give people something to talk about, newspapers something to write about, fans something to be proud of. (There may also be some consumer surplus—the amount that some people spend to attend games may be less than the maximum they would be willing to pay. But research [Alexander, Kern, and Neill 2000] indicates that demand for tickets is highly price elastic, and owners use a large variety of discriminatory pricing schemes to extract as much surplus as they can. Consumer surplus therefore is likely to be small.) See Box 18D for more on the popularity of stadiums.

Each of these external benefits is open to question: employers may not appreciate the time their workers spend chatting in front of the water cooler about last night's game; newspapers can print stories about ball games 200 miles away or about influenza, and we might all be better off if they printed more about influenza; and pride in the accomplishments of someone you never met, who grew up far away and lives most of the year in a different place, who works in the general vicinity of where you work but shows up there only fewer than a hundred times a year and rarely spends a whole workday there—this sort of pride is difficult to comprehend. The real questions to ask

Box 18D

Why Are Stadiums Popular?

☰ If building stadiums is such a miserable policy, why are they so popular, especially with elected officials? It can't be simply that the officials will get construction kickbacks—they can get those kickbacks from building more popular and more meritorious facilities. I can think of four possible answers.

First, local media may be important. A sports franchise gives them something to cover that national media won't. To the extent that people are interested in learning all the details about the local franchise, news media can see their audience increase, possibly at the expense of more national outlets. That sells advertising. Politicians want the media on their side, and so they support stadiums because they help local media.

Second, politicians like the perquisites that come with stadiums: the opportunities to attend events and sit in good seats at reduced prices, to hobnob with celebrities, to be recognized and applauded by fans. A water treatment plant may be a better investment, but it's not something you can really enjoy. Since reporters and media executives also like to attend sports events at discount prices, important media individuals may also enjoy the stadiums. This reinforces media support for stadiums.

Third, voters may systematically underestimate the size of subsidies and interpret a stadium as a sign of confidence in the city. If the stadium is there, they may think, the city is really doing well, and that means the elected officials are doing a good job.

Finally, pride, however misguided, may be important.

---

in the study of sports stadiums would be about the size and reality of these sorts of external benefits, not about economic development effects.[3] These are difficult questions to answer. And since this chapter is about economic development, not sports, I won't pursue them further.

8. Summary of Economic Development Tools

Economists don't know of any miraculous ways to spur economic development. Neither do other people. The only difference is that economists know that they don't know. We do know what retards economic development, how-

---

3. The evidence on the consumption externalities of stadiums is meager and conflicting. Carlino and Coulson (2004) find big benefits from National Football League teams—rents go up and wages go down when a team moves to town. Johnson, Groothuis, and Whitehead (2001) found in their surveys that people in Pittsburgh were not willing to pay very much to have the local hockey team around. The difference in the results of these two studies could be due either to different methods, or to different values that people place on hockey and football.

ever: lousy and expensive public services, bizarre and draconian regulatory policies, high taxes, corruption, crime and ethnic conflict, poor and capricious law enforcement, fear of expropriation.

What this review of economic development tools has shown is that governments can't do an otherwise bad job and make up for it by running great economic development programs—because there is no such thing as a great economic development program by itself.

The difficulties we have had in this section in identifying economic development policies, as opposed to policies wise or unwise on traditional cost-benefit grounds, reinforce this general advice. Section I.A showed that economic development externalities are real, because many markets don't work perfectly, but they are not huge, especially in the long run. Section I.B showed that, among otherwise reasonably wise policies, the size of economic development externalities doesn't vary much. It follows that differences in the size of economic development externalities are not often going to be decisive in choosing one policy or another. Making economic development the last chapter in this book was a deliberate decision on my part.

One obvious consequence of this analysis is that economic development—as the term has come to be used in the United States—is not a good way of improving the lives of poor people. The benefits of these programs don't generally accrue to poor people, and often the benefits are small. "Helping poor places" is not a good way of helping poor people, if only because often the best course of action for poor people is to leave poor places (almost everyone reading this book is wildly affluent by world standards, a gift we owe to whatever ancestors of ours left the Old Country or the rural South, if we didn't do it ourselves).

This is not to say that poverty is inescapable. Eating chocolate cream pie is good for some purposes, but not for cleaning carpets. The relationship between chocolate cream pie and carpet cleaning is like the relationship between economic development and poverty alleviation.

Which doesn't imply that carpets can't be cleaned. Carpet cleaning, like poverty alleviation, is just a separate topic. There are many good policies to alleviate poverty—the vast reduction in worldwide poverty over the past two centuries attests to that—and no political ideology seems to have a monopoly on these good ideas. (There are also a lot of bad ideas, and there's no ideological monopoly on them, either.) It just doesn't make a lot of sense to go looking for these good poverty-alleviation policies in the catalog of U.S.-style economic development policies, any more than it makes sense to go looking for carpet cleaner in a bakery.

What about the Erie Canal and other great episodes of economic development? The Erie Canal was good because it was a potential Pareto improvement, not because it brought a lot of jobs to New York City. By hugely reducing the cost of transporting goods to and from the Midwest (both American

and Canadian), the Erie Canal produced great benefits for huge numbers of people throughout the world.

If the leaders of New York State had thought only about the immediate benefits to people living in New York State at the time, however, they might not have gone ahead with the project. The positive role that the idea of economic development sometimes plays in public debates is that it introduces in concrete form the concerns of people who aren't there yet and may never be part of the jurisdiction making a decision. Economic development counteracts parochialism. Thus it sometimes leads government officials to make the right decisions for the wrong reasons, which is good. But it's better to make the right decisions for the right reasons.

II. Economic Development in Many Cities

Section I asked what sorts of economic development policies a wise city or state government might want to adopt—a government that cared about people living in its jurisdiction, not those living elsewhere. We didn't ask how policies in city A affect residents in city B, or vice versa. But surely such interactions occur: when North Carolina's economic development policies lure a pharmaceutical plant from New Jersey, New Jerseyans are affected, and when a Los Angeles movie studio decides to expand its operations, there's less work for New York filmmakers.

These interactions should concern a benevolent government whose citizenry includes both North Carolina and New Jersey, or both Los Angeles and New York. This section asks about policies for that higher-level government: how should it think about the losses that people in New Jersey and New York suffer? Should it encourage lower-level governments to pursue aggressive economic development policies? Should it discourage them, or remain neutral? What does neutrality mean? When we talk about higher-level governments in this section, we'll be concerned both with the federal government, in relation to states and cities, and to state governments, in relation to cities.

Questions like these are heavily debated. One camp wants strict federal controls on economic development activity. In its view, North Carolina's stealing jobs from New Jersey is not much different from any other kind of theft: North Carolina is imposing a negative externality and should not be permitted to do so. The other camp thinks that North Carolina and New Jersey are engaging in healthy competition from which will emerge a stronger nation; that the more state and local governments are forced to innovate, to hold down taxes, and to produce high-value public services, and the less they are able to exploit the businesses in their midst, the better off we'll be. This camp opposes federal controls on economic development activity.

Any discussion of these competing points of view, of course, presumes

some clear definition of what economic development activity is (and is not). We saw in section I how difficult it is to decide what activities constitute economic development, and this difficulty will continue here. If you want to prohibit cities from competing for businesses by offering more attractive tax abatement packages, do you also want to prohibit them from competing by offering better police protection, swifter approval of permits, and lower general tax rates?

Because interjurisdictional competition is a difficult issue, we'll approach it the same way we have approached all difficult issues in this book—we will pare it down into a much simpler problem by making some unrealistic assumptions, come to a good understanding of that simple problem, and then explore what happens when we drop the unrealistic assumptions.

A. The Marriage Market

One way to look at the relationships between cities and firms is to compare them with the marriage market. Just as marriages link men and women, locations link cities and businesses. Different businesses bring different costs and benefits to different cities; different cities offer different opportunities and problems for different businesses. In marriages, it's important to bring together women and men who like each other; in economics, you want to make sure the shipyards don't end up in Oklahoma City and the slaughterhouses aren't in midtown Manhattan.

Let's assume (for now) that each city can accommodate at most one business and each business can locate in at most one city. Some cities can end up without any businesses—ghost towns, and cities that never existed—and some businesses can end up without any cities—businesses that never got started. It is just like having some men or some women end up being single. Notice that in this formulation, benefits and costs are general—the benefits cities get can include economic development benefits, traditional externalities (both positive or negative), and even intangibles like pride; the benefits firms get include the usual advantages of location and climate, as well as anything more idiosyncratic, like the opportunities the CEO might have to visit her grandchildren. But I want to leave out of these costs and benefits the taxes cities collect from firms and the subsidies they pay them—I want to consider taxes and subsidies separately as outcomes of bargaining, not as inputs to the process.

What happens in a marriage market if women and men are left to bargain on their own? Under some circumstances, which I'll discuss shortly, the outcome will be Pareto optimal. Men and women will end up paired with each other in a way that maximizes total benefits realized. Any other system of matches can be broken up by someone who will promise enough to lure a more preferred partner out of whatever match he or she is in. This is a theory

of the marriage market first developed by University of Chicago economist and Nobel Prize winner Gary Becker (1981).

Intuitively, marriages in Becker's system break up when one of the partners is worth more to a potential partner outside the marriage than to the partner inside the marriage. The breakup increases total benefits. Only when total benefits are as great as they can be will no further breakups be possible.

The equilibrium is efficient, but it's not without heartbreak. When Iseult runs off with Tristan, Mark's heart is broken and he must settle for an inferior partner. Is this an externality that calls for government intervention? No—because Tristan must pay for the heartbreak he inflicts on Mark. He doesn't pay Mark, though; he pays Iseult instead. The more Mark is willing to do for Iseult, the more Tristan must do for her in order to keep her away from Mark. Remember that what's important for efficiency is that the person causing damage pays for it, not that the person suffering it be compensated. This is just like an auction: the winning bidder pays the auctioneer, not the losing bidders, but the amount the winner pays depends on how much his winning harms the losers. The pain that jilted and unrequited lovers feel is both real and great, but it's no cause for government intervention.

Similarly, when North Carolina bids a pharmaceutical plant away from New Jersey, or when Phoenix bids a football team away from St. Louis, there's no reason for federal government intervention. The business moves only if the total benefits of the new match are greater than the total benefits of the old; otherwise the old city would have been able to mount a successful counteroffer. The winning city pays for the harm it causes to the losing city—but it pays this, one way or another, to the business, not to the losing city (see Box 18E).

The conclusion about no cause for intervention, however, has merit only to the extent that interjurisdictional competition for business is really like the Becker marriage market. (A similar question arises, too, about whether courtship and marriage in Western countries is really like the Becker marriage market, but since this chapter isn't about marriage, I'll ignore it.) One assumption we made—that each city could attract at most one business—wasn't really essential: we can think about a bunch of sites rather than a bunch of cities, with the benefits realized by the city in which a site is located, and the decisions made on behalf of that site by that city.

Two other assumptions, though, are much more crucial. The first is that cities and businesses actually know and want to act upon the true benefits of every possible match. We call this the perfect agency assumption—that city governments, especially, act as perfect, selfless, knowledgeable agents of their citizens. The last section gave us considerable reason to doubt that cities actually work like this. The second crucial assumption is that cities and businesses can credibly promise to pay each other arbitrary amounts of money just to make a deal happen. We call this the perfect flexibility assumption. Since city governments are supposed to act according to laws, and because

Box 18E

See for Yourself

≡ Suppose there are two cities, Princeton and Camden, and two firms, Campbell Soup and Johnson and Johnson (J&J). Each firm is going to build a single plant, and neither city has room for more than one plant. The cities evaluate the benefits from each firm like this, excluding taxes and other special payments:

	Campbell	J&J
Evaluation by Princeton	$4	$7
Evaluation by Camden	$3	$5

The firms evaluate the profits they could make in the two cities like this:

	Evaluation by Campbell	**Evaluation by J&J**
Princeton	$10	$9
Camden	$5	$3

Find the stable matching between firms and cities. (A stable matching is one that no other pair can change to their mutual advantage.) What payments have to be made to sustain this matching?

Step 1: Find the matching that maximizes total benefits. There are two matchings:

Princeton with Campbell and Camden with J&J

and

Princeton with J&J and Camden with Campbell.

For brevity, I'll sometimes refer to the first matching as the Campbell matching and the second as the J&J matching. Princeton is the first city in each matching, and once we know which firm is in Princeton, everything else follows.

We can summarize the information about preferences by constructing a chart showing the total benefits to both partners for each possible pairing:

	Campbell	**J&J**
Princeton	4 + 10 = 14	7 + 9 = 16
Camden	3 + 5 = 8	5 + 3 = 8

Then the grand total benefits under the Campbell matching are the sum along the main diagonal (northwest to southeast): $14 + 8 = 22$. Grand total benefits under the J&J matching are along the other diagonal (northeast to southwest): $16 + 8 = 24$.

Step 2: Show that the Campbell matching is not stable. Suppose we try to operate under the Campbell matching. Then Princeton could approach J&J and say: "Why don't you move here? I'll get rid of Campbell. Right now, I'm only getting $4 in benefits and you're getting $3. If we got together, you'd get $7 and I'd get $9. We'd both be better off."

Can Campbell and Camden come up with counteroffers to quash this deal? In a nutshell, the answer is no. Princeton and J&J gain together $9 from replacing the Campbell matching with the J&J matching. Camden and Campbell together are only willing to pay $7 to stop it. They won't stop it.

Step 3: Find the range of tax or subsidy payments that have to be made to sustain the stable matching. Two different attacks against the J&J matching could be mounted: either Princeton and Campbell could try to get together, or Camden and J&J. A system of payments has to be set to repel both attacks.

Princeton and Campbell, if they got together, would realize total benefits of $14. So the J&J matching will be stable only if together Princeton and Campbell realize at least $14 in benefits. Let p denote Princeton's net benefit—the intrinsic value of the match plus any taxes it receives or minus any subsidy it pays—and let c denote Camden's net benefit. Then Campbell's net benefit is $(8 - c)$ and together in the J&J matching Princeton and Campbell realize

$$8 - c + p.$$

So the condition for stability is

$$8 - c + p \geq 14,$$

or

$$p - c \geq 6.$$

Similarly, if Camden and J&J got together they would realize benefits of $8. Camden's net benefits are c and J&J's net benefits are $(16 - p)$. So the other condition for stability is

$$c + 16 - p \geq 8,$$

or

$$p - c \leq 8.$$

In addition, Camden's benefit cannot exceed $8 and Princeton's benefit cannot ex-

ceed $16. This is because we rule out negative benefits for the companies (they could just shut down). Together these conditions give the pairs of benefits to Princeton and Camden that make the matching stable. These are illustrated in Figure 18E.1.

The parallelogram in this graph gives the combinations of net benefits that induce stability. To find the taxes or subsidies implied, just subtract $5 from Princeton's benefit and $3 from Camden's. Unless we know something more about the bargaining process or the other options the agents have, we can't specify more precisely what payments will be made.

Step 4: What happens if firms and cities always split the range of benefits down the middle? This is a simple assumption about bargaining that will allow us to compare different policies more definitely. Splitting the difference results in the benefits associated

Figure 18E.1 Possible net benefits under the stable matching.

with the point in the middle of the parallelogram. Camden's benefits lie between zero and $8; so the split gets Camden $4 in benefits. Campbell also gets $4. That means Campbell pays Camden taxes of $1. With Camden's benefits equal to $4, Princeton's benefits lie between $10 and $12. Splitting the difference gives Princeton $11; it collects $6 in taxes from J&J. J&J's benefits are $5.

governments in many cases can't be bound by their predecessors, there's also reason to doubt how applicable the perfect flexibility assumption is.

But I don't want to take up immediately an examination of these two assumptions and the consequences for marriage markets if one or both don't hold. Instead, I want to look next at what it would imply for policies of the federal government if they did hold. That way, we'll be able to understand how important these assumptions are for policies.

B. Perfect Agency and Perfect Flexibility

The obvious policy implication of the perfect agency and perfect flexibility assumptions is that the federal government should not interfere in economic development competition among states and cities. The outcome without federal (or state) interference will always be Pareto optimal, and so federal in-

volvement can't make things better. The cure would be worse than the illness, because the patient is otherwise perfectly healthy. However, it's worthwhile considering the details of how various putative cures harm the patient.

1. Helping Poor Places

Quite a few serious writers have argued that the federal government should aid the economic development efforts of places with poor and unemployed populations, but not those of richer places. (Bartik 1991 and Anderson and Wassmer 2000, for example, have espoused this view.) The argument is that economic development externalities are greater in poor localities—which may be accurate, as we saw in section I.A. Several federal programs, such as Urban Development Action Grants and parts of the Economic Development Administration, operate in this way. States have analogous programs for their poorer cities.

But if poor cities benefit more from economic development than rich cities do, they should be willing to bid more for business. With perfect agency and perfect flexibility, any greater benefits to poor cities are already incorporated in the bargaining process. Subsidizing businesses in poor cities only makes poor cities more similar to rich cities in their bidding for business. This gives businesses in other cities a better bargaining position. It might also help poor cities, but it might not, and it could lead to deadweight losses by inducing firms to locate in the wrong places.

How? Consider first a general subsidy: a federal subsidy of the same amount to any firm that locates in a poor city. A general subsidy won't change locational patterns. The matching of firms to cities that maximized total benefits before the federal subsidy still maximizes total benefits with the subsidy. Any system of matches includes the poor city exactly once, and so the grand total benefits of each system of matches increases by the amount of the subsidy; the ranking stays the same. The businesses in the poor city stay the same.

The threats and counteroffers that govern the distribution of benefits between cities and firms change, however. In the poor city, there's more for everyone, so we would expect both the city and the firm that locates there to gain something—and we have no reason to think that the city gains the whole amount of the subsidy. In the other cities, firms are in better bargaining position because their threats to leave and go to the poor city are more credible. Hence firms in nonpoor cities gain at the expense of those cities. For an example of the effect of subsidies, see Box 18F.

(Since firms gain generally, and stock ownership is usually concentrated in richer regions, subsidies for firms in poor regions can end up benefiting rich regions more than poor; see Dupont and Martin 2003.)

Instead of a general subsidy, the federal government could give a specific subsidy—a subsidy targeting only certain kinds of businesses (for instance, new or high-tech firms). Such a subsidy may or may not alter locational pat-

See for Yourself

☰ Let's take another look at the example described in Box 18E. This time, what happens when the federal government gives a subsidy of G dollars to any firm that locates in Camden?

Step 1: Revise the total benefits table and find the stable matching. If any firm locating in Camden gets G dollars from the federal government, the total benefits table looks like this:

	Campbell	J&J
Princeton	14	16
Camden	$8 + G$	$8 + G$

So the main diagonal, what we call the Campbell matching, has grand total benefits of $(22 + G)$, and the other diagonal, the J&J matching, has grand total benefits of $(24 + G)$. No matter how big or how small the federal subsidy is, the matching that has greater total benefits does not change.

Step 2: Find the parallelogram of city benefits. As before, consider the two challenges to the J&J matching. If Princeton and Campbell get together, their benefits are still $14. Princeton's benefit is p, and Campbell's benefit is $(8 + G - c)$, and so one condition for stability is

$$8 + G - c + p \geq 14$$

or

$$p - c \geq 6 - G.$$

terns. If it's not big enough to alter locational patterns, then its only effect is to change bargaining in existing matches. Firms of the targeted type get more out of their existing partners, and those partners—cities otherwise hosting the targeted firms—lose out. Since a subsidy that's not big enough to alter locations doesn't make anything happen in the poor city, the poor city is unlikely to gain.

If the specific subsidy is large enough, it can alter locational patterns. Of course, this implies a deadweight loss, since the pattern without subsidies maximized total benefits. As with general subsidies, the poor city gains, since its firm is now subsidized, but is unlikely to gain the full amount of the sub-

If Camden and J&J get together, they realize benefits of $(8 + G)$. Camden's benefits are c and J&J's are $(16 - p)$. Thus the stability condition is

$$16 - p + c \geq 8 + G$$

or

$$p - c \leq 8 - G.$$

In addition, Princeton's benefits cannot exceed 16, and Camden's cannot exceed $(8 + G)$. Figure 18F.1 illustrates the resulting parallelogram.

Step 3: What happens if firms and cities split benefits in the bargaining range evenly?

Camden's benefits lie between zero and $(8 + G)$. Thus, with splitting the difference, Camden gets $(4 + G/2)$. Half of the subsidy goes to Camden, half to Campbell. With Camden's benefits at $(4 + G/2)$, Princeton's benefits lie between

$$6 + c - G = 6 + 4 + G/2 - G$$
$$= 10 - G/2$$

and

$$8 + c - G = 8 + 4 + G/2 - G$$
$$= 12 - G/2.$$
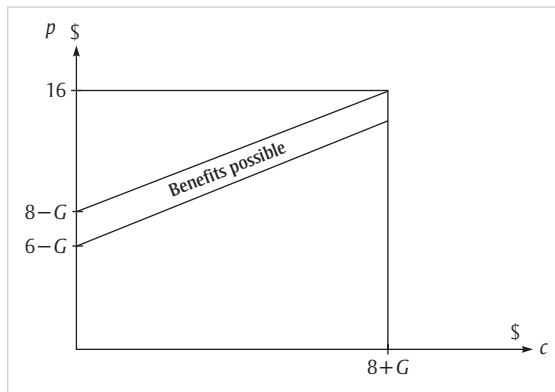
Figure 18F.1 Possible net benefits under the stable matching with a federal subsidy.

Splitting the difference gives Princeton benefits of $(11 - G/2)$ and J&J benefits of $(5 + G/2)$. Thus J&J gains half the subsidy from its threat to move to Camden, and Princeton loses half.

sidy. The firm that moves also gains—otherwise it would not have moved—and its gains come at the expense of the poor city and the federal government. Once the locational pattern has changed, additional subsidies don't necessarily affect the nonpoor cities (unless they still contain some targeted firms). In short, if the federal government wants to help cities, subsidizing economic development in them is a bad way to do it.

b. Industrial Revenue Bonds The federal income tax exempts interest from state and local bonds from taxation. This allows state and local governments to borrow at lower rates than ordinary corporations. Under some circum-

stances (describing exactly which circumstances would require another volume), states and cities can borrow money to build a plant for a business, and then get paid by the business. The bonds that states and cities use to do this are called **industrial revenue bonds.** They're a roundabout way for some firms to get the federal government to lower their interest costs.

If this policy has an impact on industrial location, it creates deadweight loss by moving businesses to where they shouldn't be. If not, it just subsidizes businesses, and may harm jurisdictions that refrain from using these bonds. Since industrial revenue bonds can be used even by rich states and cities, there's no reason to think that the policy helps poor cities. But the policy lets individual cities help individual firms at what appears to be no cost to the city, and so it is an immensely popular approach, despite its almost total lack of merit.

Getting rid of a federal policy like this presents another problem as well: how to define the objectionable activity. You can't just say, "No tax-exempt bonds for economic development purposes," because it's almost impossible to say what economic development activities are. What do you say about a police precinct house built across the street from a new office building? Or an enhanced sewer system in a town with a paper plant?

The history of sports stadiums is instructive on this point. In 1986, concerned about implicit federal subsidies because cities had been building stadiums with industrial revenue bonds, Congress removed the tax exemption from stadium bonds when cities planned to use ticket revenues to pay off more than 10 percent of the money. Cities reacted by reducing other forms of subsidies to teams, and by paying off stadium bonds without any ticket revenue. So the implicit federal subsidy for stadiums continued.

Thus, ending implicit federal subsidies for some economic development projects would probably be difficult, so long as "normal" state and city bonds are tax exempt. However, there don't seem to be any particularly compelling reasons why these bonds should be tax exempt. See, for instance, Zimmerman (1991) for a more detailed analysis of this issue.

C. Imperfect Agency and Imperfect Flexibility

Any argument for higher-level involvement in state and local economic development has to start from a rejection of either the perfect agency or the perfect flexibility assumption. In this section, we'll try to figure out whether plausible alternatives to these assumptions lead to any definite policy recommendations.

Consider first imperfect agency, the assumption that city officials don't always act in the best interest of their constituents. Many forces might lead government officials astray: ignorance, corruption, desire to get good media coverage, laziness. They may be systematically too generous to large, visible,

and new businesses; systematically too tough on small, invisible, and old businesses—or maybe not; no definitive empirical evidence exists. It has been found, however, that counties that win bidding wars for large manufacturing plants seem to do better than those that lose (Greenstone and Moretti 2003); this suggests that city officials operating under current laws are somewhat responsible.

When one city's officials misperceive the benefits particular businesses would bring, their own constituents are not the only ones who lose out. Their higher bids, even if not successful, can cause other cities to give up more to retain the businesses they are bidding for. If their higher bids are successful, they create an inefficient system of matches. In the example of Camden and Princeton, for instance (Boxes 18E and 18F), if Camden officials overvalued all businesses by $1, the location pattern wouldn't change and nothing would be different in Camden, but Princeton would have to reduce its tax collection by $1. If the Camden officials overvalued only J&J by $3, they would force the replacement of the efficient system of matches with the inefficient one.

Since everyone can lose when city officials are poor agents, some upper-level intervention might be justified. The analogy with marriage markets comes in handy here: because we believe that people are not always good agents of their own well-being in their love lives, governments restrict in many ways the marriages people can enter into. You can't marry a cousin, a child, a dog, or a married person, even if you're desperately in love and the other party wants to marry you. Most states also require waiting periods, both for marriage and divorce, and some require blood tests. Even though governments generally don't intervene in marriage markets, they do stop actions they think are serious mistakes.

Are there analogous steps that higher-level governments take to prevent lower-level governments from making huge mistakes? Some are obvious: laws against corruption, public-notice requirements, waiting periods; in general, procedural due process is imposed before a government body can be either extremely generous or extremely mean to a business or a person. Most of these are state rules binding cities, not federal rules binding states. These rules don't prevent all mistakes, and they stop some deals that in fact would be good. The case that's made for these restrictions is that, on average, they prevent more bad deals than good (and so it's ultimately empirical). Some guys really would be happiest married to their dog or their cousin, and the other party would be happy, too.

The other way higher-level governments approach the agency problem is by limiting the flexibility of city officials. Imperfect flexibility is basically a consequence of imperfect agency, a way of ameliorating the problems it causes. Thus taxes have to be imposed with some degree of uniformity; lump-sum subsidies can't be given willy-nilly; legislation has to address classes, not individuals. (Again, these are usually state rules binding cities.) A lot of the re-

strictions on flexibility can be subverted by sufficiently clever legal minds, but sufficiently clever legal minds are expensive, and so the laws form a significant disincentive to flexibility.

Losing flexibility, of course, is a problem. The economic-development marriage market leads to an optimal assignment of businesses to cities only if cities can make appropriate offers and counteroffers, and promise transfers or taxes within sometimes narrow ranges. They can't do this if they are bound up by restrictions like uniformity and generality. It's difficult to tell, though, what biases this lack of flexibility creates.

So it appears that we face a dilemma: higher-level governments can either leave city officials alone and be content with letting those officials make lots of mistakes of their own volition; or they can bind them with rules so that the mistakes are premade. In reality, though, the choice is not quite so stark.

The way to improve the trade-off between mistakes caused by discretion and mistakes caused by rules is to improve considerably the quality of rules. If operating according to uniform rules means that a city must charge every firm an average share of some huge general overhead cost, unrelated to either the costs or the benefits that a firm brings to the city, then operating according to uniform rules will cause enormous inefficiencies. A supermarket that charged everyone who entered the same price, no matter what they left with, would do poorly, and its manager would soon start cutting side deals with customers and potential customers.

But operating according to uniform rules doesn't have to mean charging average shares of general overhead, any more than operating a supermarket has to mean charging the same price for every shopper. The more money a city gets from land taxes, the less it has to raise through general uniform taxes on businesses and households. If businesses and the people associated with them pay marginal cost for the services they use and the negative externalities they cause, and if they're subsidized for the positive externalities they create (including employment), then there's a lot less about the relationship between the city and the business that has to be the subject of discretion and back-room negotiation. Rules can be reasonably finely tuned to marginal costs and marginal benefits and still be uniform. In supermarket management, careful pricing of particular items is the preferred alternative to the idiocy of one price for every shopper. Individual discretionary haggling is not the preferred alternative.

Marginal-cost prices and marginal-benefit subsidies have two other benefits for economic development: they provide better incentives for businesses to operate well once they've arrived in a city, and they affect all businesses, not just those that are big enough or loud enough to attract officials' attention. Probably the most productive intervention that higher-level governments can engage in is to permit, encourage, or require lower-level governments to establish good systems for taxes and subsidies. Like congestion pricing, these systems may not have been technically feasible fifty or even twenty years ago, but

the progress of information technology has gone a long way to reducing feasibility as a problem.

III. Conclusion

Economic development is a worthy goal for cities to pursue. Economic development externalities are real, even though they've never been measured well and are surely not so large as popularly believed. Economic development is not a good way to fight poverty, but many other worthy activities are not good ways of fighting poverty either. The tools that cities commonly employ to foster economic development are not particularly effective, and some are spectacularly counterproductive, but some policies are better than others—and most of the good ones can be identified without reference to economic development.

There don't seem to be any particularly strong arguments for upper-level governments either to restrain or to encourage the economic development activities of cities. Prohibiting interjurisdictional job stealing seems as hopeless and as unproductive as prohibiting the alienation of affections of unmarried lovers (or even married ones). Encouraging economic development activity, even on a geographically selective basis, doesn't seem to have much to commend it either. Upper-level governments should help lower-level governments use better tax systems, systems that impose smaller deadweight losses, but you don't need to study economic development to come to this conclusion.

In the final analysis, economic development is a lot like sleep. Sleep is great, but it's not a panacea. Moreover, it's a goal best pursued by not thinking about it. If you stay healthy and active and don't drink too much caffeine (and you curl up with a book like this), you'll fall asleep on your own. But if you think about whether you're falling asleep and you worry about it, you'll never get there.

Questions

1. Using the example in Boxes 18E and 18F, work through the implications of a specific subsidy to J&J if it locates in Camden.
   a. How big does the subsidy have to be to induce J&J to move to Camden? Call this the critical value.
   b. If the subsidy is less than the critical value, construct the parallelogram of benefits required for stability.
   c. Find the midpoint. Who gains as the subsidy rises? Who loses? Who is unaffected?

d. Suppose the subsidy is greater than the critical value. Construct the region of benefits required for stability. (*Hint:* It's not going to be a parallelogram this time.) Princeton's benefits cannot exceed $14. Draw this line on the diagram and note that it places a different upper bound on the benefits that Camden can get.

e. Find the midpoint. Who gains as the subsidy rises? Who loses? Who is unaffected?

2. In the same example (without federal subsidies), how do the benefits that Princeton realizes change as the benefits that Camden would enjoy if J&J located there change by a small amount? Discuss the contention that, in luring J&J away from Camden, Princeton is imposing an external cost that should be corrected.

3. Outline how you would perform a serious cost-benefit analysis of holding the 2016 Summer Olympics in Chicago. Who gains? Who loses? How much? Can the winners compensate the losers?

4. Legal pot in Quonset, Rhode Island? Ever since Quonset seceded from the United States and joined the United Nations, this issue has been intensely debated. A group of businesspeople has proposed that marijuana consumption be legalized in Quonset (exports would be prohibited). "It will be good for business and good for Quonset," they say. "Tourists will come from all over the world; they'll stay in our hotels and eat in our restaurants; they'll buy snacks and souvenirs in our shops; they'll use water and electricity; and they'll pay all sorts of taxes."

a. Consider the following three markets: marijuana (which would be produced at constant marginal cost); Quonset souvenirs (produced at constant marginal cost and sold in a competitive market); and electricity (produced by a regulated monopoly and sold at average cost, which is greater than marginal cost). In each market, identify the winners (if any) and losers (if any) from legalization. In which markets is the gain of the winners more than sufficient to compensate the losers for their losses?

b. Assume the following values. Before legalization, the price and marginal cost of marijuana in Quonset was 10 Quonset-dollars. After legalization, it will be 5 Quonset-dollars. The annual demand for marijuana in Quonset is

$$m = 1{,}000 - 50p_m - 200p_L,$$

where m is the quantity of marijuana demanded in grams, p_m is the price of marijuana per gram in Quonset dollars, and p_L is the price of Quonset land per hectare in Quonset dollars. The quantity of land in Quonset is 5,000 hectares. The annual demand for land in Quonset is

$$L = 10{,}000 - 200p_m - 3{,}000p_L.$$

Who gains from legalization? How much?

c. The U.S. government offers to pay Quonset 2,500 Quonset-dollars per year if it doesn't legalize marijuana. Should the government of Quonset accept this offer? Why or why not? If it accepts the offer, what should it do with the money?

5. How, if at all, would this chapter be different if it were about cities in developing countries rather than cities in the United States?

References

Alexander, Donald L., William Kern, and Jon Neill. 2000. "Valuing the Consumption Benefits from Professional Sports Franchises." *Journal of Urban Economics* 48(2):321–337.

Alm, James, and Julie Ann Hart. 1998. "Enterprise Zones and Economic Development in Colorado." Paper presented at the Allied Social Sciences Associations annual meeting, Chicago.

Anderson, John E., and Robert W. Wassmer. 1995. "The Decision to 'Bid for Business': Municipal Behavior in Granting Tax Abatements." *Regional Science and Urban Economics* 25:739–757.

———. 2000. *Bidding for Business: The Efficacy of Local Economic Development Incentives in a Metropolitan Area.* Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

Austrian, Ziona, and Mark S. Rosentraub. 1997. "Cleveland's Gateway to the Future." In Roger Noll and Andrew Zimbalist, eds., *Sports, Jobs, and Taxes: The Economic Impact of Sports Teams and Stadiums,* pp. 355–384. Washington, DC: Brookings Institution.

Baade, Robert. 1994. "Stadiums, Professional Sports, and Economic Development: Assessing the Reality." Heartland Institute Policy Study, April 4.

Baade, Robert, and Allen Sanderson. 1997. "The Employment Effects of Teams and Sports Facilities." In Roger Noll and Andrew Zimbalist, eds., *Sports, Jobs, and Taxes: The Economic Impact of Sports Teams and Stadiums,* pp. 92–118. Washington, DC: Brookings Institution.

Bartik, Timothy. 1991. *Who Benefits from State and Local Economic Development Policies.* Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

———. 1993. "Economic Development and Black Economic Success." Upjohn Institute Technical Report 93-001, Kalamazoo, MI.

Becker, Gary S. 1981. *A Treatise on the Family.* Cambridge, MA: Harvard University Press.

Blanchard, Olivier Jean, and Lawrence F. Katz. 1992. "Regional Evolutions." *Brookings Papers on Economic Activity* 1:1–61.

Boarnet, Marlon G., and William T. Bogart. 1996. "Enterprise Zones and Employment: Evidence from New Jersey." *Journal of Urban Economics* 40:198–215.

Bollinger, Christopher R., and Keith R. Ihlanfeldt. 2003. "The Intra-Urban Spatial Distribution of Employment: Which Government Interventions Make a Difference?" *Journal of Urban Economics* 53:396–412.

Bondonio, Daniele. 1998. "States' Enterprise Zone Policies and Local Employment: What Lessons Can Be Learned?" Paper presented at the Association for Public Policy and Management Annual Research Conference, New York.

Bound, John, and Harry Holzer. 1993. "Industrial Shifts, Skills Levels, and the Labor Market for Black and White Males." *Review of Economics and Statistics* 75(3):1258–1264.

Carlino, Gerald, and N. Edward Coulson. 2004. "Compensating Differentials and the Social Benefits of the NFL." *Journal of Urban Economics* 56(1):25–50.

Coates, Dennis, and Brad Humphreys. 1999. "The Growth Effect of Sports Franchises, Stadia, and Arenas." *Journal of Policy Analysis and Management* 14:601–624.

Courant, Paul N. 1994. "How Would You Know a Good Economic Development Policy If You Tripped Over One? Hint: Don't Just Count Jobs." *National Tax Journal* 47:863–881.

Dardia, Michael. 1998. *Subsidizing Redevelopment in California.* San Francisco: Public Policy Institute of California.

Dowall, David E. 1996. "An Evaluation of California's Enterprise Zone Programs." *Economic Development Quarterly* 10:352–368.

Dupont, Vincent, and Philippe Martin. 2003. "Subsidies to Poor Regions and Inequalities: Some Unpleasant Arithmetic." Centre for Economic Policy Research, CEPR Discussion Paper no. 4107, London.

Engberg, John, and Robert Greenbaum. 1997. "The Impact of State Enterprise Zones on Urban Housing Markets." Paper presented at the American Real Estate and Urban Economics annual midyear meetings, Washington, DC.

Fisher, Ronald. 1997. "The Effects of State and Local Public Services on Economic Development." *New England Economic Review,* March/April:53–77.

Freeman, Richard. 1991. "Employment and Earnings of Disadvantaged Young Men in a Labor Shortage Economy." In Christopher Jencks and Peter Peterson, eds., *The Urban Underclass,* pp. 103–121. Washington, DC: Brookings Institution.

Frey, William, and Alden Speare, Jr. 1988. *Regional and Metropolitan Growth and Decline in the United States.* New York: Russell Sage Foundation.

Greenbaum, Robert, and John Engberg. 2002. "An Evaluation of State Enterprise Zone Policies: Measuring the Impact on Urban Housing Market Outcomes." In F. S. Redburn and T. F. Burns, eds., *Public Policies for Distressed Communities Revisited.* Latham, MD: Lexington Books.

———. 2004. "The Impact of State Enterprise Zones on Urban Manufacturing Establishments." *Journal of Policy Analysis and Management* 23(2):315–340.

Greenstone, Michael, and Enrico Moretti. 2003. "Bidding for Industrial Plants: Does Winning a 'Million Dollar Plant' Increase Welfare?" National Bureau of Economic Research Working Paper 9844.

Helms, L. Jay. 1985. "The Effect of State and Local Taxes on Economic Growth: A Time-Series Cross-Section Approach." *Review of Economics and Statistics* 67:574–582.

Ihlanfeldt, Keith. 1992. *Job Accessibility and the Employment and School Enrollment of Teenagers.* Kalamazoo, MI: Upjohn Institute for Employment Research.

Ihlanfeldt, Keith, and David Sjoquist. 1991. "The Effect of Job Access on Black and White Youth Unemployment." *Urban Studies* 28:255–265.

Johnson, B., P. Groothuis, and J. Whitehead. 2001. "The Value of Public Goods Generated by a Major League Sports Team: The CVM Approach." *Journal of Sports Economics* 2:6–21.

Krieg, Randall. 1990. "Does Migration Function to Reduce Earnings Differentials by Race and Gender?" *Annals of Regional Science* 24(3):211–221.

Levinson, Arik. 1996. "Environmental Regulations and Manufacturers' Location

Choices: Evidence from the Census of Manufacturers." Unpublished paper, University of Wisconsin.

Marston, Stephen T. 1985. "Two Views of the Geographic Distribution of Unemployment." *Quarterly Journal of Economics* 100:57–79.

Noll, Roger, and Andrew Zimbalist. 1997. "The Economic Impact of Sports Teams and Facilities." In Noll and Zimbalist, eds., *Sports, Jobs, and Taxes: The Economic Impact of Sports Teams and Stadiums,* pp. 55–91. Washington, DC: Brookings Institution.

Oates, Wallace E., and Robert M. Schwab. 1997. "The Impact of Urban Local Taxation: The Pittsburgh Experience." *National Tax Journal* 50:1–21.

O'Flaherty, Brendan. 2000. "Newark's Non-Renaissance and Beyond." Unpublished paper, Department of Economics, Columbia University.

Papke, Leslie. 1993. "What Do We Know about Enterprise Zones?" In James M. Poterba, ed., *Tax Policy and the Economy.* Cambridge, MA: MIT Press.

———. 1994. "Tax Policy and the Economy: Evidence from the Indiana Enterprise Zone Program." *Journal of Public Economics* 54:37–49.

Peters, Alan H., and Peter S. Fisher. 2002. *State Enterprise Zones: Do They Work?* Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

Phillips, Joseph, and Ernest Goss. 1995. "The Effect of State and Local Taxes on Economic Development: A Meta-Analysis." *Southern Economic Journal* 62:320–333.

Raphael, Steven, and David A. Riker. 1999. "Geographic Mobility, Race, and Wage Differentials." *Journal of Urban Economics* 45(1):17–46.

Raphael, Steven, and Rudolf Winter-Ebmer. 2001. "Identifying the Effect of Unemployment on Crime." *Journal of Law and Economics* 44:259–283.

Rauch, James. 1993. "Does History Matter Only When It Matters Little? The Case of City-Industry Location." *Quarterly Journal of Economics* 108:843–867.

Rosentraub, Mark. 1997. "Stadiums and Urban Space." In Noll and Zimbalist, eds., *Sports, Jobs, and Taxes: The Economic Impact of Sports Teams and Stadiums,* pp. 175–207. Washington, DC: Brookings Institution.

Siegfried, John, and Andrew Zimbalist. 2000. "The Economics of Sports Facilities and Their Communities." *Journal of Economic Perspectives* 14(Summer):95–114.

Tannenwald, Robert. 1997. "State Regulatory Policy and Economic Development." *New England Economic Review,* March/April:83–99.

U.S. Bureau of the Census. 2000. *Statistical Abstract 2000.* Washington, DC: Government Printing Office.

Walden, Michael. 1997. "Don't Play Ball." *Carolina Journal,* October/November:23.

Wassmer, Robert W. 1992. "Property Tax Abatement and the Simultaneous Determination of Local Fiscal Variables in a Metropolitan Area." *Land Economics* 68:263–282.

Wasylenko, Michael. 1997. "Taxation and Economic Development: The State of the Economic Literature." *New England Economic Review,* Federal Reserve Bank of Boston, March/April:37–52.

White, James. 1988. "Property Tax Exemptions for New Development: A Test." Working Paper 88/15, Hamilton College, Clinton, New York.

Zimmerman, Dennis. 1991. *The Private Use of Tax-Exempt Bonds: Controlling Public Subsidies of Private Activities.* Washington, DC: Urban Institute Press.

Epilogue

Probably the most surprising fact about September 11, 2001, in New York City is how little the lives of most people have been affected, relative to the scale of carnage and horror that occurred at Ground Zero. Within a few weeks, almost everyone (the conspicuous exceptions being those who were in the World Trade Center towers, their immediate families, and the rescue workers) was back doing about the same things he or she usually did—going to work or school, chatting on the phone, working out, watching TV, planning how to get ahead. Life was different—some subway lines were out, commuting was tough, flags flew everywhere, and sadness and loss hung over almost every conversation and exchange. But the dramatic deaths of thousands of people and the fiery collapse of billions of dollars worth of steel, glass, and concrete did not stop most of us in New York from going about our regular business.

That's because cities are strong. Fires, floods, earthquakes, bombings usually cause very little permanent damage. Don Davis and David Weinstein have shown that Japanese cities recovered from massive bombing raids in World War II—raids that caused proportionate damage ten to sixty times greater than that of September 11—in about a decade. (Ironically, they completed their article on this topic on September 10, 2001.) The 1918 influenza pandemic killed about eight times as many New Yorkers as the terrorist attack of September 11, but it hardly interrupted the city's growth.

This book begins with an account of a minor crash—my own car accident—and ends with a major one—September 11. Both stories show that somehow cities work: Binghamton kept the car crash from having any real impact on our lives (I ran a decent 100-kilometer race thirty-six hours later), and New York gave the victims solace and kept chugging along. Cities are remarkable organizations, and crucial to our lives. In this book we have seen how they operate, and we have tried to understand how they could work better.

Reference

Davis, Donald R., and David E. Weinstein. 2002. "Bones, Bombs, and Break Points: The Geography of Economic Activity." *American Economic Review* 92(3):1269–1289.

Glossary

An asterisk denotes a word or phrase that is discussed separately in the Glossary.

advance disposal fee (ADF). A tax on material (like packaging) that is likely to enter the solid waste stream.

adverse selection. An insurance term for the problems that arise when the willingness of a potential insured to buy insurance is a signal that he or she is a relatively poor risk. Economists use this term more generally to describe situations in which A's willingness to engage in a transaction with B indicates to B that the transaction is likely to be less valuable to B than if A had not been willing. Groucho Marx's refusal to join any club that would accept him as a member is an example of adverse selection.

agglomeration economies. Production is easier if more activity is going on nearby.

antitrust. Legal actions designed to break up a monopoly,* prevent its establishment, or mitigate its ill effects.

average cost. The total cost of producing some quantity of output, divided by the number of units of output produced. If price is the same as average cost, an operation breaks even.

bottleneck. A situation in which more people want to use a particular facility at a particular time than can physically be accommodated.

bubble. When the price of an asset rises above the capitalized value* of the incomes associated with it.

capacity. A measure of the maximum amount that a facility can accommodate at one instant. For example: the number of seats in a subway car, the diameter of a water pipe, the number of spaces in a parking lot. Also, with water systems, a measure of the maximum amount that can be stored at one moment.

capital gain. The difference between the price at which someone purchases an investment like a house or a share of stock, and the price at which he or she later sells it.

capitalized value. The capitalized value of a stream of benefits extending through the future is the amount of money you could get on the market for the right to receive that stream. Sometimes referred to as present value. If ownership of a stream of benefits is legally attached to ownership of an asset, that stream of benefits is said to be "capitalized into" the value of that asset.

central business district (CBD). The area in a city where the densest office and business development is located and land rents are highest.

centralization. A group is centralized if its members on average live closer to the center of a city (the central business district) than other groups do on average.

certainty drought price. The price to import a unit of water if a reservoir runs dry, or the value of the most valuable use forgone if the reservoir runs dry, whichever is less.

closed-city case. The case in which the number of people and businesses in a particular city is fixed.

Coase theorem. Negotiation produces efficiency, provided it is costless and operates

according with well-established property rights. The physical outcome is the same no matter how property rights are assigned, so long as they are costlessly tradable.

come-as-you-please system. A system for handling a bottleneck* in which users arrive when they want to, tolls are not differentiated, and lines are allowed to form.

comparative advantage. What matters for the efficient assignment of people or nations to different activities is an actor's relative ability to perform one activity (relative to another activity), not how well she does it absolutely.

congestion pricing. A system of tolls or public transit fares in which the amount a user pays is intended to depend on the costs he imposes on other users. With dynamic congestion pricing, the toll or price changes very rapidly, every minute or every few minutes. In static congestion pricing, the price or toll changes quite slowly, maybe once or twice a day. (When I refer to congestion pricing without specifying dynamic or static, I mean dynamic.)

consumer surplus. The difference between the maximum amount that consumers in the aggregate would be willing to pay for a collection of goods—if they were forced to buy them one at a time from a mind-reading seller making take-it-or-leave-it offers—and what they actually pay.

deadweight loss. The difference between the total gains that accrue from some policy, like a sales tax, and the total losses. A synonym for excess burden.*

demand-side housing programs. Programs that try to influence directly how poor people act, rather than how housing suppliers act. See also **supply-side housing programs.**

demand smoothing. The economies of scale that arise because larger operations have less need for downtime between demands on their services.

density gradient. A graph that shows the number of residents per square meter as a function of distance from a magnet site.

deposit/refund. A system in which the purchaser of an item (a soda bottle) pays a deposit on purchase, and the deposit is returned when the purchaser or anyone else brings the bottle back.

deterrence. The idea that punishing those who do wrong makes doing wrong less attractive, and so encourages people to live better lives.

diminishing marginal returns. As more of an input is used, holding all other inputs constant, the additional output from one more unit of the input goes down.

discrimination. Differences in treatment or favor, on a basis like group membership, that are not appropriate in the case in question.

displacement cost. The loss suffered from not living in or locating a business exactly at a magnet site.

dynamic congestion pricing. See **congestion pricing.**

economic compulsion. The argument that illicit drug use causes users to commit thefts and robberies to get money to pay for their use, and that therefore drug-use regulation will reduce crime.

economies of scale. An activity has economies of scale if doubling output less than doubles cost.

efficiency. Considerations of potential Pareto improvements.*

elasticity. Demand is elastic if a small percentage change in price or cost causes a big percentage change in the quantity demanded. Demand is inelastic if a big percentage change in price or cost causes a small percentage change in the quantity demanded. Demand is perfectly inelastic if price or cost doesn't affect quantity at all.

Demand is perfectly elastic if there is a particular price or cost at which demand is arbitrarily large, and there is no demand at any price even a penny higher.

eminent domain. The power of a government to take possession of real property at the price the property is worth in the private market. An owner cannot hold out against a government.

entitlement program. A program in which everyone who meets the eligibility standards receives the service. A program like public housing is not an entitlement program because many people who meet the eligibility standards do not receive housing.

equilibrium. A situation in which everyone is doing as well as she can, given correct beliefs about how everyone else will behave.

excess burden. The extent to which a tax is inefficient. A synonym for deadweight loss.*

ex post opportunism. After one party to a contract has made a specific investment, the other party can change its demands and in general exploit the party that has made the investment.

external cost (benefit). A cost that is not borne directly by the person who decides about incurring it. For example, if I decide to play loud music at night and you live next door, you will bear the cost of lost sleep, but you will not decide about the playing of loud music.

externalities. External costs* or benefits.

factor substitution. In the decisions that a business or a household makes about how to produce something, when the price of one input goes up, less of that input and more of some other input is used to produce the same amount the cheapest way.

fixed cost. A cost you incur before you begin doing some activity that does not vary with the amount of the activity that you do. Learning to be a piano tuner is a fixed cost: it's a cost to me, and the cost is the same no matter whether I tune a million pianos or one. Similar to overhead.*

flight-from-blight hypothesis. U.S. cities are too sparsely populated because negative externalities are overconcentrated in their central parts.

fungible. Capable of being bought and sold; interchangeable.

gregariousness. The tendency of buses or subway trains to travel in groups.

horizontal equity. Equal treatment of equals.

imputed income. The value of things that you get that are not cash and so do not appear on your pay stub.

incapacitation. Prison reduces crime because prisoners cannot commit crimes while incarcerated.

inclusionary zoning. A requirement that the developer of market-rate housing set aside a certain proportion of units to sell at subsidized prices to low-income buyers.

increasing returns to scale. Doubling all inputs more than doubles the output.

index of dissimilarity. A measure of segregation between two groups. Specifically, the minimum proportion of one group that would have to move to make every neighborhood have the two groups in the same proportion.

index of isolation. A measure of segregation between one group and all others. Specifically, the average proportion of members of that group in the neighborhoods that that group lives in; the probability that a member of the group will have a co-member as his or her neighbor.

industrial revenue bonds. Tax-exempt bonds sold by state or local governments to pay for construction of private plants or office buildings. The private party benefiting from them pays them off—but at a lower rate of interest because of the tax exemption.

inefficient taxation. Taxation in which the burden is very large relative to the revenue; taxation in which relative deadweight loss* is great.

inelastic demand (supply). See **elasticity.**

judgment-proof. A person is judgment-proof if he does not have to pay a judgment against him. The most common reason for not paying is that you can't. If a court orders you to pay me $5 million, and all you have is $1.36, you are judgment-proof.

land rent. The price of using a certain piece of land for a year.

localization economies. Economies of scale within an industry; the greater the production in an industry within a city, the lower the costs.

location theory. The study of how various activities are allocated to different places and how land rents are determined.

lock-in. A situation in which, after an investment is made, some person's opportunities to deal with alternative agents are severely limited.

marginal cost. The cost of the last unit of something. If you're making fifteen loaves of bread, the marginal cost of the bread is the difference between the total cost of making fifteen loaves and the total cost of making fourteen.

metropolitan area. A core area with a large population nucleus, together with adjacent communities that have a high degree of social and economic cohesion.

monocentric city model. A model of a city in which all production is concentrated in one location and all workers commute to that location from outlying houses.

monopoly. An industry with only one seller; hence a firm with market power.

moral hazard. The possibility of loss to an insurance company because insurance might make the insured more likely to take actions that make the event insured against more likely. For instance, if I have automobile insurance, I may drive more recklessly because the insurance company, not I, will have to pay for the damage I do in an accident. Economists use the term more generally: any contract—not just an insurance contract—that creates the possibility that one party may do something that harms the other party because of the contract is called an instance of moral hazard. For example, economists would say that when professional athletes bet against their own team there is a moral hazard problem.

mortgage. A loan in which a large valuable object, such as a building, house, or boat, is pledged as collateral. If the borrower fails to pay the loan back in a timely fashion, the lender has the right to take possession of the collateral.

natural monopoly. An industry in which demand is so small relative to increasing returns to scale that if output is produced at minimal per-unit cost, then only one firm can sell its output.

open-city case. The case in which the number of people and businesses in a city is free to vary. Migration is costless, and so people and businesses with the same characteristics can be no better off in one city than in any other.

overhead. Costs of operating a firm that must be incurred every day, but that do not vary with the amount of output the firm produces.

Pareto improvement. A change that makes some people better off and no one worse off.

Pigouvian subsidies. Subsidies designed to correct externalities by being equal to the difference between private and social cost.

Pigouvian taxes. Taxes designed to correct externalities by being equal to the difference between private and social cost.

police powers. A government's powers to establish, at its discretion, laws designed to promote health, safety, morals, and general welfare.

potential Pareto improvement. A change in which the benefits to those who gain are greater than the losses to those who don't gain. Specifically, the winners could compensate the losers for all their losses (so the losers would be no worse off as a result of the change and the compensation) and still be better off than they were in the original situation.

present value. See **capitalized value.**

progressive. Term describing a government policy that helps poor people more than it helps rich people, or hurts poor people less than it hurts rich people.

public good. A good whose consumption is nonrival (your consumption doesn't reduce my consumption) and nonexcludible (if I consume it you have to consume it, too). Examples are clean air, global warming, national defense, and many types of knowledge.

racial covenant. A provision in the deed to a property that forbids the owner from selling or renting the property to minorities (or in some cases, Jews).

redlining. Unjustified refusal to make loans in certain neighborhoods (generally neighborhoods where minorities live or to which they are moving).

regressive. Term describing a government policy that helps rich people more than it helps poor people, or hurts rich people less than it hurts poor people.

rehabilitation. The idea that prison time will change a prisoner's character and make him or her less likely to want to commit crimes.

rent control. A law that regulates the rent that landlords and tenants can agree on when a new lease or renewal is put in place, and also regulates the conditions for nonrenewal of leases. **Hard rent control** is a more or less complete prohibition on rent increases. **Soft rent control** allows rent increases of certain sizes under certain conditions.

rent gradient. A graph that shows the rent of a square meter of land as a function of distance from a magnet site.

rent-seeking. Expending real resources to arrange to receive a transfer. "Rent" here doesn't refer to what people pay to use a house or a piece of land; it means, roughly, consumer surplus or producer surplus.

replacement effect. The extent to which, during one criminal's incarceration, other criminals and potential criminals will make up for the crimes he or she would have committed.

reserved power doctrine. Contract for police powers* are unenforceable. "The legislature cannot bargain away the police power of a State."

retribution. The idea that those who do wrong deserve to suffer and ought to suffer, no matter what the consequences of that suffering are.

schedule delay costs. Costs from not being able to use a facility exactly when you want to.

school sending district. A connected geographic area (that means you can walk from any place in it to any other place in it without leaving it) in which all children of the same age who go to public school go to the same school.

second-best. A policy in which a price in one market is set at something other than its optimal level because a price in a related market is not at its optimal level and cannot be changed.

segregation. Separateness.

spatial mismatch hypothesis. The idea that minorities fare poorly in the labor market because they are forced to live in locations that are not easily accessible to the jobs they would otherwise hold.

specific investment. An investment made by one party to a contract that has little or no value outside the relationship.

sprawl. A pejorative term for low-density, land-intensive, automobile-oriented development.

static congestion pricing. See **congestion pricing.**

subsidence. When a land surface falls because some of the water or minerals below it have been removed.

sunk costs. Costs that have already been incurred (and are therefore, for many purposes, irrelevant).

supply-side housing programs. Programs that try to influence directly the person who owns the housing that poor people use and how he or she acts. See also **demand-side housing programs.**

tax abatement. An agreement to reduce taxes for a specific purpose and a specific amount of time, for a specific taxpayer, to spur economic development.

tax increment financing (TIF). The government builds a private improvement and pays for it by issuing bonds; the improvement increases the taxes private entities would normally pay; the increase in tax payments is used to pay off the bonds.

tort law. The branch of law that deals with wrongful actions, except breach of contract, that are subject to civil rather than criminal redress.

transfer. An action that makes some person or group better off by precisely the amount by which it makes some other person or group worse off. For example: I take a five-dollar bill out of your wallet. In trying to decide whether a change is a potential Pareto improvement, transfers can be ignored.

two-thirds rule. Where containers are the most costly input of an activity and the content of those containers is the major output, multiplying the output by x multiplies the cost by only $x^{(2/3)}$.

urban enterprise zone. A small, usually distressed area of a city in which various kinds of taxes are lowered, regulations are loosened, and often some sort of linkage is required.

urban growth boundary (UGB). A line around a metropolitan area beyond which urban development is prohibited.

urbanization economies. Production is easier in a bigger city.

utilization ratio. The ratio of the number of people using a facility (like a subway car) to its capacity.

variable cost. A cost of production that varies with the amount produced. The opposite of fixed cost.*

zoning. A division of a town or county into districts in which certain activities are prohibited and others are permitted. Zoning generally deals with large, hard-to-move, long-lasting objects.

Index