This International Student Edition is for use outside of the U.S.

EIGHTH EDITION

LABOR ECONOMICS

GEORGE J. BORJAS

Mc Graw Hill Education

Labor Economics

Eighth Edition

George J. Borjas
Harvard University

Mc Graw Hill Education

LABOR ECONOMICS

This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 LCR 21 20 19

ISBN 978-1-260-56552-2
MHID 1-260-56552-1

Cover Image: Simfo/Getty Images

mheducation.com/highered

About the Author

George J. Borjas

George J. Borjas is the Robert W. Scrivner Professor of Economics and Social Policy at the John F. Kennedy School of Government, Harvard University. He is also a research associate at the National Bureau of Economic Research and a Research Fellow at IZA. Professor Borjas received his Ph.D. in economics from Columbia University.

Professor Borjas has written extensively on labor market issues. He is the author of several books, including *Wage Policy in the Federal Bureaucracy* (American Enterprise Institute, 1980), *Friends or Strangers: The Impact of Immigrants on the U.S. Economy* (Basic Books, 1990), *Heaven's Door: Immigration Policy and the American Economy* (Princeton University Press, 1999), *Immigration Economics* (Harvard University Press, 2014), and *We Wanted Workers: Unraveling the Immigration Narrative* (Norton, 2016). He has published more than 150 articles in books and scholarly journals, including the *American Economic Review*, the *Journal of Political Economy*, and the *Quarterly Journal of Economics*.

Professor Borjas was elected a Fellow of the Econometric Society in 1998, and a Fellow of the Society of Labor Economics in 2004. In 2011, Professor Borjas was awarded the IZA Prize in Labor Economics. He was an editor of the *Review of Economics and Statistics* from 1998 to 2006. He also has served as a member of the Advisory Panel in Economics at the National Science Foundation and has testified frequently before congressional committees and government commissions.

The McGraw-Hill Series Economics

Essentials of Economics

Brue, McConnell, and Flynn
Essentials of Economics
Fourth Edition

Mandel
Economics: The Basics
Third Edition

Schiller
Essentials of Economics
Tenth Edition

Principles of Economics

Asarta and Butters
Principles of Economics

Colander
Economics, Microeconomics, and Macroeconomics
Eleventh Edition

Frank, Bernanke, Antonovics, and Heffetz
Principles of Economics, Principles of Microeconomics, Principles of Macroeconomics
Seventh Edition

Frank, Bernanke, Antonovics, and Heffetz
Streamlined Editions: Principles of Economics, Principles of Microeconomics, Principles of Macroeconomics
Third Edition

Karlan and Morduch
Economics, Microeconomics, and Macroeconomics
Second Edition

McConnell, Brue, and Flynn
Economics, Microeconomics, Macroeconomics
Twenty-First Edition

McConnell, Brue, and Flynn
Brief Editions: Microeconomics and Macroeconomics
Third Edition

Samuelson and Nordhaus
Economics, Microeconomics, and Macroeconomics
Nineteenth Edition

Schiller
The Economy Today, The Micro Economy Today, and The Macro Economy Today
Fifteenth Edition

Slavin
Economics, Microeconomics, and Macroeconomics
Twelfth Edition

Economics of Social Issues

Guell
Issues in Economics Today
Eighth Edition

Register and Grimes
Economics of Social Issues
Twenty-First Edition

Econometrics and Data Analytics

Hilmer and Hilmer
Practical Econometrics
First Edition

Prince
Predictive Analytics for Business Strategy
First Edition

Managerial Economics

Baye and Prince
Managerial Economics and Business Strategy
Ninth Edition

Brickley, Smith, and Zimmerman
Managerial Economics and Organizational Architecture
Sixth Edition

Thomas and Maurice
Managerial Economics
Twelfth Edition

Intermediate Economics

Bernheim and Whinston
Microeconomics
Second Edition

Dornbusch, Fischer, and Startz
Macroeconomics
Thirteenth Edition

Frank
Microeconomics and Behavior
Ninth Edition

Advanced Economics

Romer
Advanced Macroeconomics
Fifth Edition

Money and Banking

Cecchetti and Schoenholtz
Money, Banking, and Financial Markets
Fifth Edition

Urban Economics

O'Sullivan
Urban Economics
Ninth Edition

Labor Economics

Borjas
Labor Economics
Eighth Edition

McConnell, Brue, and Macpherson
Contemporary Labor Economics
Eleventh Edition

Public Finance

Rosen and Gayer
Public Finance
Tenth Edition

Environmental Economics

Field and Field
Environmental Economics: An Introduction
Seventh Edition

International Economics

Appleyard and Field
International Economics
Ninth Edition

Pugel
International Economics
Seventeenth Edition

To Sarah, Timothy, and Rebecca

Preface to the Eighth Edition

The original motivation for writing *Labor Economics* grew out of my years of teaching labor economics to undergraduates. After trying out many of the textbooks in the market, it seemed to me that students were not being exposed to what the essence of labor economics was about: To try to *understand* how labor markets work. As a result, I felt that students did not really grasp *why* some persons choose to work, while other persons withdraw from the labor market; *why* some firms expand their employment at the same time that other firms are laying off workers; or *why* earnings are distributed unequally.

The key difference between *Labor Economics* and competing textbooks lies in its philosophy. I believe that knowing the *story* of how labor markets work is, in the end, more important than showing off our skills at constructing elegant models of the labor market or remembering hundreds of statistics and institutional details summarizing labor market conditions at a particular point in time.

I doubt that many students will (or should!) remember the mechanics of deriving a labor supply curve or what the unemployment rate was at the peak of the Great Recession 10 or 20 years after they leave college. However, if students could remember the *story* of how the labor market works—and, in particular, that workers and firms respond to changing incentives by altering the amount of labor they supply or demand—the students would be much better prepared to make informed opinions about the many proposed government policies that can have a dramatic impact on labor market opportunities, such as a "workfare" program requiring that welfare recipients work or a payroll tax assessed on employers to fund a national health-care program or a guest worker program that grants tens of thousands of entry visas to high-skill workers. The exposition in this book, therefore, stresses the *ideas* that labor economists use to understand how the labor market works.

The book also makes extensive use of labor market statistics and reports evidence obtained from hundreds of research studies. These data summarize the stylized facts that a good theory of the labor market should be able to explain, as well as help shape our thinking about the way the labor market works. The main objective of the book, therefore, is to survey the field of labor economics with an emphasis on *both* theory and facts. The book relies much more heavily on "the economic way of thinking" than competing textbooks. I believe this approach gives a much better understanding of labor economics than an approach that minimizes or ignores the story-telling aspects of economic theory.

Requirements

The book uses economic analysis throughout. *All* of the theoretical tools are introduced and explained in the text. As a result, the only prerequisite is that the student has some familiarity with the basics of microeconomics, particularly supply and demand curves. The exposure acquired in the typical introductory economics class more than satisfies this prerequisite. All other concepts (such as indifference curves, budget lines, production functions, and isoquants) are motivated, defined, and explained as they appear in our story. The book does not make use of any mathematical skills beyond those taught in high school algebra (particularly the notion of a slope).

Labor economists also make extensive use of econometric analysis in their research. Although the discussion in this book does not require any prior exposure to econometrics, the student will get a much better "feel" for the research findings if they know a little about how labor economists manipulate data to reach their conclusions. The appendix to Chapter 1 provides a simple (and very brief) introduction to econometrics and allows the student to visualize how labor economists conclude, for instance, that winning the lottery reduces labor supply, or that schooling increases earnings. Additional econometric concepts widely used in labor economics—such as the difference-in-differences estimator or instrumental variables—are introduced in the context of policy-relevant examples throughout the text.

Changes in the Eighth Edition

The Eighth Edition offers a thorough rewriting of the entire textbook, making it the most significant revision in quite a few years. As one edition rolls into the next and material gets added to or deleted from the textbook, I think many authors discover that the book keeps moving further away from what the author originally intended. There comes a time when one needs to take a step back, get reacquainted with the entire manuscript free from the pressures of having to get the next edition out the door, take stock of how all the pieces fit together in the context of an ever-evolving field, and do a thorough rethinking of how to best present the material once more as part of a cohesive whole. I experienced that feeling about 3 years ago, shortly after the last edition was published, and decided at the time to tackle the Eighth Edition as if I were writing the textbook for the first time. And that is precisely what I have done.

Readers will find that although much will seem familiar, big chunks of the book have been completely rewritten and streamlined. The book still offers many detailed policy discussions and still uses the evidence reported in state-of-the-art research articles to illustrate the many applications of modern labor economics. The text continues to make frequent use of such econometric tools as fixed effects, the difference-in-differences estimator, and instrumental variables—tools that play a central role in the toolkit of labor economists. And the Eighth Edition even adds to the toolkit by introducing the synthetic control method.

But the text is now much leaner, making it a shorter and easier-to-read book. And it emphasizes, from the very beginning, how these empirical tools are a central part of the methodological revolution that changed labor economics in the past two decades. Empirical analysis must be much more than calculating a correlation describing the relation between two variables. It must instead reflect a well-thought-out strategy that attempts to *identify* the direct consequences of the many shocks that continually hit the labor market.

Among the specific changes in the Eighth Edition are:

1. There are several new extensions of theoretical concepts throughout the book, including a new section on household production (Chapter 2) and on the education production function (Chapter 6). Similarly, there are more detailed discussions of some empirical applications, including the signaling value of the General Equivalency Diploma (GED) and the male–female wage gap in the "gig economy."
2. The important distinction that empirical labor economics now makes between estimating correlations and identifying consequences from specific labor market shocks is introduced early in the book. Specifically, Chapter 2 has a new section discussing the

age-old distinction between correlation and causation in the context of evidence from the labor supply literature, which measures the labor supply consequences of winning a lottery or of how taxi drivers are compensated.

3. The section on the employment effects of the minimum wage provides a detailed discussion of the studies that measure the impact of the minimum wage in Seattle, with an illustration of how empirical work in labor economics, particularly when it addresses politically contentious issues, can often lead to wildly different conclusions.
4. A reorganization of the human capital material in Chapters 6 and 7. Because of the voluminous research on the economics of education, a detailed discussion of the education decision and of how to measure the returns to education now fills up Chapter 6. Chapter 7 continues the study of the human capital model by focusing on postschool investments, on the link between human capital and the wage distribution, and on the determinants of increasing wage inequality. The discussion also introduces the canonical model used in the wage structure literature that uses the Constant Elasticity of Substitution (CES) production function to derive a relative demand curve between high- and low-skill labor. The Mathematical Appendix now includes a detailed derivation of how the model is used to estimate the elasticity of substitution between two labor inputs.
5. The material on immigration, again one of those topics where the number of studies is growing rapidly, has also been reorganized and tightened. Some users of the earlier edition suggested that because of the intimate link between the wage impact of immigration and the efficiency gains from immigration, the introduction of the immigration surplus should follow immediately after the discussion of the wage impact, and I concur. The immigration material in the geographic mobility chapter now focuses on two issues that are more directly related to the migration decision: The self-selection of immigrants and the assimilation of immigrants in the receiving labor market.

Organization of the Book

The instructor will find that this book is much shorter than competing labor economics textbooks—particularly after the thorough rewriting in the Eighth Edition. The book contains an introductory chapter, plus 11 substantive chapters. If the instructor wished to cover all of the material, each chapter could serve as the basis for about a week's worth of lectures in a typical undergraduate semester course. Despite the book's brevity, the instructor will find that all of the key topics in labor economics are covered systematically. The discussion, however, is kept to essentials as I have tried very hard not to deviate into tangential material, or into 10-page-long ruminations on my pet topics.

Chapter 1 presents a brief introduction that exposes the student to the concepts of labor supply, labor demand, and equilibrium. The chapter uses the "real-world" example of the Alaskan labor market during the construction of the oil pipeline to introduce these concepts. In addition, the chapter shows how labor economists contrast the theory with the evidence, as well as discusses the limits of the insights provided by both the theory and the data. The example used to introduce the student to regression analysis is drawn from "real-world" data—and looks at the link between differences in mean wages across occupations and differences in educational attainment as well as the "female-ness" of occupations.

The book begins the detailed analysis of the labor market with a detailed study of labor supply and labor demand. Chapter 2 examines the factors that determine whether a person chooses to work and, if so, how much, while Chapter 3 examines the factors that determine how many workers a firm wants to hire. Chapter 4 puts together the supply decisions of workers with the demand decisions of employers and shows how the labor market "balances out" the conflicting interests of the two parties. These three chapters jointly form the core of the neoclassical approach to labor economics.

The remainder of the book extends and generalizes the basic supply–demand framework. Chapter 5 stresses that jobs differ in their characteristics, so that jobs with unpleasant working conditions may have to offer higher wages in order to attract workers. Chapter 6 stresses that workers are different because they differ in their educational attainment, while Chapter 7 notes that workers also differ in how much on-the-job training they acquire. These investments in human capital help determine the shape of the wage distribution. Chapter 8 describes a key mechanism that allows the labor market to balance out the interests of workers and firms, namely labor turnover and migration.

The final section of the book discusses distortions and imperfections in labor markets. Chapter 9 analyzes how labor market discrimination affects the earnings and employment opportunities of minority workers and women. Chapter 10 discusses how labor unions affect the relationship between the firm and the worker. Chapter 11 notes that employers often find it difficult to monitor the activities of their workers, so that the workers will often want to "shirk" on the job. The chapter discusses how different types of incentive pay systems arise to discourage workers from misbehaving. Finally, Chapter 12 discusses why unemployment can exist and persist in labor markets.

The text uses a number of pedagogical devices designed to deepen the student's understanding of labor economics. A chapter typically begins by presenting a number of stylized facts about the labor market, such as wage differentials between blacks and whites or between men and women. The chapter then presents the story that labor economists have developed to understand why these facts are observed in the labor market. Finally, the chapter extends and applies the theory to related labor market phenomena. Each chapter typically contains at least one lengthy application of the material to a major policy issue, as well as boxed examples showing the "Theory at Work."

The end-of-chapter material also contains a number of student-friendly devices. There is a chapter summary describing briefly the main lessons of the chapter; a "Key Concepts" section listing the major concepts introduced in the chapter (when a key concept makes its first appearance, it appears in **boldface**). Each chapter includes "Review Questions" that the student can use to review the major theoretical and empirical issues, a set of 15 problems (many of them brand new) that test the students' understanding of the material, as well as a list of "Selected Readings" to guide interested students to many of the standard references in a particular area of study.

Supplements for the Book

There are several learning and teaching aids that accompany the eighth edition of *Labor Economics*. These resources are available to instructors for quick download and convenient access via the Instructor Resource material available through McGraw-Hill Connect®.

A *Solutions Manual* and *Test Bank* have been prepared by Robert Lemke of Lake Forest College. The Solutions Manual provides detailed answers to all of the end-of-chapter problems. The comprehensive Test Bank offers over 350 multiple-choice questions in Word and electronic format. Test questions have now been categorized by AACSB learning categories, Bloom's Taxonomy, level of difficulty, and the topic to which they relate. The computerized Test Bank is available through *McGraw-Hill's EZ Test Online*, a flexible and easy-to-use electronic testing program. It accommodates a wide range of question types and you can add your own questions. Multiple versions of the test can be created and any test can be exported for use with course management systems such as Blackboard. The program is available for Windows and Macintosh environments. *PowerPoint Presentations* prepared by Michael Welker of Franciscan University of Steubenville, contain a detailed review of the important concepts presented in each chapter. The slides can be adapted and edited to fit the needs of your course. A *Digital Image Library* is also included, which houses all of the tables and figures featured in this book.

Acknowledgments

I have benefited from countless e-mail messages sent by users of the textbook—both students and instructors. These messages often contained very valuable suggestions, most of which found their way into the Eighth Edition. I strongly encourage users to contact me (gborjas@harvard.edu) with any comments or changes that they would like to see included in the next revision. I am grateful to Robert Lemke of Lake Forest College, who updated the quiz questions for this edition, helped me expand the menu of end-of-chapter problems, and collaborated in and revised the *Solutions Manual* and *Test Bank*; and Michael Welker, Franciscan University of Steubenville, who created the PowerPoint presentation for the Eighth Edition. I am particularly grateful to many friends and colleagues who have generously shared some of their research data so that I could summarize and present it in a relatively simple way throughout the textbook, including Daniel Aaronson, David Autor, William Carrington, Chad Cotti, John Friedman, Barry Hirsch, Lawrence Katz, Alan Krueger, David Lee, Bhashkar Mazumder, and Solomon Polachek. Finally, I have benefited from the countless comments—far too numerous to mention individually—made by many colleagues on the earlier editions.

Contents in Brief

Contents

Chapter 1

Introduction

Observations always involve theory.

—*Edwin Hubble*

Most of us will allocate a substantial fraction of our time to the labor market. How we do in the labor market helps determine our wealth, what we can afford to consume, with whom we associate, where we vacation, which schools our children attend, and even who finds us attractive. Not surprisingly, we are all eager to learn how the labor market works. **Labor economics** studies how labor markets work.

Our interest in labor markets, however, is sparked by more than our personal involvement. Many of the central issues in the debate over social policy revolve around the labor market experiences of particular groups of workers or various aspects of the employment relationship between workers and firms. The policy issues examined by modern labor economics include the following:

1. Do welfare programs create work disincentives?
2. What is the impact of immigration on the wage of native-born workers?
3. Do minimum wages increase the unemployment rate of less-skilled workers?
4. What is the impact of occupational safety and health regulations on employment and earnings?
5. Do government subsidies of human capital investments improve the economic well-being of disadvantaged workers?
6. Why did wage inequality in the United States rise so rapidly after 1980?
7. What is the impact of affirmative action programs on the earnings of women and minorities and on the number of women and minorities that firms hire?
8. What is the economic impact of unions, both on their members and on the rest of the economy?
9. Would merit pay for teachers improve the academic achievement of students?
10. Do generous unemployment insurance benefits lengthen the duration of spells of unemployment?

This diverse list of questions clearly illustrates why the study of labor markets is intrinsically more important and more interesting than the study of the market for butter (unless one happens to be in the butter business!). Labor economics helps us understand and address many of the social and economic problems facing modern societies.

1-1 An Economic Story of the Labor Market

This book tells the "story" of how labor markets work. Telling this story involves much more than simply recounting the history and details of labor law or presenting reams of statistics summarizing labor market conditions. Good stories have themes, characters that come alive with vivid personalities, conflicts that have to be resolved, ground rules that limit the set of permissible actions, and events that result inevitably from the interaction among characters.

The story we will tell about the labor market has all these features. Labor economists typically assign motives to the various "actors" in the labor market. Workers, for instance, are trying to find the best possible job and firms are trying to make money. Workers and firms, therefore, enter the labor market with clashing objectives—workers are trying to sell their labor at the highest price and firms are trying to buy labor at the lowest price.

The exchanges between workers and firms are constrained by the ground rules that the government imposes to regulate transactions in the labor market. Changes in these rules and regulations obviously lead to different outcomes. For instance, a minimum wage law prohibits exchanges that pay less than a particular amount per hour worked; occupational safety regulations forbid firms from offering working conditions that are deemed too risky to the worker's health.

The deals that are struck between workers and firms determine the types of jobs that are offered, the skills that workers acquire, the extent of labor turnover, the structure of unemployment, and the observed earnings distribution. The story thus provides a theory, a framework for understanding, analyzing, and predicting a wide array of labor market outcomes.

The underlying philosophy of the book is that modern economics provides a useful story of how the labor market works. The typical assumptions we make about the behavior of workers and firms, and about the ground rules under which the labor market participants make their transactions, suggest outcomes often corroborated by what we see in real-world labor markets.

The discussion is guided by the belief that learning the story of how labor markets work is as important as knowing basic facts about the labor market. The study of facts without theory is just as empty as the study of theory without facts. Without understanding how labor markets work—that is, without having a theory of why workers and firms pursue some employment relationships and avoid others—we would be hard-pressed to predict the labor market impact of changes in government policies or of changes in the demographic composition of the workforce.

A question often asked is which are more important—ideas or facts? This book stresses that "ideas *about* facts" are most important. We do not study labor economics so that we can construct elegant mathematical theories or to remember that the unemployment rate was 6.9 percent in 1993. Rather, we want to identify which economic and social factors generate a certain level of unemployment, and why.

The main objective of this book is to survey the field of labor economics with an emphasis on *both* theory and facts: Where the theory helps us understand how the facts are generated and where the facts can help shape our thinking about the way labor markets work.

1-2 The Actors in the Labor Market

Throughout the book, we will see that there are three leading actors in our story: workers, firms, and the government.[1]

As workers, we receive top casting. Without us, after all, there is no "labor" in the labor market. We decide whether to work or not, how many hours to work, how hard to work, which skills to acquire, when to quit a job, which occupations to enter, and whether to join a labor union.

Each of these decisions is driven by the desire to *optimize,* to choose the best available option from the various choices. In our story, workers will always act in ways that maximize their well-being. Adding up the decisions of millions of workers generates the economy's labor supply in terms of the number of persons seeking work, and also in terms of the quantity and quality of skills available to employers. As we will see throughout the book, persons who want to maximize their well-being tend to supply more time and more effort to those activities that have a higher payoff. The **labor supply curve,** therefore, is often upward sloping, as illustrated in Figure 1-1.

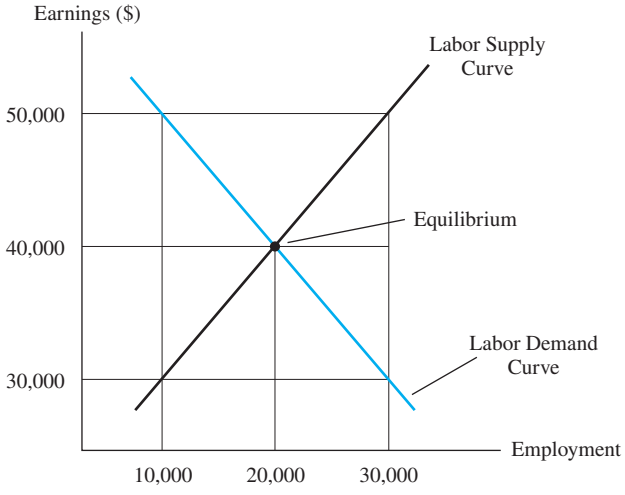
The hypothetical labor supply curve drawn in Figure 1-1 gives the number of engineers that will be forthcoming at every wage. For example, 20,000 workers are willing to supply their services to engineering firms if the engineering wage is $40,000 per year. If the engineering wage rises to $50,000, then 30,000 workers will choose to be engineers. In other words, as the engineering wage rises, more persons decide that the engineering profession is a worthwhile pursuit. More generally, the labor supply curve relates the number of person-hours supplied to the economy to the wage that is being offered. The higher the wage that is being offered, the larger the labor supplied.

Firms co-star in our story. Each firm must decide how many and which types of workers to hire and fire, the length of the workweek, how much capital to employ, and whether to offer a safe or risky working environment to its workers. Firms also have motives. We assume that firms want to maximize profits. From the firm's point of view, the consumer is king. The firm will maximize its profits by making the production decisions—and hence the hiring and firing decisions—that best serve the consumers' needs. In effect, the firm's demand for labor is a **derived demand,** a demand derived from the desires of consumers.

Adding up the hiring and firing decisions of millions of employers generates the economy's labor demand. The assumption that firms want to maximize profits implies that firms will want to hire many workers when labor is cheap but will refrain from hiring

[1] A fourth actor, trade unions, may have to be added in some countries. Unions may organize a large fraction of the workforce and represent the interests of workers in their bargaining with employers. In the United States, however, the trade union movement has been in decline for several decades. By 2016, only 6.4 percent of private-sector workers were union members.

FIGURE 1-1 Supply and Demand in the Engineering Labor Market
The labor supply curve gives the number of persons willing to supply their services to engineering firms at a given wage. The labor demand curve gives the number of engineers that firms will hire at that wage. Equilibrium occurs where supply equals demand, so that 20,000 engineers are hired at a wage of $40,000.



when labor is expensive. The relation between the price of labor and how many workers firms are willing to hire is summarized by the downward-sloping **labor demand curve** in Figure 1-1. As drawn, the labor demand curve tells us that firms in the engineering industry want to hire 20,000 engineers when the wage is $40,000 but will hire only 10,000 engineers if the wage rises to $50,000.

Workers and firms, therefore, enter the labor market with conflicting interests. Many workers are willing to supply their services when the wage is high, but few firms are willing to hire them. Conversely, few workers are willing to supply their services when the wage is low, but many firms are looking for workers. As workers search for jobs and firms search for workers, these conflicting desires are "balanced out" and the labor market reaches an **equilibrium.** In a free-market economy, equilibrium is attained when supply equals demand.

As drawn in Figure 1-1, the equilibrium wage is $40,000 and 20,000 engineers will be hired in the labor market. This wage–employment combination is an equilibrium because it balances out the conflicting desires of workers and firms. Suppose, for example, that the engineering wage was $50,000—above equilibrium. Firms would then want to hire only 10,000 engineers, even though 30,000 engineers are looking for work. The excess number of job applicants would bid down the wage as they compete for the few jobs available. Suppose, instead, that the wage was $30,000—below equilibrium. Because engineers are cheap, firms want to hire 30,000 engineers, but only 10,000 engineers are willing to work at that wage. As firms compete for the few available engineers, they bid up the wage.

There is one last major player in the labor market, the government. The government can tax the worker's earnings, subsidize the training of engineers, impose a payroll tax on

firms, demand that the racial and gender composition of engineers hired by firms exactly reflect the composition of the population, enact legislation that makes some labor market transactions illegal (such as paying engineers less than $50,000 annually), and increase the supply of engineers by encouraging their immigration from abroad. All these actions will change the equilibrium that will eventually be attained in the labor market.

The Trans-Alaska Oil Pipeline

In January 1968, oil was discovered in Prudhoe Bay in remote northern Alaska. The oil reserves were estimated to be greater than 10 billion barrels, making it the largest such discovery in North America.[2]

There was one problem with the discovery—the oil was located in a remote and frigid area of Alaska, far from where most consumers lived. To solve the daunting problem of transporting the oil to those consumers who wanted to buy it, the oil companies proposed building a 48-inch pipeline across the 789-mile stretch from northern Alaska to the southern (and ice-free) port of Valdez. At Valdez, the oil would be transferred to oil supertankers. These huge ships would then deliver the oil to consumers in the United States and elsewhere.

The oil companies joined forces and formed the Alyeska Pipeline Project. The construction project began in the spring of 1974, after Congress gave its approval in the wake of the 1973 oil embargo. Construction work continued for 3 years and the pipeline was completed in 1977. Alyeska employed about 25,000 workers during the summers of 1974 through 1977, and its subcontractors employed an additional 25,000 workers. Once the pipeline was built, Alyeska reduced its pipeline-related employment to a small maintenance crew.

Many of the workers employed by Alyeska and its subcontractors were engineers who had built pipelines across the world. Very few of those engineers were resident Alaskans. The remainder of the Alyeska workforce consisted of relatively low-skill labor such as truck drivers and excavators. Many of the low-skill workers were resident Alaskans.

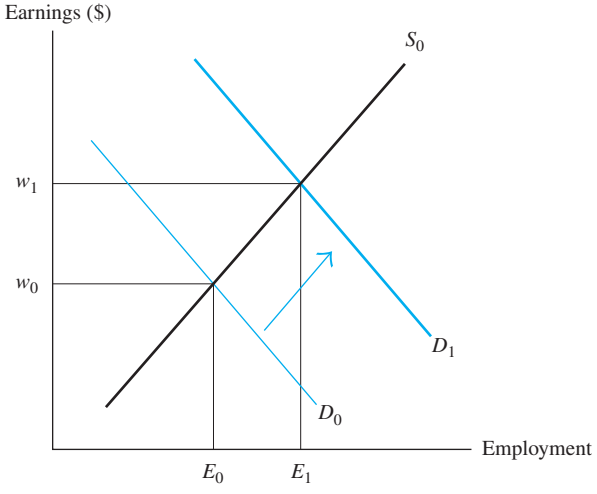
The theoretical framework summarized by the supply and demand curves can help us understand the shifts that *should* have occurred in the Alaskan labor market as a result of the Trans-Alaska Pipeline System. As Figure 1-2 shows, the labor market was initially in an equilibrium represented by the intersection of the demand curve D_0 and the supply curve S_0. A total of E_0 Alaskans were employed at a wage of w_0 in the initial equilibrium.

The construction project clearly led to a sizable increase in the demand for labor. Figure 1-2 illustrates this shift by showing the demand curve moving outward from D_0 to D_1. The outward shift in the demand curve implies that—at any given wage—Alaskan employers were looking for more workers.

The shift in demand should have moved the Alaskan labor market to a new equilibrium, represented by the intersection of the new demand curve and the original supply curve. At this new equilibrium, a total of E_1 persons were employed at a wage of w_1. The theory, therefore, predicts that the pipeline construction project should have increased *both* wages

[2] The discussion is based on William J. Carrington, "The Alaskan Labor Market during the Pipeline Era," *Journal of Political Economy* 104 (February 1996): 186–218.

FIGURE 1-2 The Alaskan Labor Market and the Construction of the Oil Pipeline
The construction of the oil pipeline shifted the labor demand curve in Alaska from D_0 to D_1, resulting in higher wages and employment. Once the pipeline was completed, the demand curve reverted back to its original level and wages and employment fell.



and employment. As soon as the project was completed, however, and the temporary need for additional workers disappeared, the demand curve would have shifted back to its original position at D_0. In the end, the wage should have gone back down to w_0 and E_0 workers would be employed. In short, the pipeline construction project should have led to a temporary increase in both wages and employment during the construction period.

Figure 1-3 shows what *actually* happened to employment and wages in Alaska between 1968 and 1983. Because the state's population was growing steadily for some decades, total employment was rising steadily even before the oil discovery in Prudhoe Bay. The data clearly show, however, that employment "spiked" in 1975, 1976, and 1977 and then went back to its long-run growth trend in 1977. The earnings of Alaskan workers also increased during the relevant period. After adjusting for inflation, monthly earnings rose from an average of $2,648 in the third quarter of 1973 to $4,140 in the third quarter of 1976, a surge of 56 percent. By 1979, real earnings were back to the level observed prior to the beginning of the pipeline construction project.

It turns out that the temporary increase in labor supply occurred for two distinct reasons. First, a larger fraction of Alaskans were willing to work when the wage increased. In the summer of 1973, about 39 percent of Alaskans worked. In the summers of 1975 and 1976, about 50 percent of Alaskans worked. Second, the rate of population growth in Alaska accelerated between 1974 and 1976, as workers living in the lower 48 states moved to Alaska to take advantage of the improved economic opportunities (despite the frigid weather conditions there). The increase in the rate of population growth, however, was temporary. Population growth reverted back to its long-run trend soon after the pipeline construction project was completed.

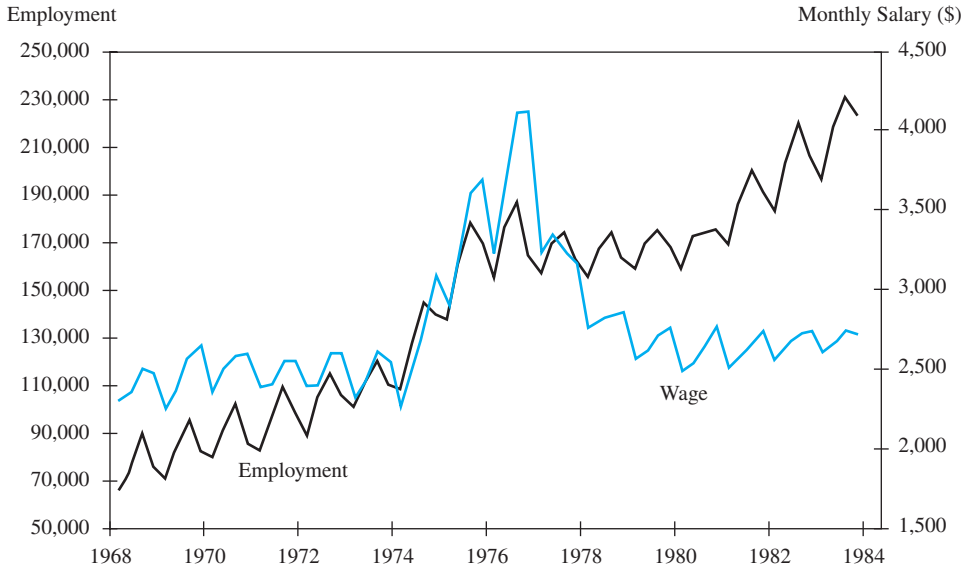


FIGURE 1-3
Wages and Employment in the Alaskan Labor Market, 1968–1984

Source: William J. Carrington, "The Alaskan Labor Market during the Pipeline Era," *Journal of Political Economy* 104 (February 1996): 199.

1-3 Why Do We Need a Theory?

We have just told a simple story of how the Trans-Alaska Pipeline System affected labor market outcomes in Alaska—and how each of the actors in our story played a major role. The government approved the pipeline project despite the potential environmental hazards involved; firms that saw income opportunities in building the pipeline increased their demand for labor; and workers responded to the change in demand by increasing the quantity of labor supplied to the Alaskan labor market.

We have, in effect, constructed a theory or **model** of the Alaskan labor market. Our model is characterized by an upward-sloping labor supply curve, a downward-sloping labor demand curve, and the assumption that an equilibrium is eventually attained that resolves the conflicts between workers and firms. This model predicts that the construction of the pipeline would temporarily increase wages and employment in the Alaskan labor market. Moreover, this prediction is testable—that is, the predictions about wages and employment can be compared with what actually happened. It turns out that the supply–demand model passes the test; the data confirm the theoretical predictions.

Needless to say, the model of the labor market illustrated in Figure 1-2 does not do full justice to the complexities of the Alaskan labor market. It is easy to come up with many variables that our simple model ignored and that could potentially change our predictions. For instance, it is possible that workers care about more than just the wage when they make labor supply decisions. The opportunity to participate in such a challenging or cutting-edge project as the construction of the Trans-Alaska Pipeline could have attracted engineers at wages lower than those offered by firms engaged in more mundane projects—despite the harsh working conditions in the field. The theoretical prediction that the construction of the pipeline project would increase wages would then be incorrect because the project could have attracted more workers at lower wages.

If the factors that we omitted from our theory play a crucial role in understanding how the Alaskan labor market operates, we might be wrongly predicting that wages and employment would rise. If these factors are only minor details, however, our model captures the essence of what goes on in the Alaskan labor market and our prediction would be valid.

We could try to build a more complex model, a model that incorporates every single one of the omitted factors. Now *that* would be a tough job! A completely realistic model would have to describe how millions of workers and firms interact and how these interactions work themselves through the labor market. Even if we knew how to accomplish such a difficult task, this "everything-but-the-kitchen-sink" approach defeats the whole purpose of having a theory. A theory that mirrored the real-world labor market in Alaska down to the minutest detail might indeed be able to explain all the facts, but it would be as complex as reality itself, cumbersome and incoherent, and would not really help us understand how the Alaskan labor market works.

There has been a long debate over whether a theory should be judged by the realism of its assumptions or by the extent to which it helps us understand and predict the labor market phenomena we are interested in. We obviously have a better shot at predicting correctly if we use more realistic assumptions. At the same time, a theory that mirrors the world too closely is too clumsy and does not isolate what *really* matters. The "art" of labor economics lies in choosing which details are essential to the story and which details are not. There is a tradeoff between realism and simplicity, and good economics hits the mark just right.

As we will see throughout this book, the supply–demand framework in Figure 1-1 helps to isolate the key factors that motivate the various actors in the labor market. The model provides a useful way of organizing our thoughts about how the labor market works. It also gives a solid foundation for building more complex and more realistic models. And, most important, the model works. Its predictions are often consistent with what is observed in the real world.

The supply–demand framework predicts that the construction of the Alaska oil pipeline would temporarily increase employment and wages in the Alaskan labor market. This prediction is an example of **positive economics.** Positive economics addresses the relatively narrow "What is?" questions, such as, What is the impact of the discovery of oil in Prudhoe Bay, and the subsequent construction of the oil pipeline, on the Alaskan labor market?

Positive economics, therefore, addresses questions that can, in principle, be answered with the tools of economics, without interjecting any value judgment as to whether the particular outcome is desirable or harmful. This book is devoted to the analysis of such positive questions as: What is the impact of the minimum wage on unemployment? What is the impact of immigration on the earnings of native-born workers? What is the impact of a tuition assistance program on college enrollment rates? What is the impact of unemployment insurance on the duration of a spell of unemployment?

These positive questions, however, beg many important issues. In fact, some would say that these positive questions beg *the* most important issues: *Should* the oil pipeline have been built? *Should* there be a minimum wage? *Should* the government subsidize college tuition? *Should* the United States accept more immigrants? *Should* the unemployment insurance system be less generous?

These questions fall in the realm of **normative economics,** which addresses much broader "What should be?" questions. By their nature, the answers to these normative questions require value judgments. Because each of us probably has different values, our answers to these normative questions may differ *regardless* of what the theory or the facts

tell us about the economic impact of the oil pipeline, the employment effects of the minimum wage, or the impact of immigration on the well-being of native workers.

Normative questions force us to make value judgments about the type of society we wish to live in. Consider, for instance, the impact of immigration on a particular host country. As we will see, the supply–demand framework implies that an increase in the number of immigrants lowers the income of competing workers but raises the income of the firms that hire those workers by even more. On net, therefore, the receiving country gains. Moreover, because immigration is typically a voluntary supply decision, it also makes the immigrants better off.

Suppose, in fact, that the evidence for a particular host country was consistent with the model's predictions. In particular, the immigration of 10 million workers improved the well-being of the immigrants (relative to their well-being in their country of birth); reduced the income of native workers by $25 billion annually; and increased the income of employers by $40 billion. Let's now ask a normative question: *Should* the country admit 10 million more immigrants?

This normative question cannot be answered solely on the basis of the theory or the facts. Even though total income in the host country has increased by $15 billion, there also has been a redistribution of wealth. Some persons are worse off and others are better off.

To answer the question of whether the country should continue to admit immigrants, one has to decide whose economic welfare we should care most about: that of immigrants, who are made better off; that of native workers, who are made worse off; or that of employers, who are made better off. One might even bring into the discussion the well-being of the people left behind in the source countries, who are clearly affected by the emigration of their compatriots. It is clear that any resolution of this issue requires clearly stated assumptions about what constitutes the "national interest," about who matters more.

Many economists often take a fallback position when these types of problems are encountered. Because the immigration of 10 million workers increases the *total* income in the destination country by $15 billion, it is then possible to redistribute income so that every person in that country is made better off. A policy that can *potentially* improve the well-being of everyone in the economy is said to be "efficient"; it increases the size of the economic pie available to the country. The problem, however, is that this type of redistribution seldom occurs in the real world; the winners typically remain winners and the losers remain losers. Our answer to a normative question, therefore, forces us to confront the tradeoff between efficiency and distributional issues.

As a second example, we will see that the supply–demand framework predicts that unionization transfers wealth from firms to workers, but that unionization also shrinks the size of the economic pie. Suppose that the facts unambiguously support these theoretical implications, unions increase the total income of workers by, say, $40 billion, but the country as a whole is poorer by $20 billion. Let's now ask a normative question: *Should* the government pursue policies that discourage workers from forming labor unions?

Our answer to this normative question again depends on how we balance the gains to the unionized workers with the losses to the employers who must pay higher wages and to the consumers who must pay higher prices for union-produced goods.

The lesson should be clear. As long as there are winners and losers—and government policies inevitably leave winners and losers in their wake—neither the theoretical implications of economic models nor the facts are sufficient to answer the normative question of whether a particular policy is desirable.

Despite the fact that economists cannot answer what many would consider to be the "big questions," there is an important sense in which framing and answering positive questions is crucial for any policy discussion. Positive economics tells us how particular government policies affect the well-being of different segments of society. Who are the winners, and how much do they gain? Who are the losers, and how much do they lose?

In the end, any informed policy discussion requires that we be fully aware of the price that has to be paid when making particular choices. The normative conclusion that one might reach may well depends on the magnitude of the costs and benefits associated with a particular policy. For example, the redistributive impact of unions (that is, the transfer of income from firms to workers) could easily dominate the normative discussion if unions generated only a small decrease in the size of the economic pie. The distributional impact, however, might be less relevant if unions greatly reduced the size of the economic pie.

Summary

- Labor economics studies how labor markets work. Topics addressed by labor economics include the determination of the income distribution, the economic impact of unions, the allocation of a worker's time to the labor market, the hiring and firing decisions of firms, labor market discrimination, the determinants of unemployment, and the worker's decision to invest in human capital.
- Models in labor economics typically contain three actors: workers, firms, and the government. It is typically assumed that workers maximize their well-being and that firms maximize profits. Governments influence the decisions of workers and firms by imposing taxes, granting subsidies, and regulating the "rules of the game" in the labor market.
- A good theory of the labor market should have realistic assumptions, should not be clumsy or overly complex, and should provide empirical implications that can be tested with real-world data.
- The tools of economics are helpful for answering positive questions. The answer to a normative question, however, typically requires that we impose a value judgment on the desirability of particular economic outcomes.

Review Questions

1. What is labor economics? Which types of questions do labor economists analyze?
2. Who are the key actors in the labor market? What motives do economists typically assign to workers and firms?
3. Why do we need a theory to understand real-world labor market problems?
4. What is the difference between positive and normative economics? Why are positive questions easier to answer than normative questions?

Key Concepts

derived demand, *3*
equilibrium, *4*
labor demand curve, *4*
labor economics, *1*
labor supply curve, *3*
model, *7*
normative economics, *8*
positive economics, *8*

Appendix

An Introduction to Regression Analysis

Labor economics is an empirical science. It makes extensive use of **econometrics,** the application of statistical techniques to study relationships in economic data. For example, we will be addressing such questions as

1. Do higher levels of unemployment benefits lead to longer spells of unemployment?
2. Do higher levels of welfare benefits reduce work incentives?
3. Does going to school one more year increase a worker's earnings?

The answers to these three questions ultimately depend on a correlation between pairs of variables: the level of unemployment benefits and the duration of unemployment spells; the level of welfare benefits and labor supply; and educational attainment and wages. We also will want to know not only the *sign* of the correlation, but the *size* as well. In other words, by how many weeks does a $50 increase in unemployment benefits lengthen the duration of unemployment spells? By how many hours does an increase of $200 per month in welfare benefits reduce labor supply? And by how much do our earnings increase if we get a college education?

Although this book does not use the technical details of econometric analysis in the discussion, the student can better appreciate both the usefulness *and* the limits of empirical research by knowing how labor economists manipulate the available data to answer the questions we are interested in. The main statistical technique used by labor economists is **regression analysis.**

An Example

There are sizable wage differences across occupations. We are interested in determining why some occupations pay more than others. One obvious factor that determines the average wage in an occupation is the level of education of workers in that occupation.

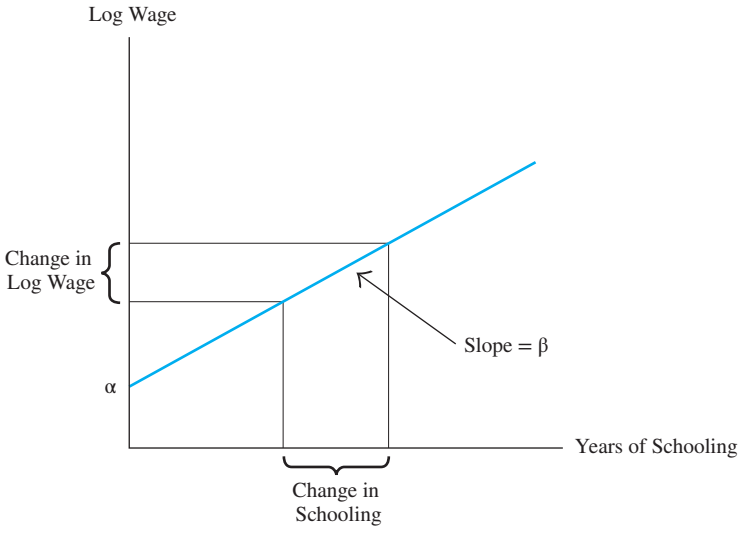
It is common in labor economics to conduct empirical studies of earnings by looking at the logarithm of earnings, rather than the actual level of earnings. There are sound theoretical and empirical reasons for this practice, one of which will be described shortly. Suppose there is a linear equation relating the average log wage in an occupation ($\log w$) to the mean years of schooling of workers in that occupation (s). We write this line as

$$\log w = \alpha + \beta s \qquad (1\text{-}1)$$

The variable on the left-hand side—the average log wage in the occupation—is called the **dependent variable.** The variable on the right-hand side—average years of schooling in the occupation—is called the **independent variable.** The main objective of regression analysis is to obtain numerical estimates of the coefficients α and β by using actual data on the mean log wage and mean schooling in each occupation. It is useful, therefore, to spend some time interpreting these **regression coefficients.**

Equation (1-1) traces out a line, with intercept α and slope β; this line is drawn in Figure 1-4. As drawn, the regression line makes the sensible assumption that the slope β is

FIGURE 1-4 The Regression Line
The regression line gives the relationship between the average log wage rate and the average years of schooling of workers across occupations. The slope of the regression line gives the change in the log wage resulting from a one-year increase in years of schooling. The intercept gives the log wage for an occupation where workers have zero years of schooling.



positive, so wages are higher in occupations where the typical worker is better educated. The intercept α gives the log wage that would be observed in an occupation where workers have zero years of schooling. Elementary algebra teaches us that the slope of a line is given by the change in the vertical axis divided by the corresponding change in the horizontal axis or

$$\beta = \frac{\text{Change in log wage}}{\text{Change in years of schooling}} \qquad (1\text{-}2)$$

Put differently, the slope β gives the change in the log wage associated with a one-year increase in schooling. It turns out that a small change in the log wage approximates the percent change in the wage. For example, if the difference in the mean log wage between two occupations is 0.051, we can say that there is approximately a 5.1 percent wage difference between the two occupations. This property is one of the reasons why labor economists typically conduct studies of salaries using the logarithm of the wage; they can then interpret changes in this quantity as a percent change in the wage. This mathematical property of logarithms implies that the coefficient β can be interpreted as giving the percent change in earnings resulting from one more year of education.

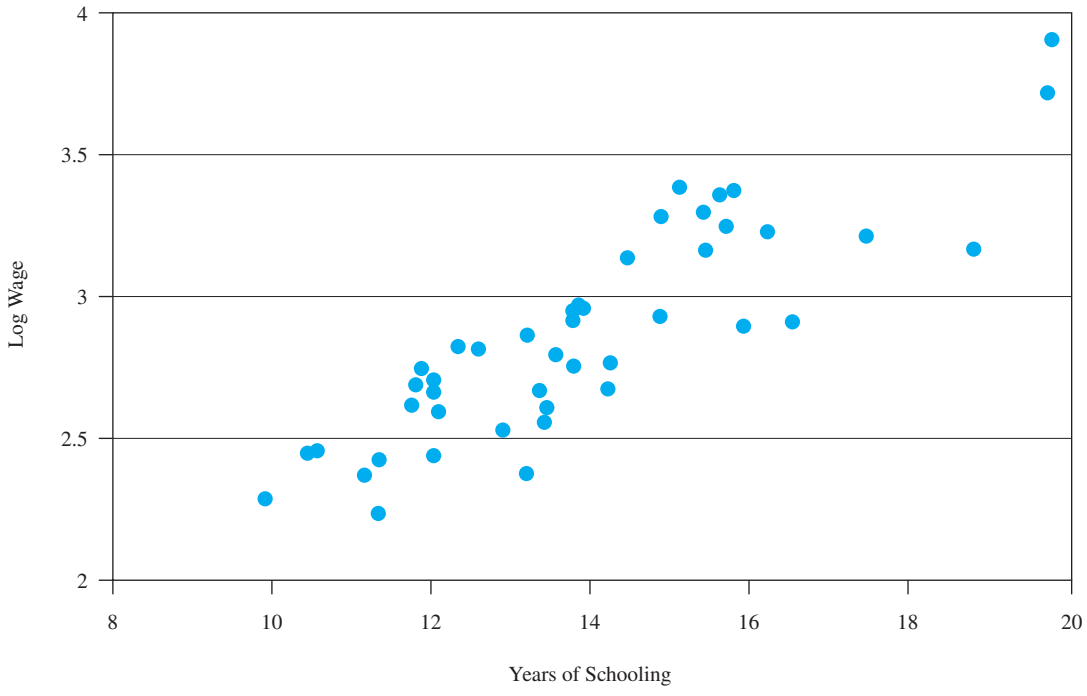
To estimate the parameters α and β, we first need to obtain data on the average log wage and average years of schooling by occupation. These data can be easily calculated using the Annual Social and Economic Supplement of the Current Population Surveys (CPS). These data, collected in March of every year by the Bureau of Labor Statistics, report employment conditions and salaries for tens of thousands of workers. One can use the data to compute the average log hourly wage and the average years of schooling for men working in each of 45 different occupations. The resulting data are reported in Table 1-1. The typical male engineer had a log wage of 3.37 and 15.8 years of schooling. In contrast, the typical construction laborer had a log wage of 2.44 and 10.5 years of schooling.

TABLE 1-1 Characteristics of Occupations, 2001

Source: Annual Demographic Files of the Current Population Survey, 2002.

Occupation	Mean Log Hourly Wage of Male Workers	Mean Years of Schooling for Male Workers	Female Share (%)
Administrators and officials, public administration	3.24	15.7	52.4
Other executives, administrators, and managers	3.29	14.9	42.0
Management-related occupations	3.16	15.4	59.4
Engineers	3.37	15.8	10.7
Mathematical and computer scientists	3.36	15.6	32.2
Natural scientists	3.22	17.4	34.2
Health diagnosing occupations	3.91	19.8	31.2
Health assessment and treating occupations	3.23	16.2	86.2
Teachers, college and university	3.17	18.8	44.7
Teachers, except college and university	2.92	16.5	75.8
Lawyers and judges	3.72	19.7	29.3
Other professional specialty occupations	2.90	15.9	54.0
Health technologists and technicians	2.76	14.2	83.1
Engineering and science technicians	2.97	13.8	26.0
Technicians, except health, engineering, and science	3.30	15.4	48.5
Supervisors and proprietors, sales occupations	2.96	13.9	37.6
Sales representatives, finance and business services	3.39	15.1	44.7
Sales representatives, commodities, except retail	3.14	14.4	25.4
Sales workers, retail and personal services	2.61	13.4	64.0
Sales-related occupations	2.93	14.8	72.4
Supervisors, administrative support	2.94	13.8	61.2
Computer equipment operators	2.91	13.8	57.1
Secretaries, stenographers, and typists	2.75	13.8	98.0
Financial records, processing occupations	2.67	14.2	92.9
Mail and message distributing	2.87	13.2	41.9
Other administrative support occupations, including clerical	2.66	13.4	79.2
Private household service occupations	2.46	10.6	96.0
Protective service occupations	2.80	13.6	18.7
Food service occupations	2.23	11.4	60.0
Health service occupations	2.38	13.2	89.1
Cleaning and building service occupations	2.37	11.2	48.2
Personal service occupations	2.55	13.4	80.4
Mechanics and repairers	2.81	12.6	5.2
Construction trades	2.74	11.9	2.4
Other precision production occupations	2.82	12.3	22.5
Machine operators and tenders, except precision	2.62	11.8	35.2
Fabricators, assemblers, inspectors, and samplers	2.65	12.0	36.2
Motor vehicle operators	2.59	12.1	12.7
Other transportation occupations and material moving	2.68	11.8	6.3
Construction laborer	2.44	10.5	3.9
Freight, stock, and material handlers	2.44	12.0	30.4
Other handlers, equipment cleaners, and laborers	2.42	11.3	28.0
Farm operators and managers	2.52	12.9	20.5
Farm workers and related occupations	2.29	9.9	18.5
Forestry and fishing occupations	2.70	12.0	3.7

FIGURE 1-5 **Scatter Diagram Relating Wages and Schooling by Occupation, 2001**

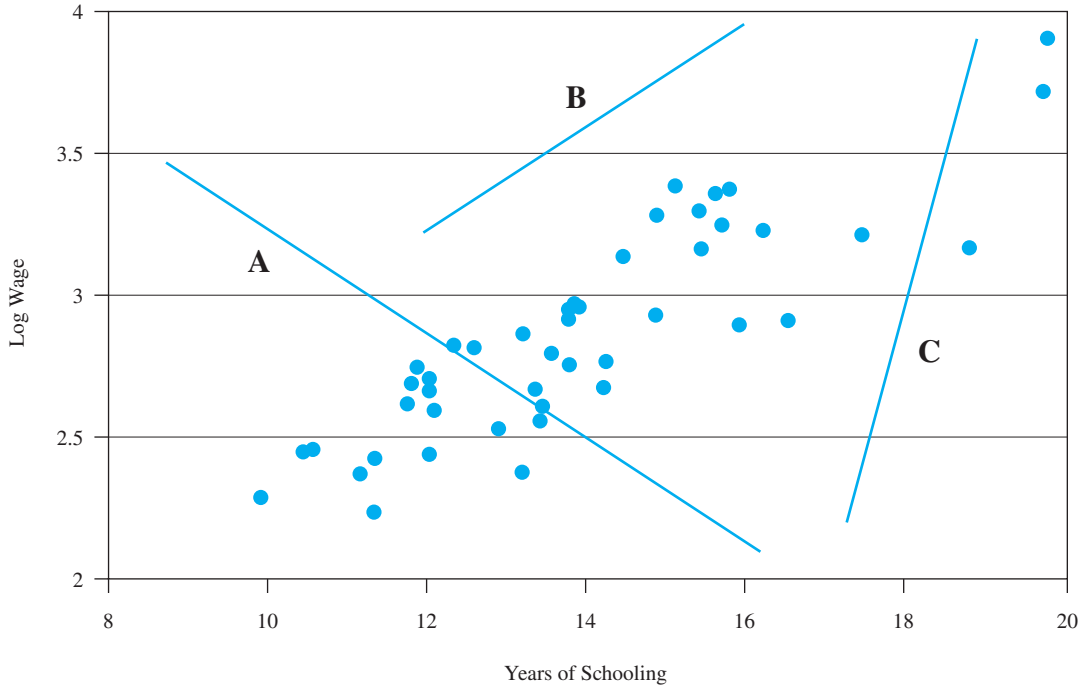


The plot of these data in Figure 1-5 is called a **scatter diagram** and describes the relation found between the average log wage and the average years of schooling in the real world. The relation between the two variables does not look anything like the regression line that we hypothesized. Instead, it is a scatter of points. But the points are not randomly scattered on the page; they have a noticeable upward-sloping drift. The raw data, therefore, suggest a positive correlation between wages and schooling, but nothing as simple as an upward-sloping line.

We have to recognize, however, that education is not the only factor that determines the average wage in an occupation. There is probably a great deal of error when workers report their salary to the Bureau of Labor Statistics. This measurement error disperses the points on a scatter diagram away from the line that we believe represents the "true" data. There also might be other factors that affect average earnings in an occupation, such as the average age of the workers or perhaps a variable indicating the "female-ness" of the occupation. It is often argued that jobs that are predominantly done by men (for example, welders) tend to pay more than jobs that are predominantly done by women (for example, kindergarten teachers). All of these factors would again disperse our data points away from the line.

The objective of regression analysis is to find the *best* line that goes through the scatter diagram. Figure 1-6 redraws our scatter diagram and inserts a few of the many lines that we could draw through the scatter. Line *A* does not represent the general trend very well; after all, the raw data suggest a positive correlation between wages and education, yet line *A*

FIGURE 1-6 Choosing among Lines Describing the Trend in the Data
There are many lines that can be drawn through the scatter diagram. Lines *A*, *B*, and *C* provide three such examples. None of these lines "fit" the trend in the scatter diagram very well.



has a negative slope. Both lines *B* and *C* are upward sloping, but they are both a bit "off"; line *B* lies above all of the points in the scatter diagram and line *C* is too far to the right.

The **regression line** is the line that best summarizes the data.[3] The formula that calculates the regression line is included in every statistics and spreadsheet software program. If we apply the formula to the data in our example, we obtain the regression line.

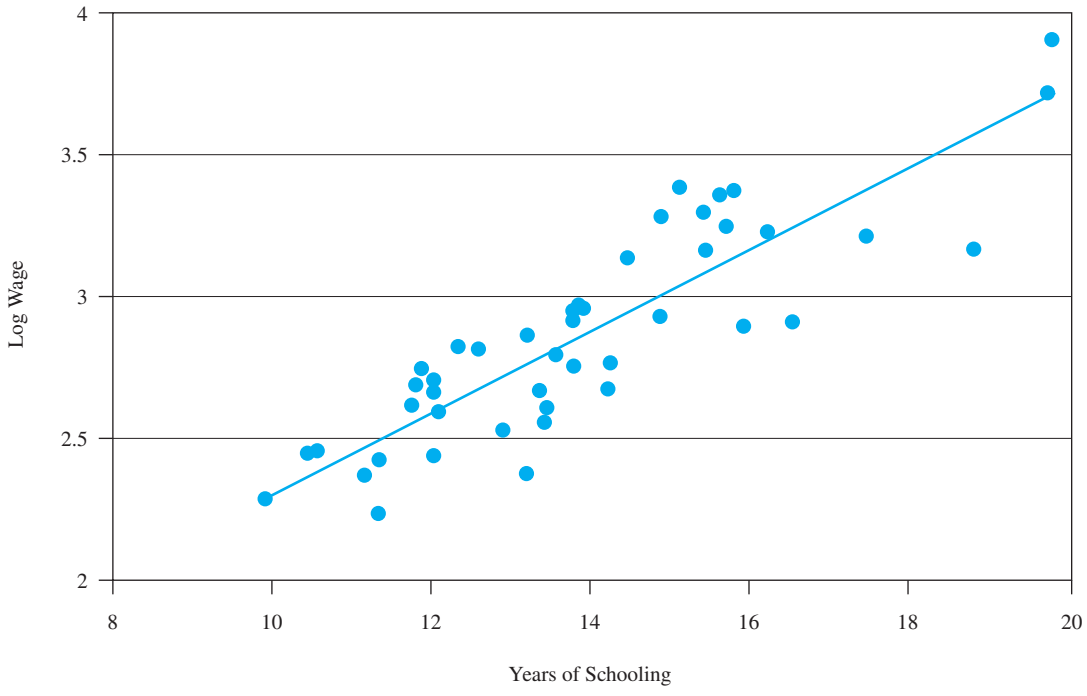
$$\log w = 0.869 + 0.143s \qquad (1\text{-}3)$$

This estimated regression line is superimposed on the scatter diagram in Figure 1-7. We interpret the regression line reported in equation (1-3) as follows. The estimated slope is positive, indicating that the average log wage is indeed higher in occupations where workers are more educated. The 0.143 slope implies that each one-year increase in the mean schooling of workers in an occupation raises the wage by approximately 14.3 percent.

The intercept indicates that the log wage would be 0.869 in an occupation where the average worker had zero years of schooling. We have to be very careful when we use this result. After all, no occupation has a workforce with zero years of schooling. In fact, the

[3] More precisely, the regression line is the line that minimizes the sum of the square of the vertical differences between every point in the scatter diagram and the corresponding point on the line. This method of estimating the regression line is called *least squares*.

FIGURE 1-7 The Scatter Diagram and the Regression Line



smallest value of *s* is 9.9 years. The intercept is obtained by extrapolating the regression line to the left until it hits the vertical axis. In other words, we are using the regression line to make an out-of-sample prediction. It is easy to get absurd results when we do this type of extrapolation: After all, what does it mean to say that the typical person in an occupation has no schooling whatsoever? An equally silly extrapolation takes the regression line and extends it to the right until, say, we wish to predict what would happen if the average worker had 25 years of schooling. Put simply, it is problematic to predict outcomes that lie outside the range of the data.

"Margin of Error" and Statistical Significance

If we plug the data reported in Table 1-1 into a statistics or spreadsheet program, we will find that the program reports many more numbers than just the intercept and the slope of a regression line. The program also reports what are called **standard errors,** or a measure of the statistical precision with which the coefficients are estimated. When poll results are reported in the media, we often hear, for instance, that 52 percent of the population believes that tomatoes should be bigger and redder, with a margin of error of ±3 percent. We use standard errors to calculate the margin of error for our estimated regression coefficients.

In our data, it turns out that the standard error for the intercept α is 0.172 and that the standard error for the slope β is 0.012. *The margin of error that is used commonly in econometric work is twice the standard error.* We can then say that a one-year increase in average

schooling increases the log wage by 0.143 ± 0.024 (or twice the standard error of 0.012). In other words, our data suggest that a one-year increase in schooling increases the average wage in an occupation by as little as 11.9 percent or by as much as 16.7 percent. Statistical theory tells us that the *true* impact of the one-year increase in schooling lies within this range with a 95 percent probability.

The regression program will also report a **t statistic** for each regression coefficient. The *t* statistic helps us assess the **statistical significance** of the estimated coefficients. The *t* statistic is defined as

$$t \text{ statistic} = \frac{\text{Absolute value of regression coefficient}}{\text{Standard error of regression coefficient}} \qquad (1\text{-}4)$$

If a regression coefficient has a *t* statistic above the "magic" number of 2, the regression coefficient is said to be significantly different from zero. In other words, it is very likely that the true value of the coefficient is not zero, so there is some correlation between the two variables that we are interested in. If a *t* statistic is below 2, the coefficient is said to be insignificantly different from zero, so we cannot conclude that there is a correlation between the two variables of interest.

Note that the *t* statistic associated with our estimated slope is 11.9 (or 0.143 ÷ 0.012), which is certainly above 2. Our estimate of the slope is significantly different from zero. It is extremely likely that there is indeed a positive correlation between the average log wage in an occupation and the average schooling of workers.

Finally, the statistical software will also report a number called the ***R*-squared**. This statistic gives the fraction of the dispersion in the dependent variable that is "explained" by the dispersion in the independent variable. The *R*-squared of the regression reported in equation (1-3) is 0.762. In other words, 76.2 percent of the variation in the mean log wage across occupations can be attributed to differences in educational attainment across the occupations. Put differently, our very simple regression model seems to do a very good job at explaining why engineers earn more than construction laborers—it is largely because one group of workers has a lot more education than the other.

Multiple Regression

Up to this point, the regression model contains only one independent variable, mean years of schooling. As noted above, the average log wage of men in an occupation will depend on many other factors. The simple correlation between wages and schooling implied by the regression model in equation (1-3) could be confounding the effect of some of these other variables. To isolate the relationship between the log wage and schooling (and avoid what is called "omitted variable bias"), it is important to control for other variables that also might generate wage differences across occupations.

Suppose we believe that occupations that are predominantly held by men tend to pay more—for given schooling—than occupations that are predominantly held by women. We can then write an expanded regression model as

$$\log w = \alpha + \beta s + \gamma p \qquad (1\text{-}5)$$

where the variable p gives the percent of workers in an occupation that are women.

We now wish to interpret the coefficients in this multiple regression model—a regression that contains more than one independent variable. Each coefficient in the **multiple regression** measures the impact of a particular variable on the log wage, *other things being equal.* For instance, the coefficient β gives the change in the log wage resulting from a one-year increase in mean schooling, holding constant the relative number of women in the occupation. Similarly, the coefficient γ gives the change in the log wage resulting from a one-percentage-point increase in the share of female workers, holding constant the average schooling of the occupation. Finally, the intercept α gives the log wage in a fictional occupation that employs only men and where the typical worker has zero years of schooling.

The last column in Table 1-1 reports the values of the female share p for the occupations in our sample. The representation of women varies significantly across occupations: 75.8 percent of teachers below the university level are women, as compared to only 5.2 percent of mechanics and repairers.

Because we now have two independent variables, our scatter diagram is three dimensional. The regression "line" is now the plane that best fits the data in this three-dimensional space. If we plug these data into a computer program to estimate the regression model in equation (1-5), the estimated regression line is given by

$$\begin{array}{l} \log w = 0.924 + 0.150s - 0.003p \qquad R\text{-squared} = 0.816 \\ \qquad\quad\ (0.154)\ \ (0.011)\ \ (0.001) \end{array} \qquad \textbf{(1-6)}$$

where the standard error of each of the coefficients is reported in parentheses below the coefficient.

A one-year increase in the occupation's mean schooling raises weekly earnings by approximately 15.0 percent. In other words, if we compare two occupations that have the same female share but differ in years of schooling by one year, workers in the better educated occupation earn 15 percent more.

We also find that the female share of the occupation has a statistically significant negative impact on the log wage. In other words, men who work in predominantly female occupations earn less than men who work in predominantly male occupations—even if both occupations have the same mean schooling. The regression coefficient, in fact, implies that a 10-percentage-point increase in the female share lowers the average earnings of an occupation by 3.0 percent.

The multiple regression model can, of course, be expanded to incorporate many more independent variables. As we will see throughout this book, labor economists put a lot of effort into defining and estimating regression models that isolate the correlation between the two variables of interest *after controlling for all other relevant factors.* Regardless of how many independent variables are included in the regression, however, all the regression models are estimated in essentially the same way: The regression line best summaries the trends in the underlying data.

Key Concepts

dependent variable, *11*
econometrics, *11*
independent variable, *11*
multiple regression, *18*
regression analysis, *11*
regression coefficients, *11*
regression line, *15*
R-squared, *17*
scatter diagram, *14*
statistical significance, *17*
standard errors, *16*
t statistic, *17*

Chapter 2

Labor Supply

It's true hard work never killed anybody, but I figure, why take the chance?

—Ronald Reagan

Each of us must decide whether to work and, once employed, how many hours to work. At any point in time, the aggregate labor supply in the economy is given by adding the work choices made by all person in the population.

The economic and social consequences of these decisions vary dramatically over time. In 1948, 84 percent of American men and 31 percent of American women aged 16 or over worked. By 2017, the proportion of working men had declined to 66 percent, whereas the proportion of working women had risen to 55 percent. Similarly, the length of the average workweek in manufacturing fell from 55 to 42 hours over the past century.[1] These labor supply trends have surely altered the nature of the American family as well as greatly affected the economy's productive capacity.

This chapter develops the framework economists use to study labor supply decisions. In this framework, individuals seek to maximize their well-being by consuming goods (such as fancy cars and nice homes) and leisure. Goods have to be purchased in the marketplace. Because most of us are not independently wealthy, we must work in order to earn the cash required to buy the desired goods. The economic trade-off is clear: If we do not work, we can consume a lot of leisure, but we have to do without the goods and services that make life more enjoyable. If we do work, we will be able to afford many of these goods and services, but we must give up some of our valuable leisure time.

The economic model of labor–leisure choice isolates the person's wage rate and income as the key variables that guide the allocation of time between the labor market and leisure activities. In this chapter, we initially use the framework to analyze "static" labor supply decisions, the factors that determine a person's labor supply at a point in time. We then extend the basic model to explore how the work decision changes as a person ages.

This economic framework not only helps us understand why women's work propensities rose and hours of work declined, but also allows us to address a number of questions with important policy implications. For example, do welfare programs reduce incentives to work? Or do cuts in the income tax rate increase hours of work?

[1] The Bureau of Labor Statistics website contains a vast collection of employment statistics; see www.bls.gov/data/home.htm.

2-1 Measuring the Labor Force

On the first Friday of every month, the Bureau of Labor Statistics (BLS) releases its estimate of the unemployment rate for the previous month. This statistic is widely regarded as a measure of the overall health of the U.S. economy. The media often interpret minor month-to-month blips in the unemployment rate as a sign of either a precipitous decline in economic activity or a surging recovery.

The unemployment rate is tabulated from the responses to a monthly BLS survey called the *Current Population Survey* (CPS). In this survey, nearly 60,000 households are questioned about their work activities during a particular week of the month (that week is called the reference week). Almost everything we know about trends in the U.S. labor force comes from tabulations of CPS data. The survey instrument used by the CPS also influenced the development of comparable surveys in other countries. In view of the importance of the CPS in the calculation of labor force statistics both in the United States and abroad, it is crucial to review the definitions of labor force activities that are routinely used by the BLS to generate its statistics.

The CPS classifies all persons aged 16 or older into one of three categories: The *employed,* the *unemployed,* and the residual group that is said to be *out of the labor force.* To be employed, a person must have been at a job with pay for at least 1 hour or worked at least 15 hours on a nonpaid job (such as the family farm). To be unemployed, a person must either be on a temporary layoff from a job or have no job but be actively looking for work in the four-week period prior to the reference week.

Let E be the number of persons employed and U the number of persons unemployed. A person participates in the **labor force** if he or she is either employed or unemployed. The size of the labor force (LF) is given by

$$LF = E + U \tag{2-1}$$

The vast majority of employed persons (those who work at a job with pay) are counted as being in the labor force regardless of how many hours they work. The size of the labor force, therefore, does not say anything about the "intensity" of work.

The **labor force participation rate** gives the fraction of the population (P) that is in the labor force and is defined by

$$\text{Labor force participation rate} = \frac{LF}{P} \tag{2-2}$$

The **employment rate** (also called the "employment–population ratio") gives the fraction of the population that is employed, or

$$\text{Employment rate} = \frac{E}{P} \tag{2-3}$$

Finally, the **unemployment rate** gives the fraction of labor force participants who are unemployed:

$$\text{Unemployment rate} = \frac{U}{LF} \tag{2-4}$$

Note a crucial detail: The number of persons who are out of the labor force does not play *any* role in the calculation of the official unemployment rate.

The Hidden Unemployed

The BLS calculates an unemployment rate based on a subjective measure of what it means to be unemployed. To be considered unemployed, a person must either be on temporary layoff or claim that he or she has "actively looked for work" in the past 4 weeks. Persons who have given up and stopped looking for work are not counted as unemployed, but are "out of the labor force." At the same time, some persons who have little intention of working may claim to be actively looking for a job in order to qualify for unemployment benefits.

The unemployment numbers, therefore, can be interpreted in different ways. During the severe recession that began in 2009, for instance, it was argued that the official unemployment rate (that is, the BLS statistic) understated the economic hardships. Because it was so hard to find work, many laid-off workers became discouraged with their futile job search activity, dropped out of the labor market, and stopped being counted *as* unemployed. A more sensible approach would perhaps add this army of **hidden unemployed** to the pool of unemployed workers, making the unemployment rate far higher than it appeared from the BLS data. For example, if the "unemployed" included persons who are out of the labor force because they are "discouraged over job prospects" as well as persons who are only "marginally attached" to the labor force, the unemployment rate in March 2011 would have been 15.7 percent, rather than the official 8.8 percent.

Some analysts believe that a more objective measure of aggregate economic activity may be given by the employment rate. The employment rate gives the fraction of the population at a job. But this statistic has the drawback that it lumps together persons who say they are unemployed with everyone who is out of the labor force. Although the latter group includes the hidden unemployed, it also includes many individuals who have little intention of working, including retirees, some women with small children, and many students enrolled in school.

A decrease in the employment rate could then be attributed to either increases in unemployment or unrelated increases in fertility or school enrollment rates. It is far from clear, therefore, that the employment rate provides a better measure of fluctuations in economic activity than the unemployment rate. We will return to some of these issues in the unemployment chapter.

2-2 Basic Facts about Labor Supply

This section summarizes some of the key trends in labor supply in the United States.[2] These facts have motivated much of the research in recent decades. Table 2-1 documents the historical trends in the labor force participation rate of men. There was a slight fall in the labor force participation rates of men in the twentieth century, from 80 percent in 1900 to 71 percent by 2010. The decline is particularly steep for men near or above age 65, as more men choose to retire earlier. The labor force participation rate of men aged 45–64, for example, declined by 12 percentage points between 1950 and 2010, while the participation rate of men over 65 declined from 46 to 22 percent. Moreover, the labor force participation

[2] More detailed discussions of labor supply trends are given by John H. Pencavel, "Labor Supply of Men: A Survey," in Orley C. Ashenfelter and Richard Layard, editors, *Handbook of Labor Economics,* vol. 1, Amsterdam: Elsevier, 1986, pp. 3–102; and Mark R. Killingsworth and James J. Heckman, "Female Labor Supply: A Survey," in ibid., pp. 103–204.

TABLE 2-1 Labor Force Participation Rates of Men, 1900–2010

Sources: U.S. Bureau of the Census, *Historical Statistics of the United States, Colonial Years to 1970,* Washington, DC: Government Printing Office, 1975; U.S. Bureau of the Census, *Statistical Abstract of the United States,* Washington, DC: Government Printing Office, various issues.

Year	All Men	Men Aged 25–44	Men Aged 45–64	Men Aged over 65
1900	80.0	94.7	90.3	63.1
1920	78.2	95.6	90.7	55.6
1930	76.2	95.8	91.0	54.0
1940	79.0	94.9	88.7	41.8
1950	86.8	97.1	92.0	45.8
1960	84.0	97.7	92.0	33.1
1970	80.6	96.8	89.3	26.8
1980	77.4	93.0	80.8	19.0
1990	76.4	93.3	79.8	16.3
2000	74.8	93.1	78.3	17.5
2010	71.2	90.6	78.4	22.1

TABLE 2-2 Labor Force Participation Rates of Women, 1900–2010

Sources: U.S. Bureau of the Census, *Historical Statistics of the United States, Colonial Years to 1970,* Washington, DC: Government Printing Office, 1975, p. 133; and U.S. Department of Commerce, *Statistical Abstract of the United States, 2011,* Washington, DC: Government Printing Office, 2011, Table 596.

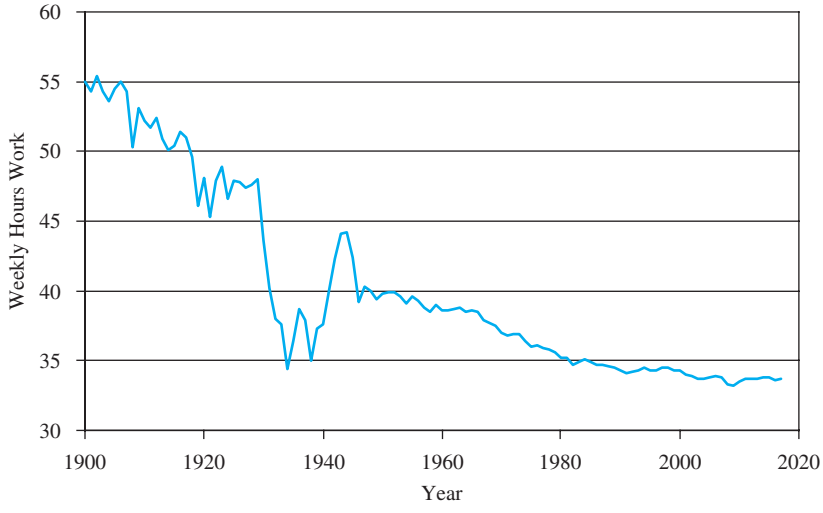
Year	All Women	Single Women	Married Women	Widowed, Divorced, or Separated
1900	20.6	43.5	5.6	32.5
1910	25.4	51.1	10.7	34.1
1930	24.8	50.5	11.7	34.4
1940	25.8	45.5	15.6	30.2
1950	29.0	46.3	23.0	32.7
1960	34.5	42.9	31.7	36.1
1970	41.6	50.9	40.2	36.8
1980	51.5	64.4	49.9	43.6
1990	57.5	66.7	58.4	47.2
2000	59.9	68.9	61.1	49.0
2010	58.6	63.3	61.0	48.8

rate of men in their prime working years (ages 25–44) also declined, from 97 percent in 1950 to 91 percent in 2010. Note, however, that the labor force participation rate of men in their retirement years has stabilized and even begun to increase in the past two decades.

As Table 2-2 shows, there also has been a huge increase in the labor force participation rate of women. At the beginning of the twentieth century, only 21 percent of women were in the labor force. As late as 1950, even after the social and economic disruptions caused by two world wars and the Great Depression, only 29 percent of women were in the labor force. During the past 50 years, however, the labor force participation rate of women grew dramatically. By 2010, almost 60 percent of all women were in the labor force. It is worth noting that the increase in female labor force participation was particularly steep among married women. Their labor force participation rate almost doubled in recent decades, from 32 percent in 1960 to 61 percent in 2010.

FIGURE 2-1 Average Weekly Hours of Work, 1900–2013

Sources: The pre-1947 data refer to workers in manufacturing and are drawn from Ethel Jones, "New Estimates of Hours of Work per Week and Hourly Earnings, 1900–1957," *Review of Economics and Statistics* 45 (November 1963): 374–385. The post-1947 data are drawn from U.S. Department of Labor, Bureau of Labor Statistics, *Employment, Hours, and Earnings from the Current Employment Statistics Survey,* "Table B-7. Average Weekly Hours of Production or Nonsupervisory Workers on Private Nonfarm Payrolls by Industry Sector and Selected Industry Detail."



These dramatic shifts in labor force participation rates were accompanied by a sizable decline in average hours of work per week. Figure 2-1 shows that the typical person employed in production worked 55 hours per week in 1900, 40 hours in 1940, and just under 34 hours in 2010.

There exist sizable differences in the various dimensions of labor supply across demographic groups at a particular point in time. As Table 2-3 shows, men not only have larger participation rates than women, but are also less likely to be employed in part-time jobs. Only 4 percent of working men are in part-time jobs, as compared to 13 percent of working women. The table also documents a strong positive correlation between labor supply and educational attainment for both men and women. In 2017, 90 percent of male college graduates and 81 percent of female college graduates were in the labor force, as compared to only 72 and 46 percent of male and female high school dropouts, respectively. There are also racial differences in labor supply, between whites and minorities as well as within the minority population itself, with blacks tending to have the lowest participation rates and Asian men the highest.

2-3 The Worker's Preferences

The framework that economists typically use to analyze labor supply behavior is called the **neoclassical model of labor–leisure choice.** This model isolates the factors that determine whether a particular person works and, if so, how many hours she chooses to work. The model tells a simple "story" that helps us understand many of the stylized facts discussed above. More importantly, it lets us predict how changes in economic conditions or in government policies will affect work incentives.

TABLE 2-3 Labor Supply in the United States, 2017 (Persons Aged 25–64)

Source: U.S. Bureau of Labor Statistics, *Current Population Survey,* Annual Social and Economic Supplement, March 2017. The average number of hours worked is calculated in the subsample of workers. The percent of workers in part-time jobs refers to the proportion working fewer than 30 hours per week.

	Labor Force Participation Rate		Annual Hours of Work		Percent of Workers in Part-Time Jobs	
	Men	Women	Men	Women	Men	Women
All persons	83.1	71.4	2,170	1,933	4.3	12.9
Educational attainment:						
Less than 12 years	72.1	45.6	2,033	1,753	5.4	19.7
12 years	79.1	63.3	2,124	1,875	4.7	14.1
13–15 years	82.5	73.5	2,166	1,906	4.8	13.4
16 years or more	90.4	80.5	2,235	2,000	3.4	11.2
Age:						
25–34	87.1	75.6	2,101	1,904	5.7	12.0
35–44	89.2	75.1	2,201	1,928	2.7	13.0
45–54	85.3	74.7	2,221	1,978	2.9	12.0
55–64	70.5	60.0	2,160	1,922	6.2	15.2
Race:						
White	83.8	73.1	2,208	1,933	4.1	13.8
Black	74.9	72.0	2,096	1,963	6.0	9.6
Hispanic	85.6	64.7	2,086	1,882	4.0	12.7
Asian	87.5	68.2	2,121	1,961	3.1	11.3

The representative person in our model receives satisfaction both from the consumption of goods (which we denote by C) and from the consumption of leisure (L) in a particular time period. Obviously, the person consumes many different types of goods. To simplify, we aggregate the dollar value of all the goods that the person consumes and define C as the total dollar value of all goods purchased. For example, if the person spends $1,000 weekly on food, rent, car payments, movie tickets, and other items, the variable C would take on the value of $1,000. The variable L gives the number of hours of leisure that a person consumes during that same period.

Utility and Indifference Curves

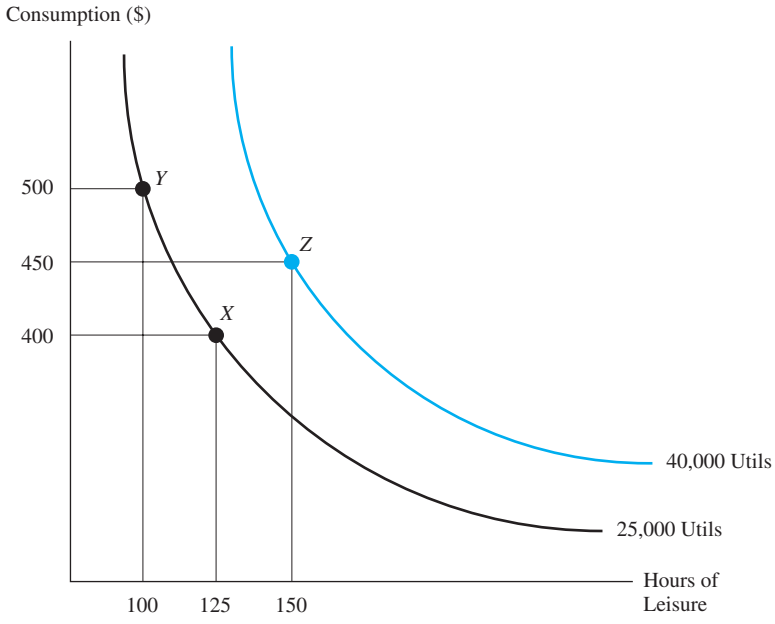
The notion that individuals get satisfaction from consuming goods and leisure is summarized by the **utility function:**

$$U = f(C, L) \qquad (2\text{-}5)$$

The utility function transforms the person's consumption of goods and leisure into an index U that measures the individual's level of satisfaction or happiness. This index is called *utility.* The higher the index U is, the happier the person will be. We make the sensible assumption that buying more goods or having more leisure hours both increase a person's utility. In the jargon of economics, C and L are "goods," not "bads."

Suppose that a person is consuming $500 worth of consumption goods and 100 hours of leisure weekly (point Y in Figure 2-2). This particular consumption basket yields a particular level of utility to the person, say 25,000 utils. It is easy to imagine that different combinations

FIGURE 2-2 Indifference Curves
Points *X* and *Y* lie on the same indifference curve and yield the same utility (25,000 utils); point *Z* lies on a higher indifference curve and yields more utility.



of goods and leisure might yield the same level of utility. For example, the person might say that she would be indifferent to consuming $500 worth of goods and 100 hours of leisure or consuming $400 worth of goods and 125 hours of leisure. Figure 2-2 illustrates the many combinations of *C* and *L* that generate this particular level of utility. The locus of such points is called an **indifference curve**—and all points along this curve yield 25,000 utils.

Suppose that the person was instead consuming $450 worth of goods and 150 hours of leisure (point *Z* in the figure). This consumption basket would certainly make the person happier, placing her on the higher indifference curve with 40,000 utils. We can then construct an indifference curve for that level of utility. In fact, we can construct an indifference curve for every level of utility. As a result, the utility function can be represented graphically in terms of a family (or a "map") of indifference curves.

Indifference curves have four important properties:

1. *Indifference curves are downward sloping.* We assumed that individuals prefer more of both *C* and *L*. If indifference curves were upward sloping, a consumption basket with more *C* and more *L* would yield the same level of utility as a consumption basket with less *C* and less *L*. This clearly contradicts our assumption that the individual likes both goods and leisure. The only way that we can offer a person a few more hours of leisure, and still hold utility constant, is to take away some of the goods.
2. *Higher indifference curves indicate higher levels of utility.* The consumption bundles lying on the indifference curve that yields 40,000 utils are preferred to the bundles lying on the curve that yields 25,000 utils. To see this, note that point *Z* in the figure

must yield more utility than point *X,* simply because the bundle at point *Z* allows the person to consume more goods and more leisure.

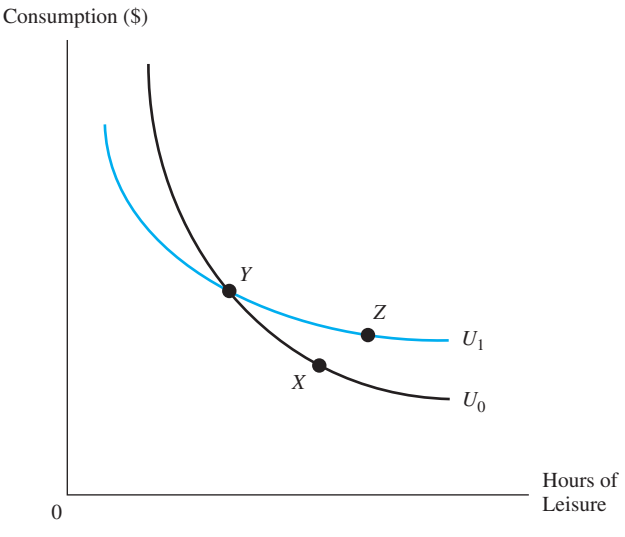
3. *Indifference curves do not intersect.* To see why, consider Figure 2-3, where indifference curves are allowed to intersect. Because points *X* and *Y* lie on the same indifference curve, the individual would be indifferent between the bundles *X* and *Y.* Because points *Y* and *Z* lie on the same indifference curve, the individual would be indifferent between bundles *Y* and *Z.* The person would then be indifferent between *X* and *Y,* and between *Y* and *Z,* so that she should also be indifferent between *X* and *Z.* But *Z* is clearly preferable to *X,* because *Z* has more goods and more leisure. Indifference curves that intersect contradict our assumption that individuals like to consume both goods and leisure.
4. *Indifference curves are convex to the origin.* The convexity of indifference curves does not follow from either the definition of indifference curves or the assumption that both goods and leisure are "goods." The convexity reflects an additional assumption about the shape of the utility function. It turns out (see Problem 2-1 at the end of the chapter) that indifference curves must be convex to the origin if we are ever to observe a person both working and consuming some leisure in the same period.

The Slope of an Indifference Curve

What happens to a person's utility as she allocates one more hour to leisure or buys an additional dollar's worth of goods? The **marginal utility** of leisure is defined as the change in utility resulting from an additional hour devoted to leisure activities, holding constant the amount of goods consumed. We denote the marginal utility of leisure as MU_L. Similarly,

FIGURE 2-3 Indifference Curves Do Not Intersect

Points *X* and *Y* yield the same utility because they are on the same indifference curve; points *Y* and *Z* should also yield the same utility. Point *Z,* however, is preferable to point *X.*



the marginal utility of consumption gives the change in utility if the individual consumes one more dollar of goods, holding constant the number of hours of leisure. We denote the marginal utility of consumption by MU_C. Because both leisure and the consumption of goods are desirable activities, the marginal utilities of leisure and consumption must be positive numbers.

As we move along an indifference curve, say from point *X* to point *Y* in Figure 2-2, the slope of the indifference curve measures the rate at which a person is willing to give up some leisure time in return for additional consumption, *while holding utility constant.* Put differently, the slope tells us how many additional dollars' worth of goods it would take to "bribe" the person into giving up some leisure time. It can be shown that the slope of an indifference curve equals[3]

$$\frac{\Delta C}{\Delta L} = -\frac{MU_L}{MU_C} \qquad \textbf{(2-6)}$$

The absolute value of the slope of an indifference curve, which is called the **marginal rate of substitution (MRS) in consumption,** is the ratio of marginal utilities.

The assumption that indifference curves are convex to the origin is essentially an assumption about how the marginal rate of substitution changes as the person moves along an indifference curve. Convexity implies that the slope of an indifference curve is steeper when the worker is consuming a lot of goods and little leisure, and flatter when the worker is consuming few goods and a lot of leisure. As a result, the absolute value of the slope of an indifference curve declines as the person "rolls down" the curve. The assumption of convexity, therefore, is equivalent to an assumption of *diminishing* marginal rate of substitution.

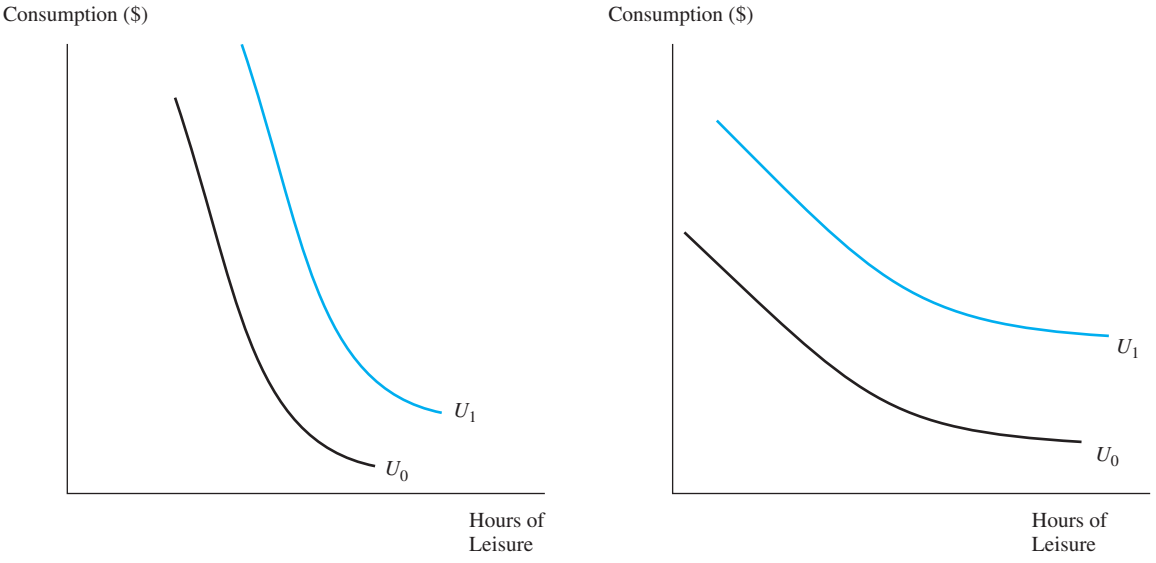
Differences in Preferences across Workers

The map of indifference curves presented in Figure 2-2 illustrates the way a *particular* worker views the trade-off between leisure and consumption. Different workers will view this trade-off differently. Some of us may like to devote a lot of time to our jobs, while others would prefer to devote most of their time to leisure. These differences in preferences imply that the indifference curves may look quite different for different workers.

Figure 2-4 shows the indifference curves for two workers, Cindy and Mindy. Cindy's indifference curves tend to be very steep, indicating that her marginal rate of substitution takes on a very high value (see Figure 2-4*a*). In other words, she requires a sizable monetary bribe (in terms of additional consumption) to convince her to give up an additional hour of leisure. Cindy obviously likes leisure a lot. Mindy, on the other hand, has flatter indifference curves, indicating that her marginal rate of substitution takes on a low value (see Figure 2-4*b*). Mindy, therefore, does not require a large bribe to convince her to give up an additional hour of leisure.

[3] To show that the slope of an indifference curve equals the ratio of marginal utilities, suppose that points *X* and *Y* in Figure 2-2 are very close to each other. When going from *X* to *Y,* the person is giving up ΔL hours of leisure, and each hour of leisure given up has a marginal utility of MU_L. Therefore, the loss in utility associated with moving from *X* to *Y* is given by $\Delta L \times MU_L$. The move from *X* to *Y* also involves a gain in utility. After all, the worker is not just giving up leisure time; she is consuming an additional ΔC dollars of goods. Each additional dollar of consumption increases utility by MU_C units. The total gain in utility is given by $\Delta C \times MU_C$. All points along an indifference curve yield the same utility. This implies that the utility loss in moving from *X* to *Y* must be exactly offset by the gain, or $(\Delta L \times MU_L) + (\Delta C \times MU_C) = 0$. Equation (2-6) follows by rearranging terms.

FIGURE 2-4 Differences in Preferences across Workers
(*a*) Cindy's indifference curves are steep; she requires a substantial bribe to give up an hour of leisure. (*b*) Mindy's indifference curves are flatter; she attaches a much lower value to her leisure time.



(*a*) Cindy's Indifference Curves

(*b*) Mindy's Indifference Curves

Interpersonal differences in the "tastes for work" are obviously important determinants of differences in labor supply in the population. For the most part, economic models gloss over these differences in preferences. The reason for this omission is that differences in tastes, although probably very important, are hard to observe and measure. It would be extremely difficult, if not impossible, to conduct surveys that attempt to measure differences in indifference curves across workers. Moreover, the reliance on taste differences provides an easy way out for anyone who wishes to explain why different workers behave differently. One can always assert that the different behavior patterns of two workers arise because worker *A* likes leisure more than worker *B,* and there would be no way of proving whether such a claim is correct.

Economic models instead emphasize the impact of variables that are easily observable—such as wages and incomes—on the labor supply decision. Because these variables can be observed, the predictions made by the model about which types of persons will tend to work more are testable and refutable.

2-4 The Budget Constraint

The person's consumption of goods and leisure is constrained by her income and by the fact there are only 24 hours in a day. Part of a person's income (such as property income, dividends, and lottery prizes) is independent of how many hours she works. We denote this

"nonlabor income" by *V.* Let *h* be the number of hours the person will allocate to the labor market during the period and *w* be the hourly wage rate. The person's **budget constraint** can be written as

$$C = wh + V \qquad \textbf{(2-7)}$$

In words, the dollar value of expenditures on goods (C) must equal the sum of labor earnings (wh) and nonlabor income (V).[4]

The wage rate plays a central role in the labor supply decision. Initially, we assume that the wage rate is constant *for a particular person,* so the person receives the same hourly wage regardless of how many hours she works. In fact, the "marginal" wage rate (that is, the wage rate received for the last hour worked) generally depends on how many hours a person works. Persons who work over 40 hours per week typically receive an overtime premium, and the wage rate in part-time jobs is often lower than that in full-time jobs.[5] For now, we ignore the possibility that a worker's marginal wage may depend on how many hours she chooses to work.

It is then easy to graph the budget constraint. The person has two alternative uses for her time: Work or leisure. The total time allocated to each of these activities must equal the total time available in the period, say T hours per week, so that $T = h + L$. We can rewrite the budget constraint as

$$C = w\,(T - L) + V \qquad \textbf{(2-8)}$$

or

$$C = (wT + V) - wL$$

This last equation is in the form of a line, and the slope is the negative of the wage rate (or $-w$).[6] The **budget line** is illustrated in Figure 2-5. Point E in the graph indicates that if the person decides not to work at all and devotes T hours to leisure, she can still purchase V dollars' worth of consumption goods. Point E is the *endowment point.* If the person is willing to give up 1 hour of leisure, she can then move up the budget line and purchase an additional w dollars' worth of goods. In fact, each additional hour of leisure that the person is willing to give up allows her to buy an additional w dollars' worth of goods. In other words, each hour of leisure consumed has a price, and the price is given by the wage rate. If the worker gives up all her leisure activities, she ends up at the intercept of the budget line and can buy ($wT + V$) dollars' worth of goods.

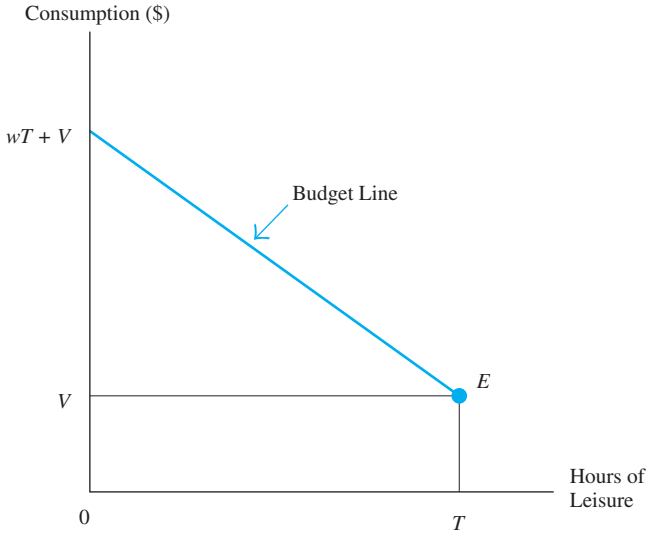
The consumption and leisure bundles that lie below the budget line are available to the worker; the bundles that lie above the budget line are not. The budget line, therefore, gives the frontier of the worker's **opportunity set**—the set of all the consumption baskets that a particular worker could afford to buy.

[4] The budget constraint implies that the worker spends all her income in the period, so there are no savings.

[5] Shelly Lundberg, "Tied Wage-Hours Offers and the Endogeneity of Wages," *Review of Economics and Statistics* 67 (August 1985): 405–410.

[6] The equation of a line relating the variables y and x is $y = a + bx$, where a is the intercept and b is the slope.

FIGURE 2-5 The Budget Line Is the Boundary of the Worker's Opportunity Set
Point *E* is the endowment point, telling the person how much she can consume if she does not work at all. The worker moves up the budget line as she trades an hour of leisure for consumption of goods. The absolute value of the slope of the budget line is the wage rate.



2-5 The Hours of Work Decision

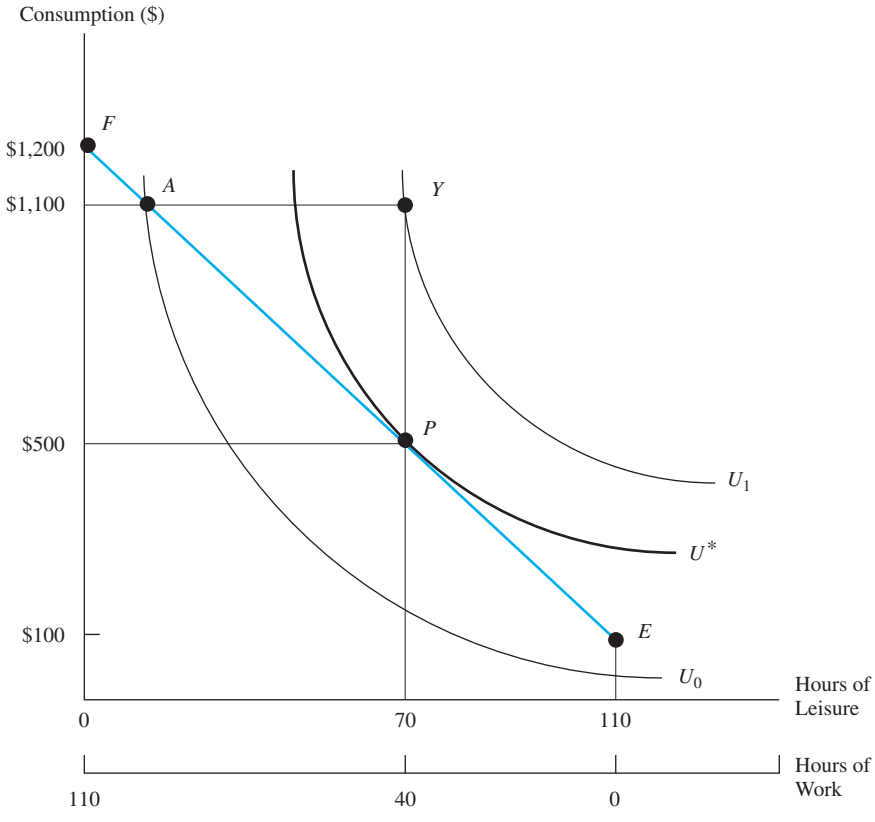
We make one important assumption about the person's behavior: She chooses the particular combination of goods and leisure that maximizes her utility. This means that the person will choose the level of goods and leisure that lead to the highest possible level of the utility index *U*—given the limitations imposed by the budget constraint.

Figure 2-6 illustrates the solution to this problem. As drawn, the budget line *FE* describes the opportunities available to a worker who has $100 of nonlabor income per week, faces a market wage rate of $10 per hour, and has 110 hours of nonsleeping time to allocate between work and leisure activities (assuming she sleeps roughly 8 hours per day).

Point *P* gives the optimal bundle of goods and hours of leisure chosen by the utility-maximizing worker. The highest indifference curve attainable places her at point *P* and gives her U^* units of utility. The worker then consumes 70 hours of leisure per week, works a 40-hour workweek, and buys $500 worth of goods weekly. The worker would obviously prefer to choose a point on indifference curve U_1, which provides a higher level of utility. For example, the worker would prefer to be at *Y,* where she works a 40-hour workweek and buys $1,100 worth of goods. Given her wage and nonlabor income, however, the worker could never afford this outcome. In contrast, the worker could choose a point such as *A,* which lies on the budget line, but she would not do so. After all, point *A* gives her less utility than point *P.*

The optimal consumption of goods and leisure, therefore, is given by the point where the budget line is tangent to the indifference curve. This type of solution is called an *interior solution* because the worker is not at either corner of the opportunity set (that is, at point *F,* working all available hours, or at point *E,* working no hours whatsoever).

FIGURE 2-6 Interior Solution to the Labor–Leisure Decision
A utility-maximizing worker chooses the consumption–leisure bundle at point *P,* where the indifference curve is tangent to the budget line.



Interpreting the Tangency Condition

At the optimal point *P,* the slope of the indifference curve equals the slope of the budget line. This implies that[7]

$$\frac{MU_L}{MU_C} = w \qquad \textbf{(2-9)}$$

At the chosen level of consumption and leisure, the marginal rate of substitution (the rate at which a person is willing to give up leisure hours in exchange for additional consumption) equals the wage rate (the rate at which the market allows the worker to substitute one hour of leisure time for consumption).

[7] Although the slope of the indifference curve and the slope of the budget line are both negative numbers, the minus signs cancel out when the two numbers are set equal to each other, resulting in equation (2-9).

The economic intuition behind this condition is easier to grasp if we rewrite it as

$$\frac{MU_L}{w} = MU_C \qquad \textbf{(2-10)}$$

The quantity MU_L gives the additional utility received from consuming an extra hour of leisure. This extra hour costs w dollars. The left-hand side of equation (2-10), therefore, gives the number of utils received from spending an additional dollar on leisure. Because C is defined as the dollar value of expenditures on consumption goods, MU_C gives the number of utils received from spending an additional dollar on goods. The tangency solution at point P implies that the last dollar spent on leisure buys the same number of utils as the last dollar spent on goods. If this equality did not hold (so that, for example, the last dollar spent on consumption buys more utils than the last spent on leisure), the worker would not be maximizing utility. She could rearrange her consumption plan so as to purchase more of the commodity that yields more utility for the last dollar.

What Happens to Hours of Work When Nonlabor Income Changes?

We want to know what happens to hours of work when the worker's nonlabor income V increases. The increase in V might be triggered by the payment of higher dividends on the worker's stock portfolio or because some distant relatives named the worker as the beneficiary in their will.

Figure 2-7 illustrates what happens to hours of work when the worker has an increase in V, *holding the wage constant*. Initially, the worker's nonlabor income equals \$100 weekly, which is associated with endowment point E_0. Given the worker's wage rate, the budget line is then given by F_0E_0. The worker maximizes utility by choosing the bundle at point P_0. At this point, the worker consumes 70 hours of leisure and works 40 h.

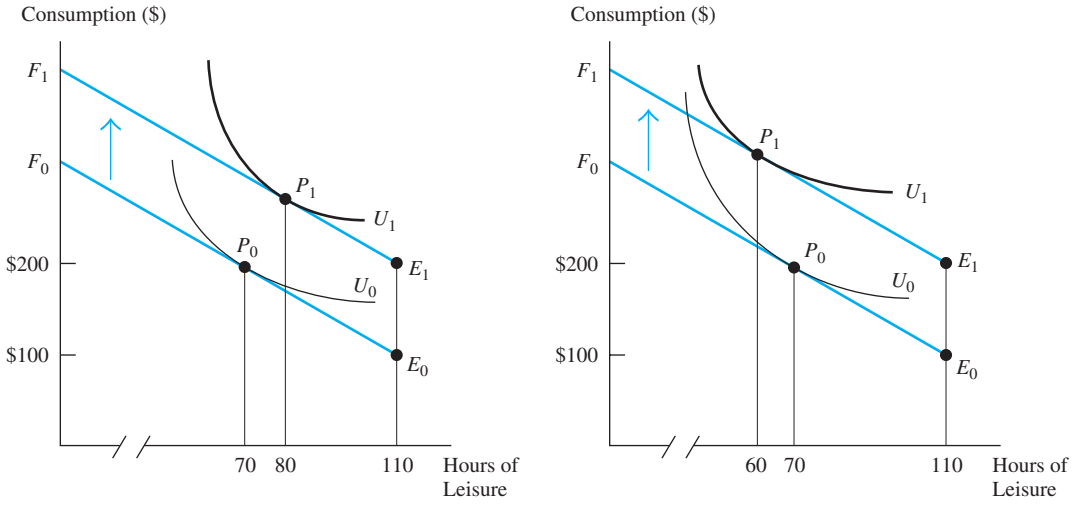
The increase in nonlabor income to \$200 weekly shifts the endowment point to E_1, so that the new budget line is given by F_1E_1. Because the worker's wage rate is being held constant, the slope of the new budget line is the same as the slope of the budget line that originated at point E_0. An increase in nonlabor income that holds the wage constant expands the worker's opportunity set through a parallel shift in the budget line.

The increase in nonlabor income allows the worker to jump to a higher indifference curve, such as point P_1 in Figure 2-7. Increases in nonlabor income necessarily make the worker better off. After all, the expansion of the opportunity set opens up many additional opportunities for the worker. Figure 2-7*a* draws point P_1 so that the additional nonlabor income increases both purchases on goods and leisure hours. As a result, the length of the workweek falls to 30 hours. Figure 2-7*b* draws point P_1 so that the additional nonlabor income reduces leisure hours, increasing the length of the workweek to 50 hours. The impact of the change in nonlabor income (holding wages constant) on the number of hours worked is called an **income effect.**

Both panels in Figure 2-7 draw "legal" indifference curves. The indifference curves are downward sloping, do not intersect, and are convex to the origin. We cannot predict how an increase in nonlabor income affects hours of work unless we make an additional restriction on the shape of indifference curves. The additional restriction we make is that leisure is a "normal" good (as opposed to leisure being an "inferior" good).

We define a commodity to be a normal good when increases in income, holding the prices of all goods constant, increase its consumption. A commodity is an inferior good

FIGURE 2-7 The Effect of a Change in Nonlabor Income on Hours of Work
An increase in nonlabor income leads to a parallel, upward shift in the budget line, moving the worker from point P_0 to point P_1. (*a*) If leisure is a normal good, hours of work fall. (*b*) If leisure is an inferior good, hours of work rise.



(*a*) Leisure Is a Normal Good

(*b*) Leisure Is an Inferior Good

when increases in income, holding prices constant, decrease its consumption. Low-priced subcompact cars, for instance, are typically thought of as inferior goods, whereas BMWs are typically thought of as normal goods. In other words, we would expect the demand for low-quality subcompacts to fall as nonlabor income increases, and the demand for BMWs to increase.

If we reflect on whether leisure is a normal or an inferior good, most of us would probably conclude that leisure is a normal good. Put differently, if we were wealthier, we would surely demand a lot more time off. We could then visit Aspen in December, Rio in February, and exotic beaches in the summer.

Because it seems reasonable to assume that leisure is a normal good and because there is some evidence (discussed below) supporting this assumption, our discussion focuses on this case. The assumption that leisure is a normal good resolves the conflict between the two panels in Figure 2-7 in favor of the one on the left-hand side. *The income effect, therefore, implies that an increase in nonlabor income, holding the wage rate constant, reduces hours of work.*

What Happens to Hours of Work When the Wage Changes?

Consider a wage increase from $10 to $20 an hour, holding nonlabor income V constant. The wage increase rotates the budget line around the endowment point, as illustrated in Figure 2-8. The rotation of the budget line shifts the opportunity set from *FE* to *GE*. It should be obvious that a wage increase does not change the endowment point: The dollar value of the goods that can be consumed when one does not work is the same regardless of whether the wage rate is $10 or $20 an hour.

The two panels presented in Figure 2-8 illustrate the possible effects of a wage increase on hours of work. In Figure 2-8*a*, the wage increase shifts the optimal consumption bundle from point *P* to point *R*. At the new equilibrium, the individual consumes more leisure (from 70 to 75 hours), so that hours of work fall from 40 to 35 hours.

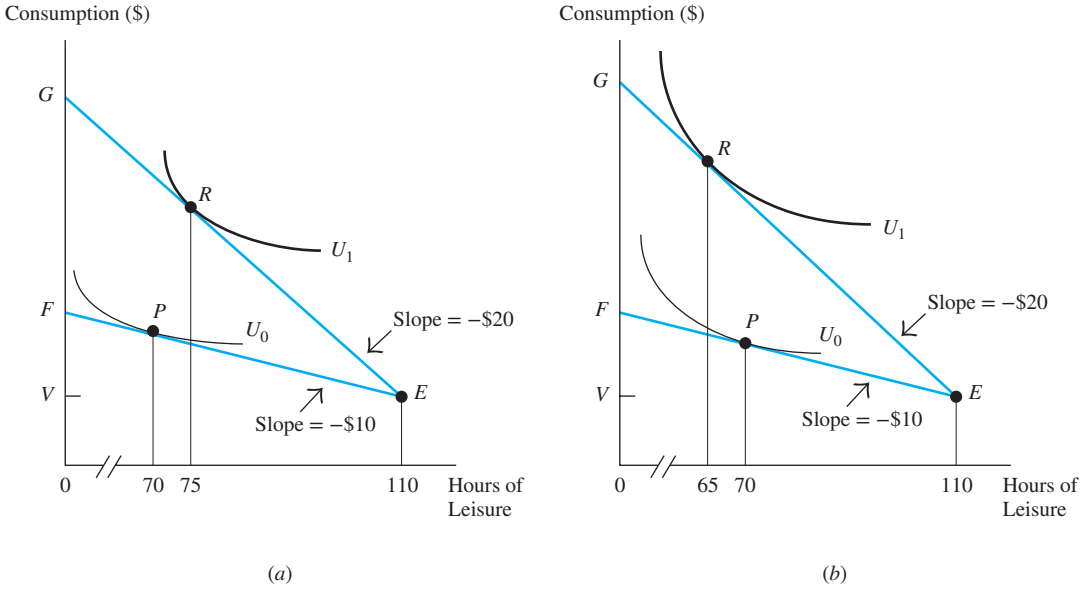
Figure 2-8*b*, however, shows the opposite result. The wage increase again moves the worker to a higher indifference curve and shifts the optimal consumption bundle from point *P* to point *R*. This time, however, the wage increase reduces leisure hours (from 70 to 65 hours), so the length of the workweek increases from 40 to 45 hours. It seems, therefore, that we cannot make an unambiguous prediction about an important question without making even more assumptions.

The reason for the ambiguity in the relation between hours of work and the wage rate is of fundamental importance and introduces tools and ideas that play a central role in all of economics. Both panels in Figure 2-8 show that, regardless of what happens to hours of work, a wage increase expands the worker's opportunity set. Put differently, a worker has more opportunities when she makes $20 an hour than when she makes $10 an hour. We know that an increase in income increases the demand for all normal goods, including leisure. The increase in the wage thus increases the demand for leisure, which reduces hours of work.

But this is not all that happens. The wage increase also makes leisure more expensive. When the worker earns $20 an hour, she gives up $20 every time she decides to take an hour off. Leisure time is a very expensive commodity for high-wage workers and is relatively cheap for low-wage workers. High-wage workers would have strong incentives to cut

FIGURE 2-8 The Effect of a Change in the Wage Rate on Hours of Work

A change in the wage rate rotates the budget line around the endowment point *E*. A wage increase moves the worker from point *P* to point *R*, and can either decrease or increase hours of work.



back on their consumption of leisure. A wage increase thus reduces the demand for leisure and increases hours of work.

This discussion highlights the source of the ambiguity in the relation between hours of work and the wage rate. A high-wage worker wants to enjoy the rewards of her high income, and would like to consume more leisure. The same worker, however, finds that leisure is very expensive and that she simply cannot afford to take time off from work.

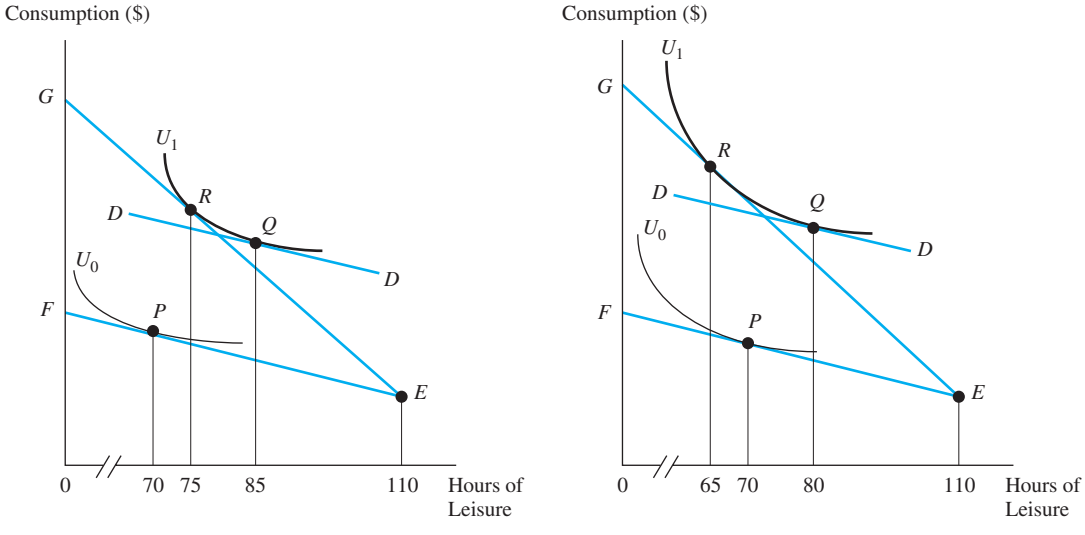
These two conflicting forces are illustrated in Figure 2-9*a*. The initial wage rate is \$10 per hour. The worker maximizes her utility by choosing the consumption bundle given by point *P,* where she consumes 70 hours of leisure and works 40 hours per week. Suppose the wage increases to \$20. The budget line rotates and the new consumption bundle is given by point *R.* The worker is now consuming 75 hours of leisure and working 35 h. As drawn, the person is working fewer hours at the higher wage.

It helps to think of the move from point *P* to point *R* as a two-stage move. The two stages correspond exactly to our discussion that the wage increase generates two effects: It increases the worker's income and it raises the price of leisure. In particular, suppose we draw a budget line that is parallel to the old budget line (so that its slope is also −\$10), but tangent to the new indifference curve. This budget line (*DD*), also illustrated in Figure 2-9*a*, generates a new tangency point *Q*.

The move from initial position *P* to final position *R* can then be decomposed into a first-stage move from *P* to *Q* and a second-stage move from *Q* to *R*. It is easy to see that the move from point *P* to point *Q* is an income effect. In particular, the move from *P* to *Q* arises from a change in the worker's income, holding wages constant. The income effect

FIGURE 2-9 Income and Substitution Effects

An increase in the wage rate generates both income and substitution effects. The income effect (the move from point *P* to point *Q*) reduces hours of work; the substitution effect (the move from *Q* to *R*) increases hours of work.



(*a*) Income Effect Dominates

(*b*) Substitution Effect Dominates

gives the change in the consumption bundle induced by the additional income resulting from the wage increase. Because both leisure and goods are normal goods, point Q must lie to the northeast of point P (so that more is consumed of both goods and leisure). The income effect increases the demand for leisure (from 70 to 85 hours) and reduces hours of work by 15 hours per week.

The second-stage move from Q to R is called the **substitution effect.** It illustrates what happens to the worker's consumption bundle as the wage increases, holding utility constant. By moving along an indifference curve, the worker's utility or "real income" is held fixed. The substitution effect isolates the impact of the increase in the price of leisure on hours of work, holding real-income constant.

The move from point Q to point R shows a substitution away from leisure and toward goods. In other words, as the wage rises, the worker devotes less time to expensive leisure activities (from 85 to 75 hours) and increases her consumption of goods. Through the substitution effect, therefore, hours of work rise by 10 hours. *The substitution effect implies that an increase in the wage rate, holding real income constant, increases hours of work.*

As drawn in Figure 2-9*a*, the decrease in hours of work generated by the income effect (15 hours) exceeds the increase in hours of work associated with the substitution effect (10 hours). The stronger income effect thus leads to a negative relationship between hours of work and the wage rate. In Figure 2-9*b*, the income effect (again the move from point P to point Q) decreases hours of work by 10 hours, whereas the substitution effect (the move from Q to R) increases hours of work by 15 hours. Because the substitution effect dominates, there is a positive relationship between hours of work and the wage rate.

The reason for the ambiguity in the relationship between hours of work and the wage rate should now be clear. As the wage rises, a worker faces a larger opportunity set and that income effect increases her demand for leisure and reduces hours of work. As the wage rises, however, leisure becomes more expensive and the substitution effect encourages the worker to switch away from the consumption of leisure and instead consume more goods. This shift frees up leisure hours and increases hours of work.

To summarize:

- An increase in the wage rate increases hours of work if the substitution effect dominates the income effect.
- An increase in the wage rate decreases hours of work if the income effect dominates the substitution effect.

2-6 To Work or Not to Work?

Our analysis of the relation between nonlabor income, the wage rate, and hours of work assumed that the person worked both before and after the change in nonlabor income or the wage. Hours of work then adjusted to the change in the opportunity set. But what factors motivate a person to work in the first place?

To illustrate the nature of this decision, consider Figure 2-10. The figure draws the indifference curve that goes through the endowment point E. This indifference curve indicates that a person who does not work at all receives U_0 units of utility. The woman, however, can choose to enter the labor market and trade some of her leisure time for earnings that will allow her to buy goods. The decision of whether to work or not boils down to

Theory at Work

DOLLARS AND DREAMS

The fact that our consumption of leisure responds to its price is not surprising. When the wage rate is high, we will find ways of minimizing the use of our valuable time. We will go through a ticket broker and pay high prices for concert and theater tickets, rather than stand in line for hours to buy a ticket at face value. We will hire a nanny or send our children to day care, rather than withdraw from the labor market. And we will consume preprepared meals and order pizza or take-out Chinese, rather than engage in lengthy meal preparations.

It turns out that how we allocate our time responds to economic incentives even when there are no easy substitutes available, such as when we decide how many hours to sleep. Sleeping takes a bigger chunk of our time than any other activity. The typical person sleeps around 57 hours a week. Although most of us believe that how long we sleep is biologically (and perhaps even culturally) determined, there is evidence that hours sleeping can also be viewed as another activity that responds to economic incentives. As long as some minimum biological threshold for the length of a sleeping spell is met, the demand for sleep time seems to respond to changes in the price of time.

In particular, there is a negative correlation between a person's earnings capacity and the number of hours spent sleeping. More highly educated persons, for example, sleep less—an additional four years of school reduces sleep time by about an hour per week. Similarly, a 20 percent wage increase reduces sleep time by 1 percent, or about 34 minutes per week. Even dreaming of a nice vacation in a remote island becomes expensive when our time is valuable.

Source: Jeff E. Biddle and Daniel S. Hamermesh, "Sleep and the Allocation of Time," *Journal of Political Economy* 98 (October 1990): 922–943.

a simple question: Are the "terms of trade"—the rate at which leisure can be traded for goods—sufficiently attractive to bribe her into entering the labor market?

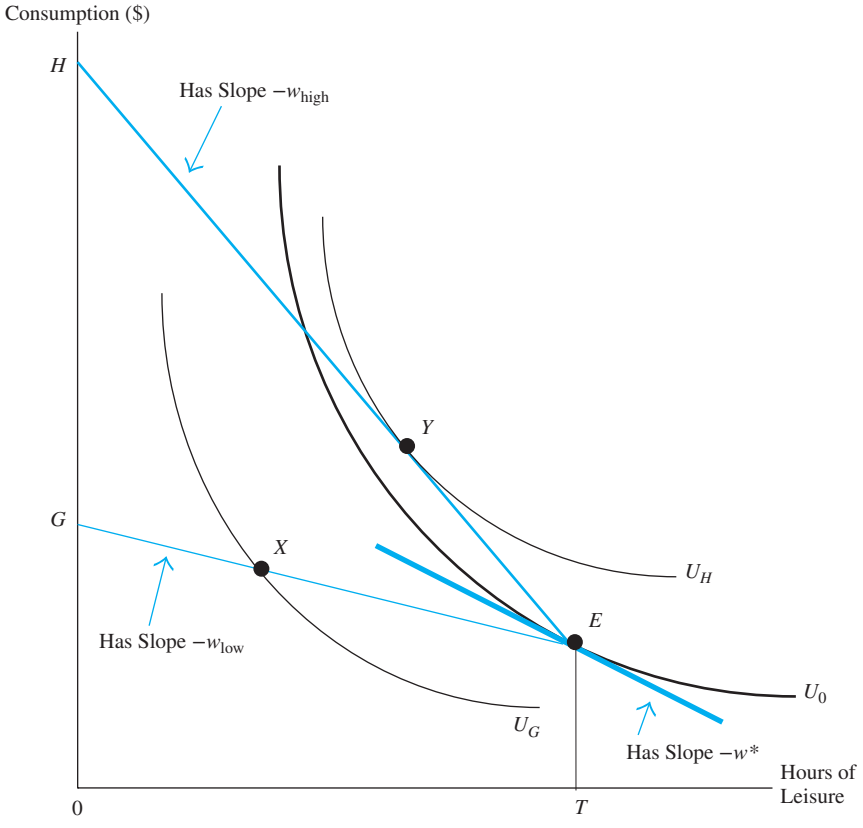
Suppose initially that the person's wage rate is given by w_{low} so that the woman faces budget line *GE* in Figure 2-10. No point on this budget line can give her more utility than U_0. At this low wage, the person's opportunities are quite meager. If the worker were to move from the endowment point *E* to any point on the budget line *GE,* she would be moving to a lower indifference curve. For example, at point *X* the woman gets only U_G utils. At wage w_{low}, therefore, the woman chooses not to work.

In contrast, suppose that the wage rate was given by w_{high}, so that the woman faces budget line *HE.* Moving to any point on this steeper budget line would increase her utility. At point *Y,* the woman gets U_H utils. At the wage w_{high}, therefore, the woman is better off working.

In sum, Figure 2-10 indicates that the woman does not work at low-wage rates (such as w_{low}), but does work at high-wage rates (such as w_{high}). As we rotate the budget line from wage w_{low} to wage w_{high}, we will typically encounter a wage rate, call it w^*, that makes her indifferent between working and not working. We call w^* the **reservation wage.** The reservation wage gives the minimum increase in income that would make a person indifferent between remaining at the endowment point *E* and working that first hour. It is given by the absolute value of the slope of the indifference curve at point *E.*

The definition of the reservation wage implies that the person will not work when the market wage is less than the reservation wage; but the person will work when the market

FIGURE 2-10 The Reservation Wage
If the person chooses not to work, she can remain at the endowment point E and get U_0 units of utility. At a low wage (w_{low}), the person is better off not working. At a high wage (w_{high}), she is better off working. The reservation wage w^* is given by the slope of the indifference curve at the endowment point.



wage exceeds the reservation wage. The decision to work, therefore, depends entirely on a comparison of the market wage, which indicates how much employers are willing to pay for an hour of work, and the reservation wage, which indicates how much the worker requires to be bribed into working that first hour.

The theory obviously implies that a high reservation wage makes it less likely that a person will work. The reservation wage will typically depend on the person's tastes for work, which helps to determine the slope of the indifference curve, as well on many other factors. For instance, the assumption that leisure is a normal good implies that the reservation wage rises as nonlabor income increases.[8] Because workers want to consume more leisure as nonlabor income increases, a larger bribe will be required to convince a wealthier person to enter the labor market.

[8] Try to prove this statement by drawing a vertical line through the endowment point in Figure 2-6. Because of convexity, the indifference curves will get steeper as we move to higher indifference curves.

Holding the reservation wage constant, the theory also implies that high-wage persons are more likely to work. A rise in the wage rate, therefore, increases the labor force participation rate of a group of workers. As we shall see, this positive correlation between wage rates and labor force participation helps explain the rapid increase in the labor force participation rate of women observed in the United States and in many other countries in the past century.[9]

In sum, the theory predicts a positive relation between the person's wage rate and her probability of working. It is of interest to contrast this strong prediction with our earlier result that a wage increase has an ambiguous effect on hours of work, depending on whether the income or substitution effect dominates.

The disparity arises because an increase in the wage generates an income effect *only if the person is already working.* A wage increase from $10 to $20 per hour for a person working 40 hours per week makes leisure more expensive (so that she wants to work more) *and* makes the person wealthier (so that she wants to work less). In contrast, if the person is not working at all, an increase in the wage rate has no effect on her real income. The quantity of goods that a nonworker can buy is independent of whether her potential wage rate is $10 or $20 an hour. An increase in the potential wage of a nonworker, therefore, does not generate an income effect. It simply makes leisure time more expensive and is likely to draw the nonworker into the labor force.

2-7 The Labor Supply Curve

The predicted relation between hours of work and the wage rate is called the **labor supply curve.** Figure 2-11 illustrates how a person's labor supply curve can be derived from the utility-maximization problem that we solved earlier.

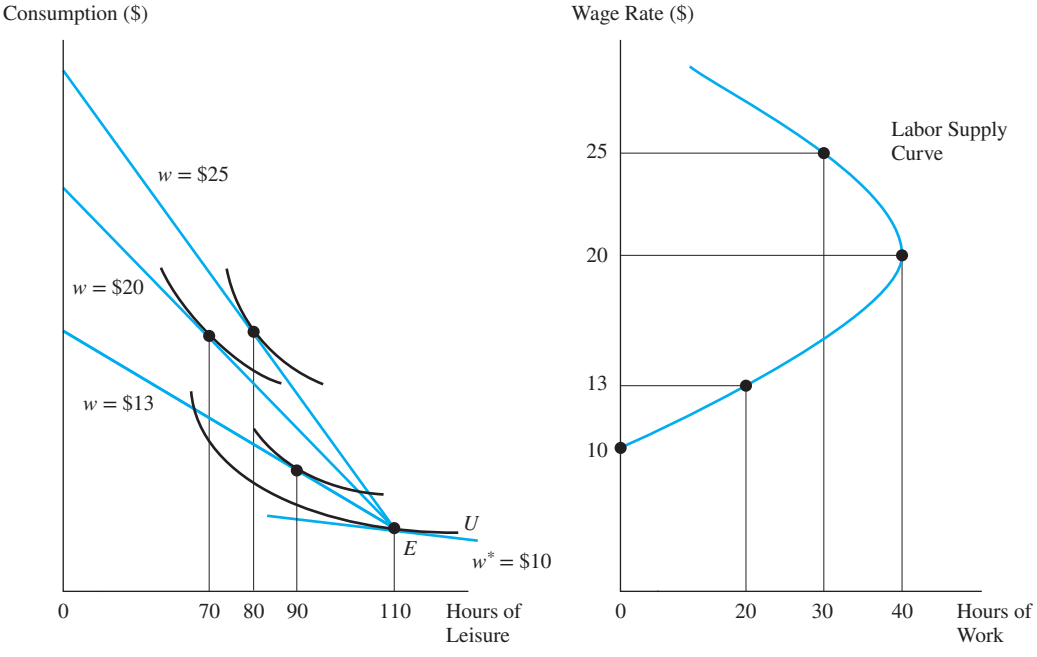
The left panel of the figure shows the person's optimal consumption bundle at a number of alternative wage rates. As drawn, the reservation wage is $10, the wage at which she is indifferent between working and not working. This person, therefore, supplies zero hours to the labor market at any wage less than or equal to $10. Once the wage rises above $10, the person chooses to work for some hours. For example, she works for 20 hours when the wage is $13; 40 hours when the wage is $20; and 30 hours when the wage is $25. Note that, as drawn, the figure implies that substitution effects dominate at lower wages and that income effects dominate at higher wages.

The right panel of the figure traces out the labor supply curve, the relation between hours worked and the wage rate. Initially, the labor supply curve is positively sloped. Once the wage rises above $20, however, the income effect dominates and hours of work decline as the wage rises, creating a segment of the labor supply curve that has a negative slope. The type of labor supply curve illustrated in Figure 2-11*b* is called a *backward-bending* labor supply curve because it eventually bends around and has a negative slope.

[9] The modern analysis of labor force participation decisions within an economic framework began with the classic work of Jacob Mincer, "Labor Force Participation of Married Women," in H. Gregg Lewis, editor, *Aspects of Labor Economics,* Princeton, NJ: Princeton University Press, 1962, pp. 63–97. An important study that stresses the comparison between reservation and market wages is given by James J. Heckman, "Shadow Prices, Market Wages and Labor Supply," *Econometrica* 42 (July 1974): 679–694.

FIGURE 2-11 Deriving a Labor Supply Curve for a Worker

The labor supply curve traces out the relationship between the wage rate and hours of work. At wages below the reservation wage ($10), the person does not work. At wages higher than $10, the person enters the labor market. The upward-sloping segment of the labor supply curve implies that substitution effects are stronger initially; the backward-bending segment implies that income effects may dominate eventually.



(*a*) Optimal Consumption Bundles

(*b*) Relation between Optimal Hours of Work and the Wage Rate

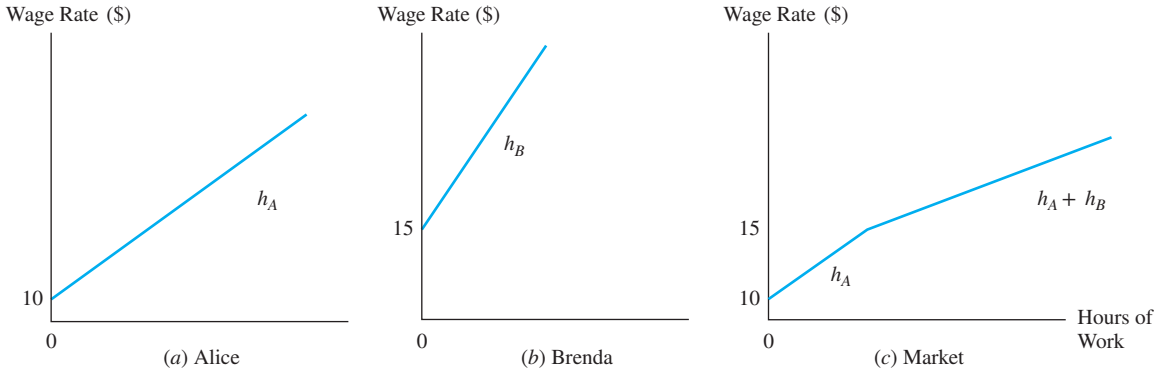
We can use the utility-maximization framework to derive a labor supply curve for every person in the economy. The labor supply curve in the aggregate labor market is then given by adding up the hours that all persons in the economy are willing to work at a given wage. Figure 2-12 illustrates how this "adding up" is done in an economy with two workers, Alice and Brenda. Alice's reservation wage is $15 and Brenda's is $20. It should be clear that no one would work if the wage is below $15, and that only Alice would work if the wage is between $15 and $20. At wages higher than $20, market labor supply is given by the total number of hours worked by Alice and Brenda, or $h_A + h_B$. The labor supply curve in the market, therefore, is obtained by adding up the supply curves of all workers *horizontally.*

To measure the responsiveness of hours of work to changes in the wage rate, we define the **labor supply elasticity** as

$$\sigma = \frac{\Delta h/h}{\Delta w/w} = \frac{\Delta h}{\Delta w} \cdot \frac{w}{h} \qquad \textbf{(2-11)}$$

The labor supply elasticity σ gives the percentage change in hours of work associated with a 1 percent change in the wage rate. The sign of the labor supply elasticity depends

FIGURE 2-12 Derivation of the Market Labor Supply Curve
The market labor supply curve "adds up" the supply curves of individual workers. When the wage is below $15, no one works. At a wage of $15, Alice enters the labor market. If the wage rises above $20, Brenda also enters the market.



on whether the labor supply curve is upward sloping ($\Delta h/\Delta w > 0$) or downward sloping ($\Delta h/\Delta w < 0$), and is positive when substitution effects dominate and negative when income effects dominate. Hours of work are more responsive to changes in the wage the greater the absolute value of the labor supply elasticity.

To see how the labor supply elasticity is calculated, consider the following example. Suppose that the worker's wage is initially $10 per hour and that she works 1,900 hours per year. The worker gets a raise to $20 per hour, and she decides to work 2,090 hours per year. This worker's labor supply elasticity can then be calculated as

$$\sigma = \frac{\%\ \Delta h}{\%\ \Delta w} = \frac{10\%}{100\%} = 0.1 \qquad \textbf{(2-12)}$$

A labor supply curve is said to be *inelastic* when the labor supply elasticity is less than one in absolute value. In other words, there is relatively little change in hours of work for a given change in the wage rate. If the labor supply elasticity is greater than one in absolute value—indicating that hours of work are greatly affected by the change in the wage—the labor supply curve is said to be *elastic*. Labor supply is inelastic in the numerical example in equation (2-12); a doubling of the wage (a 100 percent increase) raised labor supply by only 10 percent.

2-8 Estimates of the Labor Supply Elasticity

Few topics in applied economics have been as thoroughly researched as the empirical relationship between hours of work and wages. We begin our review of this literature by focusing on the estimates of the labor supply elasticity for men. The typical study uses the sample of working men to correlate a particular person's hours of work with his wage rate and nonlabor income. In particular, the generic regression model estimated in these studies is

$$h_i = \beta w_i + \gamma V_i + \text{Other variables} \qquad \textbf{(2-13)}$$

where h_i gives the number of hours that person i works; w_i gives his wage rate; and V_i gives his nonlabor income. The coefficient β measures the impact of a one-dollar wage increase

on hours of work, holding nonlabor income constant; and the coefficient γ measures the impact of a one-dollar increase in nonlabor income, holding the wage constant. The sign of the coefficient β depends on whether income or substitution effects dominate: β is negative if income effects dominate and positive if substitution effects dominate. The estimate of β can be used to calculate the labor supply elasticity defined by equation (2-11). Assuming leisure is a normal good, the theory also predicts that the coefficient γ should be negative because workers with more nonlabor income consume more leisure.

There is a lot of variation in existing estimates of the labor supply elasticity. Some studies report the elasticity to be zero; other studies report it to be large and negative; still others report it to be large and positive. There have been some attempts to determine which estimates are most credible.[10] These surveys suggest that the elasticity of male labor supply is roughly around -0.1. In other words, a 10 percent increase in the wage leads, on average, to a 1 percent decrease in hours of work for men. In terms of the decomposition into income and substitution effects, a 10 percent increase in the wage raises hours of work by about 1 percent because of the substitution effect, but also leads to a 2 percent decrease because of the income effect.

Three points are worth noting about the -0.1 estimate of the labor supply elasticity. First, it is negative, so income effects dominate. As a result, the observed decline in hours of work between 1900 and 2000 is often attributed to the income effects resulting from rising real wages.[11] Second, the labor supply curve is inelastic. Hours of work for men do not seem to be very responsive to changes in the wage. One would not be stretching the truth too much by claiming that the male labor supply elasticity is essentially zero. After all, most prime-age men work a full workweek every week of the year.[12] And, third, it is important to keep in mind that the labor supply elasticity probably differs greatly between men and women and between younger and older workers.

Problems with the Estimated Elasticities

It turns out that much of the empirical research is marred by a number of statistical and measurement problems. In fact, each of the three variables that are crucial for estimating the labor supply model—the person's hours of work, the wage rate, and nonlabor income—introduces difficult estimation problems.

[10] Richard Blundell and Thomas MaCurdy, "Labor Supply: A Review of Alternative Approaches," in Orley C. Ashenfelter and David Card, editors, *Handbook of Labor Economics,* vol. 3A, Amsterdam: Elsevier, 1999, pp. 1559–1695. Some of the variation in the arises because of conceptual differences in what is meant by the "labor supply elasticity"; see Raj Chetty, John N. Friedman, Tore Olsen, and Luigi Pistaferri, "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records," *Quarterly Journal of Economics* 126 (May 2011): 749–804.

[11] Thomas J. Kniesner, "The Full-Time Workweek in the United States: 1900–1970," *Industrial and Labor Relations Review* 30 (October 1976): 3–15. In recent years, hours of work have begun to rise for highly educated men, perhaps due to a strong substitution effect resulting from a rapidly rising real wage; see Peter Kuhn and Fernando Luzano, "The Expanding Workweek? Understanding Trends in Long Work Hours among U.S. Men, 1979–2006," *Journal of Labor Economics* 26 (April 2008): 311–343.

[12] Recall, however, that the labor force participation rate of men fell throughout much of the twentieth century; see Chinhui Juhn, "The Decline of Male Labor Market Participation: The Role of Market Opportunities," *Quarterly Journal of Economics* 107 (February 1992): 79–121.

Hours of Work

What precisely do we mean by hours of work when we estimate a labor supply regression: Is it hours of work per day, per week, or per year? The elaborate theoretical apparatus that we developed does not tell us what the span of the time period should be. Not surprisingly, the observed responsiveness of hours of work to a wage change depends crucially on whether we look at a day, a week, or a year. The labor supply curve becomes more elastic the longer the time period over which the hours-of-work variable is defined, so labor supply is almost completely inelastic if we analyze hours of work per week, but it is a bit more responsive if we analyze hours of work per year. The conclusion that the labor supply elasticity may be around -0.1 is based on studies that look at variation in annual hours of work.

There is also substantial measurement error in the hours-of-work measure that is typically available in survey data.[13] Workers who are paid by the hour know quite well how many hours they worked last week; after all, their take-home pay depends on the length of the workweek. Many of us, however, are paid an annual salary and we make little effort to track exactly how many hours we work in any given week. When we are asked how many hours we work, many of us will respond "40 hours" because that is the easy answer. Actual hours of work, however, may have little to do with the mythical 40-hour workweek for many salaried workers. As we will see shortly, this measurement error biases the regression estimates of the labor supply elasticity.

The Wage Rate

The typical salaried worker is paid an annual salary, regardless of how many hours she puts into her job, so that survey data does not typically report an hourly wage rate. It is instead customary to define the wage rate as the average wage, the ratio of annual earnings to annual hours worked. This calculation transmits any measurement errors in the reported measure of hours of work to the measure of the hourly wage rate.

To illustrate the problem introduced by these measurement errors, suppose that a worker overreports her hours of work. Because of the way the wage rate is constructed (that is, as the ratio of annual earnings to annual hours of work), the denominator of this ratio is too large and we calculate an artificially low-wage rate. High reported hours of work are then associated with low-wage rates, generating a spurious negative correlation between hours and average wages. Suppose instead that the worker underreports her hours of work. The constructed wage rate is then artificially high, again generating a spurious negative correlation between hours of work and the wage. In short, measurement error exaggerates the importance of income effects.[14]

Even if there were no measurement error, there is still a conceptual problem in defining the hourly wage rate as the ratio of annual earnings to annual hours of work for salaried workers. The correct price of leisure in our model is the marginal wage, the increase in earnings associated with an additional hour of work. The relevant marginal wage for salaried workers may have little to do with the average wage earned per hour.

[13] John Bound, Charles Brown, Greg Duncan, and Willard Rogers, "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data," *Journal of Labor Economics* 12 (July 1994): 345–368.

[14] George J. Borjas, "The Relationship between Wages and Weekly Hours of Work: The Role of Division Bias," *Journal of Human Resources* 15 (Summer 1980): 409–423.

Finally, anyone who attempts to estimate the labor supply regression in equation (2-13) quickly encounters the problem that the wage rate is not observed for people who are not working. However, a person who is not working does not really have a zero wage rate. All that we really know is that this person's wage is below the reservation wage. Many studies avoid the problem by simply throwing the nonworkers out of the sample that is used for calculating the labor supply elasticity.

This procedure, however, is fundamentally flawed. The decision of whether to work depends on a comparison of market wages and reservation wages. Persons who do not work have either very low-wage rates or very high-reservation wages. The sample of workers, therefore, is *not* a random sample of the population. Because most of the econometric techniques and statistical tests assume that the sample under analysis is a random sample, these methods cannot be used to examine the labor supply behavior of a sample that only includes workers. As a result, the estimated labor supply elasticities are not calculated correctly. This problem is typically referred to as "selection bias."[15]

Nonlabor Income

We would ideally like V to measure that part of the worker's income stream that has nothing to do with how many hours she works. For most people, however, the current level of nonlabor income partly represents the returns to past savings and investments. Suppose that some workers have a "taste for work." They worked long hours, had high-labor earnings, and were able to save and invest a large fraction of their income in the past. These are the workers who will have high levels of nonlabor income today. As long as the taste for work does not change too much over time, these are also the workers who will tend to work more hours today. The correlation between nonlabor income and hours of work will then be positive, simply because persons who have large levels of nonlabor income are the persons who tend to work many hours. Not surprisingly, some studies in the literature report that workers who have more nonlabor income work more hours. This finding would suggest either that leisure is an inferior good or that the biases introduced by the correlation between tastes for work and nonlabor income are sufficiently strong to switch the sign of the estimated income effect. As we will see shortly, the income effect is indeed negative when the worker's nonlabor income can be attributed to purely random shocks.

2-9 Household Production

The neoclassical model of labor–leisure choice assumed that a person can allocate her time to either leisure activities or to work in the labor market. Much of our time, however, is devoted to a different type of work, where we produce various commodities in the household or nonmarket sector. These commodities include such things as childbearing, cooking, and cleaning the house. Not surprisingly, there are differences in how men and women allocate their time among the labor market, the household sector, and leisure activities. Women allocate more hours to the household than men. In 2013, the average man allocated 33.8 hours per week

[15] The classic study of the selection bias problem is given by James J. Heckman, "Sample Selection Bias as a Specification Error," *Econometrica* 47 (January 1979): 153–162.

to "paid work," but only 11.8 hours to housework and child care. In contrast, the average woman allocated 23.9 hours to paid work, and 24.3 hours to housework and child care.[16]

Unlike hours worked in the labor market, hours worked in the household do not lead to higher earnings. The end-product of our household production (such as a well-behaved child or a good meal) is seldom sold in the marketplace. Instead, household production makes us better off because it yields commodities that we consume at home.[17] By examining how various household members allocate their time among various uses, we can address a number of questions. Most important, why do some household members specialize in paid work and other members specialize in household production?

The Household Production Function

Consider the two-person household of Jack and Jill. This married couple would like to maximize utility, which depends on the dollar value of the goods they can buy in the marketplace, and the dollar value of the commodities they produce in the household. To buy goods in the marketplace, Jack and Jill need cash, and the only way to get cash is to get a job. To consume household commodities, Jack and Jill must spend some time in household production. Suppose that Jack and Jill each have 10 hours a day to devote to both types of work activities (the other 14 hours are allocated to personal care and sleeping). How should Jack and Jill divide their 10-hour workday between the market and nonmarket sectors?

The **household production function** tells us how much household output Jack and Jill can generate for any given allocation of time. It is possible that Jack and Jill have different aptitudes for producing commodities in the household sector. Consider a very simple example. Suppose, in particular, that Jack produces $10 worth of output for each hour he spends in household production, while Jill produces $25 of output per hour.

Jack's hourly wage rate is $20. Figure 2-13*a* illustrates the budget line that Jack would face *if he were a single man.* If Jack devoted all of his 10 available hours to the labor market, he would be able to purchase $200 worth of market goods. Because Jack's marginal product in the household sector is only $10 per hour, he can produce only $100 worth of household commodities if he devotes all his time to household production. This argument derives the two corners of Jack's "single" budget line.

Suppose that Jill's hourly wage rate is $15. If Jill were single, she could allocate all of her available time to the market sector and purchase $150 worth of market goods. If she allocated all her time to the household sector, she could generate $250 worth of household commodities. Jill's "single" budget line is drawn in Figure 2-13*b*.

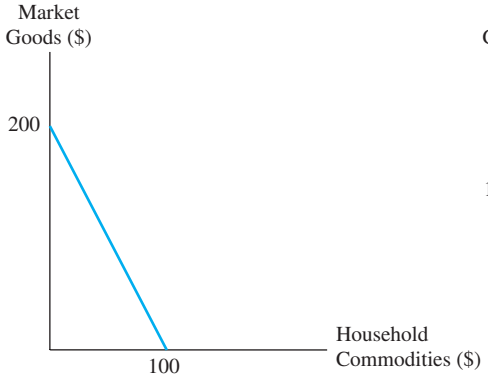
If Jack and Jill were a couple, they would no longer be constrained by these budget lines. The household's opportunity set would expand because each of them could specialize in the sector where they are relatively more productive. To see this, let's derive the

[16] Kim Parker and Wendy Wang, *Modern Parenthood: Roles of Moms and Dads Converge as They Balance Work and Family,* Washington, DC: Pew Research Center, 2013.

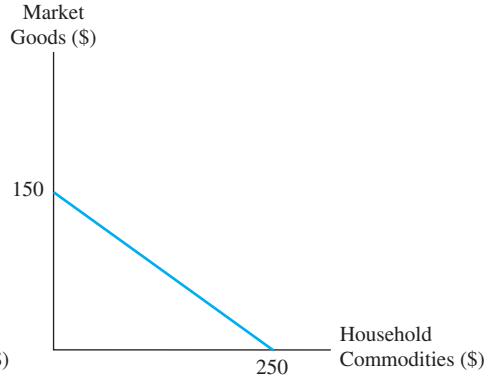
[17] The classic analysis of how we allocate our time among the various sectors is presented in Gary S. Becker, "A Theory of the Allocation of Time," *Economic Journal* 75 (September 1965): 493–517. See also Reuben Gronau, "Leisure, Home Production and Work—The Theory of the Allocation of Time Revisited," *Journal of Political Economy* 85 (December 1977): 1099–1123; and Francine D. Blau, Marianne A. Ferber, and Anne E. Winkler, *The Economics of Women, Men, and Work,* 7th Edition, Boston, MA: Pearson, 2013.

FIGURE 2-13 Budget Lines and Opportunity Frontier of Married Couple

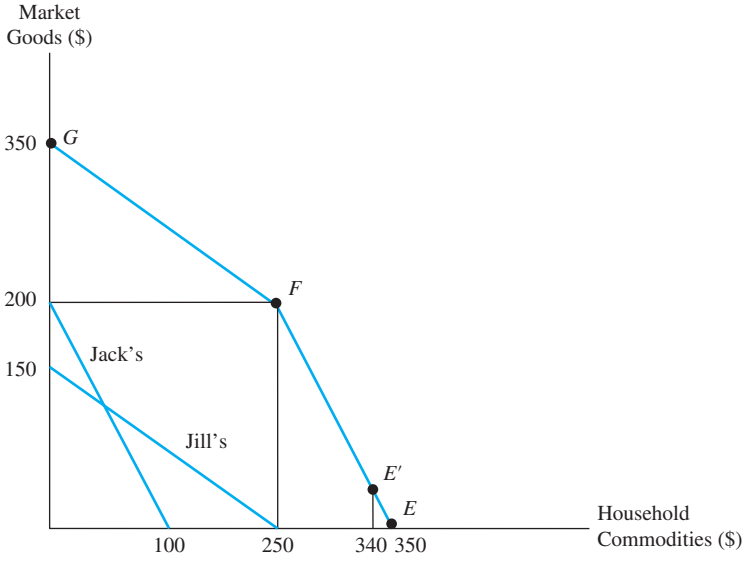
At point *E,* Jack and Jill allocate all their time to the household sector. If they wish to buy market goods, Jack gets a job because he has a relatively higher wage, generating segment *FE* of the frontier. After he allocates all his time to the labor market, Jill then gets a job, generating segment *GF* of the frontier.



(*a*) Jack's Single Budget Line



(*b*) Jill's Single Budget Line



(*c*) Jack and Jill's Opportunity Frontier

opportunity set for the Jack and Jill household. In particular, suppose that both Jack and Jill decide to allocate all of their time to the household sector. Jack could then produce $100 worth of household commodities (10 hours times $10 per hour) and Jill could generate $250 worth (10 hours times $25 per hour), for a total of $350. This joint decision generates point *E* in Figure 2-13*c*.

Suppose Jack and Jill want to buy some goods in the marketplace. They will have to get a job to generate the cash needed to make those purchases. But who should allocate that first hour to the labor market? If Jack allocates that first hour, the household gives up

$10 worth of household commodities and gains $20 worth of market goods. If Jill allocates that first hour, the household gives up $25 worth of household commodities and gains $15 worth of market goods. It makes economic sense for Jack and Jill to decide that it should be Jack who enters the labor market and works that first hour. For every dollar's worth of household commodities they give up, Jack gains $2 worth of market goods (or 20 ÷ 10). If Jill were to enter the labor market, for every dollar's worth of household commodities the household gives up, Jill can get only 60 cents worth of market goods (or 15 ÷ 25).

This exercise generates point E' on the household's opportunity frontier, the boundary of the set of choices available to the Jack and Jill household. Let's now consider what would happen if Jack and Jill wanted more market goods and decided to allocate a second hour to the labor market. The same arithmetic still applies: For every dollar's worth of household commodities they give up, they get more market goods if it is Jack who allocates that second hour to the labor market. In fact, given the numerical values in our example, Jack will always be the one chosen to allocate each additional hour to the labor market, until he has no more hours left. Jack exhausts his 10 available hours at point *F* in the figure, where Jack is devoting all 10 hours to the labor market and Jill is devoting all 10 hours to the household sector (allowing them to purchase $200 worth of market goods and $250 worth of household goods).

If the household then wished to buy even more market goods, Jill would have to enter the labor market. Because Jill has a relatively low-wage rate, however, each hour that Jill allocates to the labor market generates only $15 worth of market goods. As a result, the slope of the opportunity frontier flattens to the left of the "kink" at point *F.* In the end, if Jack and Jill allocate all their available time to the labor market, they would be able to purchase $350 worth of market goods, as at point *G.*

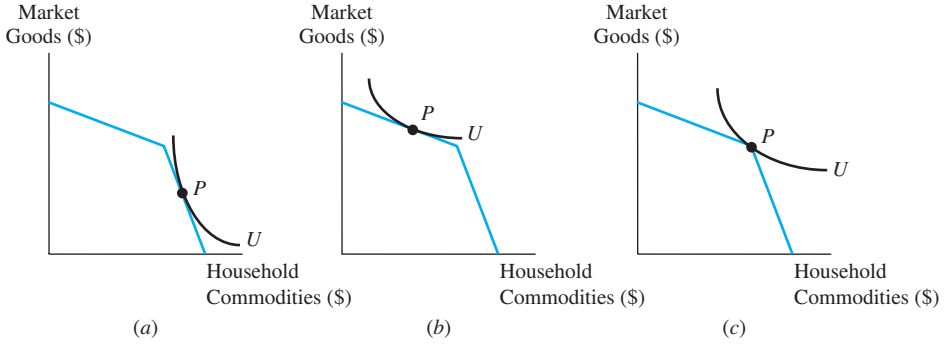
The household's opportunity set, therefore, is bounded by the frontier *GFE.* It is composed of two segments: a relatively steep segment (*FE*) where Jill devotes all her time to the household and Jack shares his time between the market and household sectors; and a flatter segment (*GF*) where Jack devotes all his time to the market sector and Jill is sharing her time between the market and the household.

Who Works Where?

Which point on the opportunity frontier will the household choose? A utility-maximizing household chooses the point that places the household on the highest possible indifference curve. Figure 2-14 shows three possible solutions. In Figure 2-14*a*, the household chooses a point along the steeper segment of the opportunity frontier (and gets *U* units of utility). In other words, Jack and Jill decide that Jill devotes all her time to household production and Jack divides his time between the market and household sectors. In Figure 2-14*b*, the household chooses a point along the flatter segment of the frontier. Jack now allocates all of his time to the market sector, and Jill divides her time between the market and the household. In Figure 2-14*c*, point *P* is located at the kink, and Jack and Jill completely specialize. He devotes all his time to the market sector; she devotes all her time to the household sector.

The discussion suggests that differences in the market wage among household members help determine the allocation of time within the household. Figure 2-15*a* shows how a large wage differential between husband and wife creates incentives for specialization. At the initial point *P,* Jack divides his time between the market and nonmarket sectors. Suppose that Jack's wage increased substantially. The steep segment of the opportunity

FIGURE 2-14 The Division of Labor in the Household
The indifference curve U is tangent to the opportunity frontier at point P. (*a*) Jill specializes in the household sector and Jack divides his time between the labor market and the household. (*b*) Jack specializes in the labor market and Jill divides her time between the two sectors. (*c*) Jack specializes in the labor market and Jill specializes in the household sector.



frontier would now become much steeper, as illustrated in the figure. If the wage increase were sufficiently large, the household would move from point P to point P', at the kink of the higher opportunity frontier. The wage increase encourages Jack to withdraw entirely from the household sector and to specialize in the market sector. This result is easy to understand. Even if both parties in the household were equally efficient at household production, the household could expand its opportunity set by having the person with the lowest wage rate devote the most time to household production.

The allocation of time also depends on the relative aptitudes of Jack and Jill for household production. Consider Figure 2-15*b*. At the initial point P, Jack allocates all his time to the market sector, and Jill divides her time between the two sectors. Suppose that Jill's productivity in the household sector increased substantially. The opportunity frontier shifts out, moving the household to the new kink point P' on the higher opportunity frontier. Jill now allocates all her time to household production.

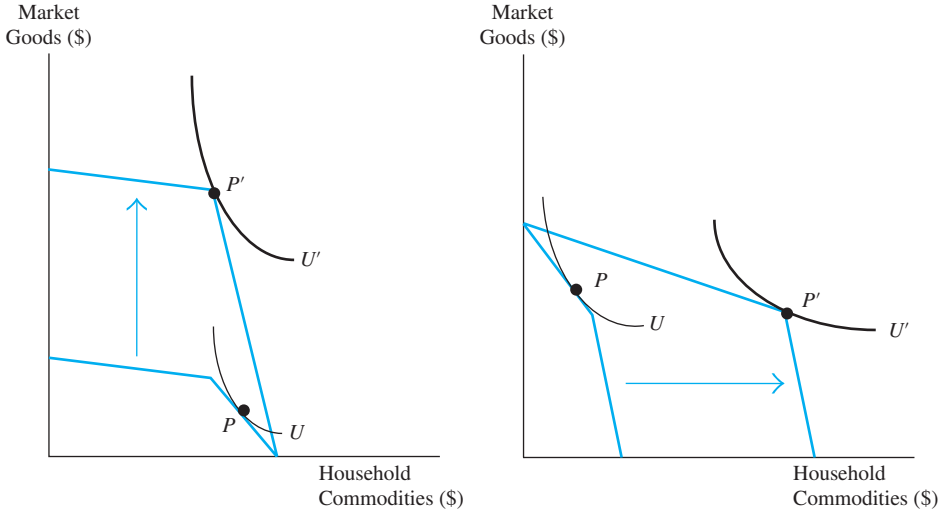
Trends in Female Labor Force Participation

The theory of time allocation between leisure, market work, and household production plays an important role in explaining trends in female labor force participation. As we saw earlier, the labor force participation rate of women increased dramatically in recent decades.

Our discussion emphasizes changes in the wage rate and in productivity in household production as key determinants of the increase in female labor force participation. As the wage rises, women have an incentive to reduce the time they allocate to the household sector and are more likely to enter the labor market. The real wage of women increased substantially in many countries in the past few decades, and there is strong evidence that participation rates grew fastest in those developed countries that experienced the highest wage increase.[18]

[18] Jacob Mincer, "Intercountry Comparisons of Labor Force Trends and of Related Developments: An Overview," *Journal of Labor Economics* 3 (January 1985, Part 2): S1–S32.

FIGURE 2-15 Increases in Wage Rate or Household Productivity Lead to Specialization
(*a*) An increase in Jack's wage moves the household from point P to P' and Jack specializes in the labor market. (*b*) An increase in Jill's household productivity moves the household from point P to P' and Jill specializes in the household sector.



(*a*) Increase in Jack's Wage

(*b*) Increase in Jill's Marginal Product in the Household

The labor force participation decision also depends on the value of a woman's time in household production. A fall in the number of children probably reduces the value of a woman's time in the household. Between 1950 and 2000, the total lifetime fertility of the average adult woman in the United States fell from 3.3 to 2.1 children, so the reduction in fertility probably contributed to the increase in female labor force participation.[19] At the same time, female participation rates were boosted by time-saving technological advances in household production, such as stoves, washing machines, and the microwave oven. The amount of time required to produce many household commodities was cut drastically, freeing up the scarce time for work in the labor market.

Of course, changes in cultural and legal attitudes toward working women also drove many women into the labor market. A fascinating example is that unmarried young women living in states that granted them an early right to obtain oral contraceptives without parental consent experienced a faster increase in labor force participation rates.[20]

Many studies estimate the responsiveness of female labor supply to changes in the wage rate. Unlike the estimate of the labor supply elasticity for prime-age men (that is, an elasticity

[19] U.S. Bureau of the Census, Statistical Abstract of the United States, Washington, DC: Government Printing Office, various issues; Joshua D. Angrist and William N. Evans, "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review* 88 (June 1998): 450–477.

[20] Martha J. Bailey, "More Power to the Pill: The Effect of Contraceptive Freedom on Women's Life Cycle Labor Supply," *Quarterly Journal of Economics* 121 (February 2006): 289–320.

of about −0.1), most studies of female labor supply find a *positive* correlation between a woman's hours of work and her wage rate, suggesting substitution effects dominate for working women. The size of the female labor supply elasticity, however, is not very large, perhaps around 0.2.[21]

Because of the huge increase in the number of women in the workforce in recent decades, there is a perception that female labor supply is more elastic than male labor supply. This perception, however, is mostly due to the fact that female labor force participation rates increased dramatically. The hours of work of working women, like those of working men, do not seem to be very responsive to changes in the wage. Put differently, female labor supply responds to economic factors mainly at the margin of deciding whether or not to work, rather than at the margin of deciding how many hours to work once in the labor force.

2-10 Correlation versus Causation: Searching for Random Shocks

As noted earlier, the correlation between hours of work and nonlabor income does not necessarily measure the income effect implied by our model of labor–leisure choice. High levels of nonlabor income today may reflect the returns to savings accumulated because we worked hard in the past, and the observed correlation could well be positive, rather than negative, if this "taste for work" persists over time. Similarly, a wage increase today may reflect the payoff to the work effort in past years, As long as the personality traits that lead to hard work persist over time, it is unlikely that the correlation between wages and hours of work today would measure the income and substitution effects implied by the model.

Put bluntly, the correlations estimated by the labor supply regression in equation (2-13) may provide no information whatsoever about how workers respond to income or wage changes. An increased appreciation for the fact that correlation does not imply causation led to a revolution in empirical labor economics in the past three decades, as researchers began to search for *random* increases in nonlabor income or in wages that would allow us to observe the impact of that random shock on a person's labor supply.

Much of the research attempting to estimate income effects has focused on situations where workers suddenly come across a large sum of money. Not surprisingly, this line of research has paid a great deal of attention to documenting the labor supply effects of winning a lottery.

In 1970, there were only two state lotteries in the United States. These lotteries sold $100 million in tickets during that year. By 2014, 43 states and the District of Columbia offered government-operated lotteries, selling $78 billion in tickets. The first prize in these lotteries often reaches astronomical amounts. The largest jackpot in U.S. history (the Powerball drawing held on January 13, 2016) divided a first prize of $1.6 billion among three lucky tickets.

The labor supply behavior of 1,000 lottery winners is revealing.[22] Nearly 25 percent of the winners left the labor force within a year, and an additional 9 percent reduced

[21] Francine D. Blau and Lawrence M. Kahn, "Changes in the Labor Supply of Married Women: 1980–2000," *Journal of Labor Economics* 25 (July 2007): 393–438; and Bradley T. Haim, "The Incredible Shrinking Elasticities: Married Female Labor Supply, 1978–2002," *Journal of Human Resources* 42 (Fall 2007): 881–918.

[22] Roy Kaplan, "Lottery Winners and Work Commitment: A Behavioral Test of the American Work Ethic," *Journal of the Institute for Socioeconomic Studies* 10 (Summer 1985): 82–94.

the number of hours they worked or quit a second job. Not surprisingly, the labor supply effects depended on the size of the jackpot. Only 4 percent of the winners who won a jackpot between $50,000 and $200,000 left the labor force, but nearly 40 percent of those whose jackpot exceeded $1 million retired to the "easy life." A more recent study focuses on the behavior of lottery winners in Massachusetts, confirming that the decline in labor supply resulting from new-found wealth is larger for those who win larger prizes.[23]

The labor supply of male taxi drivers provides an analogous opportunity to measure how workers respond to sudden and random changes in the wage rate.[24] In the pre-Uber days, the taxi industry in New York City was tightly regulated by the New York City Taxi and Limousine Commission (TLC). The TLC limited the number of licensed taxis in the city and set the fares. By examining the miles driven by a particular taxi driver before and after a fare change (which is a good measure of hours worked if driving speed is relative constant), it is possible to estimate the labor supply elasticity. In 1996, the TLC changed fares so that there was a 17 percent increase in revenue per mile driven. This change in the wage rate reduced the number of miles driven by a taxi driver by 3.2 percent, implying a labor supply elasticity is −0.19. The male labor supply elasticity is negative, implying that income effects dominate. And it is small, implying that labor supply is inelastic.

As we will see throughout the book, modern empirical research in labor economics follows this basic template of searching for random shocks that induce behavioral changes. These random shocks are sometimes due to luck (such as winning a lottery), or policy shifts, or even to human-made experiments where a random subset of the population is offered a particular set of opportunities while another subset is not. The typical empirical study then traces the impact of the random shock on the behavior of the targeted population to determine if the behavioral changes conform to the theoretical predictions.

Although this approach is obviously far better than interpreting correlations as estimates of the parameters that underlie the economic model, there is an important drawback. Are the results obtained by looking at lottery winners generalizable to a larger population? After all, not everyone buys lottery tickets in the first place. Does the labor supply response of taxi drivers tell us all that much about how a comparable wage increase would change hours worked for computer programmers, construction workers, or government bureaucrats? Tracing the behavioral response of a particular random shock measures precisely what happened as a result of *that* shock. Unfortunately, it is far from clear that the observed response can be generalized to other groups or other contexts.

[23] Guido W. Imbens, Donald B. Rubin, and Bruce Sacerdote, "Estimating the Effect of Unearned Income on Labor Supply, Earnings, Savings, and Consumption: Evidence from a Survey of Lottery Players," *American Economic Review* 91 (September 2001): 778–794; and David Cesarini, Erik Lindqvist, Matthew J. Notowidigdo, and Robert Östling, "The Effect of Wealth on Individual and Household Labor Supply: Evidence from Swedish Lotteries," *American Economic Review* 107 (December 2017): 3917–3946.

[24] Orley Ashenfelter, Kirk Doran, and Bruce Schaller, "A Shred of Credible Evidence on the Long-Run Elasticity of Labour Supply," *Economica* 77 (October 2010): 637–650.

2-11 Policy Application: Welfare Programs and Work Incentives

The impact of income maintenance programs, such as Temporary Assistance for Needy Families (TANF), on the work incentives of recipients has been hotly debated since the days when the United States declared a war on poverty in the mid-1960s. In fact, much of the opposition to welfare programs was motivated by the conjecture that these programs encourage recipients to "live off the dole" and foster dependency on public assistance. The perception that welfare does not work and that the so-called War on Poverty was lost found a sympathetic ear on all sides of the political spectrum and led to President Clinton's promise to "end welfare as we know it." This political consensus culminated in the enactment of the Personal Responsibility and Work Opportunity Reconciliation Act (PRWORA) in August 1996. This reform legislation included lifetime limits on the receipt of various types of welfare programs, tightened eligibility requirements for most families, and mandated that many benefit-receiving families engage in work-related activities.

Cash Grants and Labor Supply

To illustrate how welfare programs alter work incentives, let's first consider a simple program that grants eligible persons a cash grant. In particular, suppose that eligible persons (such as unmarried women with children) are given a check for, say, $1,000 per month as long as they remain outside the labor force. If these persons enter the labor market, the government officials assume that the women no longer need assistance and are dropped from the welfare rolls regardless of how much they earned.

Figure 2-16 illustrates the impact of this cash grant on work incentives. In the absence of the program, the budget line is given by *FE* and leads to an interior solution at point *P,* in which the person consumes 70 hours of leisure and works 40 h.

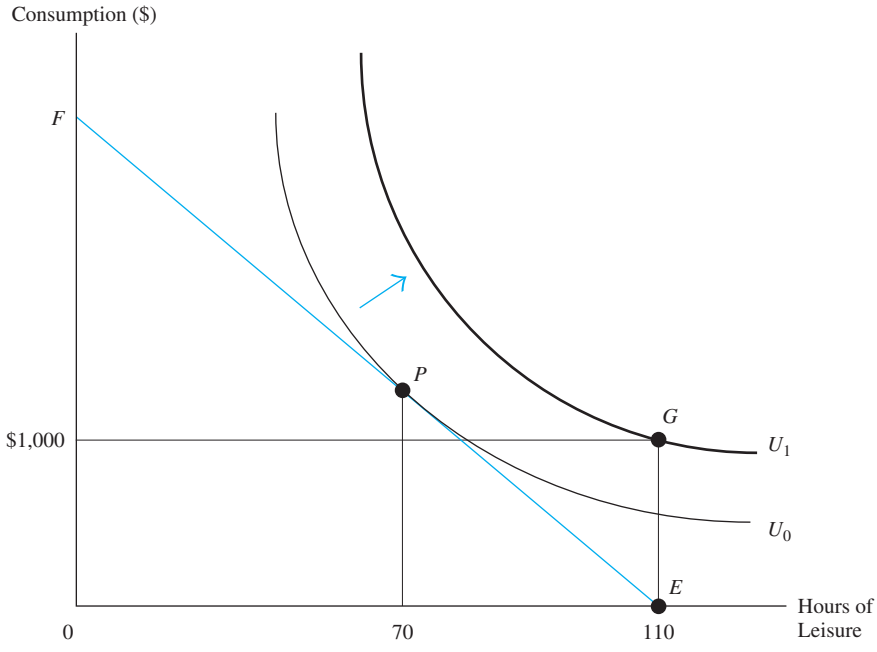
For simplicity, assume that the woman does not have any nonlabor income. The introduction of a cash grant of $1,000 to nonworkers then introduces point *G* into the opportunity set. At this point, the woman can purchase $1,000 worth of goods if she participates in the welfare program and does not work. Once the woman enters the labor market, however, the welfare grant is taken away and the opportunity set switches back to the original budget line *FE.*

The existence of the welfare benefit at point *G* can greatly reduce work incentives. As drawn, the woman attains a higher level of utility by choosing the corner solution at point *G* (that is, the welfare solution) than by choosing the interior solution at point *P* (that is, the work solution).

This type of "take-it-or-leave-it" cash grant can induce many workers to drop out of the workforce. It should be clear that low-wage women are most likely to choose the welfare solution. An improvement in the endowment point (from point *E* to point *G*) increases the worker's reservation wage, reducing the likelihood that a low-wage person will enter the labor market.

It is important to emphasize that welfare programs do not lower the work propensities of low-wage workers because these workers lack a "work ethic." After all, we have implicitly assumed that the preferences of low-wage workers (as represented by the family of indifference curves) are identical to the preferences of high-wage workers. Rather, the welfare program reduces the work incentives of low-wage workers because it is these workers who are most likely to find that the economic opportunities provided by the welfare system are better than those available in the labor market.

FIGURE 2-16 Effect of a Cash Grant on Work Incentives
A take-it-or-leave-it cash grant of $1,000 per month moves the worker from point *P* to *G,* and she leaves the labor force.



The Impact of Welfare on Labor Supply

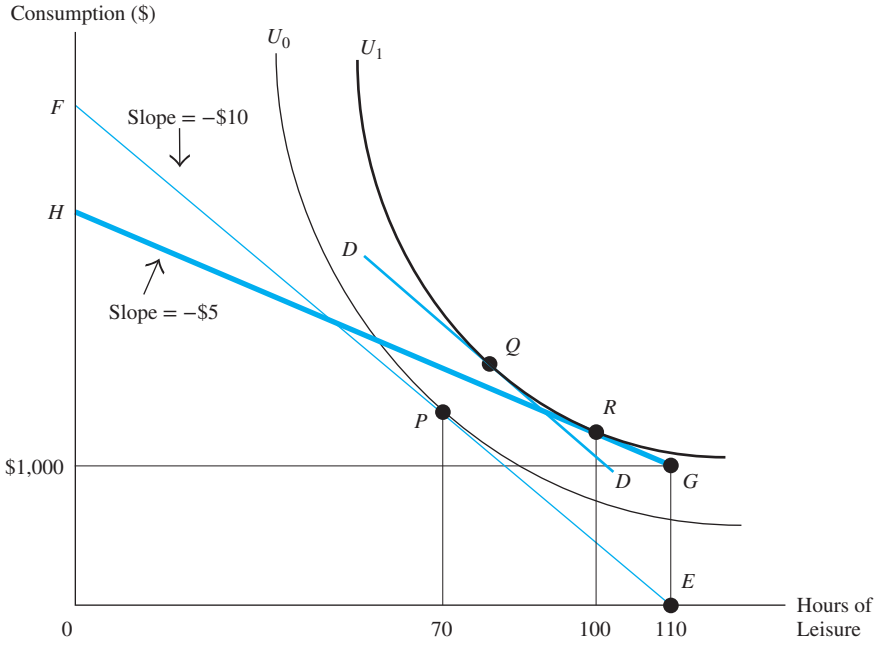
Because of the extreme disincentive effects of the program illustrated in Figure 2-16, real-world social assistance programs typically allow welfare recipients to remain employed. Although welfare recipients can work, the amount of the cash grant is often reduced by some amount for every dollar earned in the labor market. For example, in 2010 the TANF grant to a Maryland mother with two children who worked part-time at a minimum wage job was reduced by 60 cents for every dollar earned on the job.[25]

It is instructive to describe with a numerical example how this type of welfare program alters the person's opportunity set. Suppose that a woman's monthly income is $1,000 if she does not work at all and goes on welfare (assuming that she does not have any other nonlabor income). For the purpose of this example, suppose that the government takes away 50 cents from the cash grant for every dollar earned in the labor market. This means that, if the woman works one hour at a wage of $10, her labor earnings increase by $10 but her grant is reduced by $5. Her total income, therefore, is $1,005. If she decides to work 2 hours, her labor earnings are $20 but her grant is reduced by $10. Total income would then be $1,010. Every additional hour of work increases income by only $5. Under the guise of reducing the size of the welfare grant, the government is actually taxing the

[25] Austin Nichols and David Kassabian, "TANF Recipients' Implicit Tax Rates from Earnings Disregard Policies," Urban Institute, Washington, DC, 2011.

FIGURE 2-17 Effect of a Welfare Program on Hours of Work

The welfare program in budget line HG gives the worker a cash grant of $1,000 and imposes a 50 percent tax on labor earnings. In the absence of welfare, the worker is at point *P.* The income effect resulting from the program moves the worker to point *Q;* the substitution effect moves the worker to point *R.* Both income and substitution effects reduce hours of work.



welfare recipient's wage at a 50 percent rate. Therefore, it becomes important to differentiate between the woman's *actual* wage rate (which is $10 an hour) and the *net* wage (which is only $5 an hour.

Figure 2-17 shows the budget line created by this type of welfare program. In the absence of the program, the budget line is given by *FE* and the woman chooses point *P.* She would consume 70 hours of leisure and work 40 h.

The welfare program shifts the budget line in two important ways. Because of the $1,000 monthly grant when the woman does not work, the endowment point changes from point *E* to point *G.* The program also changes the slope of the budget line. We have seen that the reduction of the grant by 50 cents for every dollar earned is equivalent to a 50 percent tax on her earnings. The relevant slope of the budget line, therefore, is the net wage rate. Hence the welfare program cuts the (absolute value of the) slope by half, from $10 to $5. The budget line associated with the welfare program is given by *HG.*

As drawn, when given the choice between the budget line *FE* and the budget line *HG,* the woman opts for the welfare system and chooses the consumption bundle at point *R.* She consumes 100 hours of leisure and works 10 h. Even this "workfare" program, therefore, seems to create work disincentives because she works fewer hours than she would have worked in the absence of welfare.

In fact, we can demonstrate that a welfare program that includes both a cash grant and a tax on labor earnings *must* reduce hours of work. In particular, point *R* must be to the right of point *P.* To see why, draw a hypothetical budget line parallel to the prewelfare budget line, but tangent to the new indifference curve. This line is labeled *DD* in Figure 2-17. It is easy to see that the move from *P* to *Q* is an income effect and represents the impact of the cash grant on hours of work. This income effect increases the demand for leisure. In other words, point *Q* must be to the right of point *P.* The move from *Q* to *R* represents the substitution effect induced by the 50 percent tax on labor earnings, and point *R* must be to the right of point *Q.* The tax reduces the price of leisure by half for welfare recipients. As a consequence, the welfare recipient will demand even more leisure.

This stylized example neatly describes the work incentive problems introduced by welfare programs. If our model adequately represents how persons make labor supply decisions, it is impossible to formulate a relatively generous welfare program without substantially reducing work incentives. Awarding cash grants to recipients, as welfare programs inevitably do, reduces both the probability of a person working and the number of hours worked by those who remain on the job. In addition, efforts to recover some of the grant money as a worker accumulates labor earnings effectively impose a tax on work activities. This tax reduces the price of leisure and further lowers the number of hours that the welfare recipient will work.

The study of how welfare programs affect work incentives shows how the neoclassical model of labor–leisure choice serves as a point of departure for analyzing more complex situations. By incorporating details about how government policies alter a person's opportunity set, we can analyze important policy questions. The beauty of the economic approach is that we do not need different models to analyze labor supply decisions under alternative government policies or social institutions. In the end, we are always analyzing the *same* model—how workers allocate their limited time and money to maximize their utility—but we keep feeding the model more information about the person's opportunities.

Welfare Reform and Labor Supply

The theory predicts that welfare programs create work disincentives. Much of the research that examined the impact of the Aid to Families with Dependent Children program (AFDC), the key assistance program prior to the 1990s, found that the program reduced labor supply.[26]

On August 22, 1996, President Clinton signed into law the welfare reform legislation that fundamentally changed the welfare system of the United States. A key provision gave states freedom to set their own eligibility rules and benefit levels.[27] For example,

[26] Comprehensive reviews of this literature are given by Alan B. Krueger and Bruce D. Meyer, "Labor Supply Effects of Social Insurance," in Alan Auerbach and Martin Feldstein, editors, *Handbook of Public Economics,* Vol. 4, Amsterdam: North-Holland, 2002; and Robert A. Moffitt, "Welfare Programs and Labor Supply," in Alan Auerbach and Martin Feldstein, editors, *Handbook of Public Economics,* Vol. 4, Amsterdam: North-Holland, 2002.

[27] Robert A. Moffitt, "The Temporary Assistance for Needy Families Program," in Robert A. Moffitt, *Means-Tested Transfer Programs in the United States,* Chicago: University of Chicago Press, 2003, pp. 291–363.

California allowed a TANF recipient to earn up to $225 per month without affecting the size of the welfare benefit, but any additional earnings were taxed at a 50 percent rate. In contrast, Illinois taxed all labor earnings at a 33 percent rate, while Mississippi applied a 100 percent tax rate on any labor earnings above $90 per month.

Many studies exploit this interstate variation to determine the impact of welfare programs on labor supply and other variables, including the size of the welfare population itself. One problem is that the period immediately following the enactment of PRWORA coincided with a historic economic boom, making it difficult to determine how much of the subsequent decline in the size of the welfare caseload (from 4.4 million families receiving TANF in August 1996 to 2.2 million in June 2000) can be attributed to the economic boom and how much can be attributed to the change in welfare policy.[28]

Several states also conducted large-scale experiments. In the typical experiment, a group of randomly chosen families is offered a particular set of benefits, whereas other families are offered a different set. By investigating the variation in labor supply among the different groups of families, it is possible to document how labor supply responds to the changed opportunity. These experiments often confirm the theoretical predictions.[29] One well-known experiment, the Minnesota Family Investment Program, allowed women to keep some of the cash grant even if their earnings were relatively high. The results indicated that reducing the tax on labor earnings indeed encouraged the welfare recipients to work more.

There also has been interest in determining the impact of "time limits" on welfare participation. A key provision of PRWORA limits the amount of time that families can receive federal assistance to 60 months over their lifetimes, and many states have set even shorter time limits.

The time limits introduce interesting strategic choices for an eligible family: A family may choose to "bank" its benefits in order to maintain eligibility further into the future. However, federal law permits welfare payments only to families that have children younger than 18 years of age. As a result, the family's choice of whether to receive assistance today (and use up some of its 60 eligible months) or to save for a later period depends crucially on the age of the youngest child. Families with older children might as well use up their benefits now because it is unlikely that they can qualify for benefits in the future. But families with younger children have a longer time span over which they must allow for the possibility that they might need assistance, and they have an incentive not to use up the 60 months of lifetime benefits too soon. It turns out that time limits indeed discourage the welfare participation rates of families with small children.[30]

[28] Jeffrey Grogger, "The Effects of Time-Limits, the EITC, and Other Policy Changes on Welfare Use, Work, and Income among Female-Headed Families," *Review of Economics and Statistics* 85 (May 2003): 394–408.

[29] Jeffrey Grogger; Lynn A. Karoly, and Jacob Alex Klerman, *Consequences of Welfare Reform: A Research Synthesis,* Santa Monica, CA: The Rand Corporation, July 2002; and Rebecca Blank, "Evaluating Welfare Reform in the U.S.," *Journal of Economic Literature* 40 (December 2002): 1105–1166.

[30] Jeffrey Grogger, "Time Limits and Welfare Use," *Journal of Human Resources* 39 (Spring 2004): 405–424; and Jeffrey Grogger and Charles Michalopoulos, "Welfare Dynamics under Time Limits," *Journal of Political Economy* 111 (June 2003): 530–554.

2-12 Policy Application: The Earned Income Tax Credit

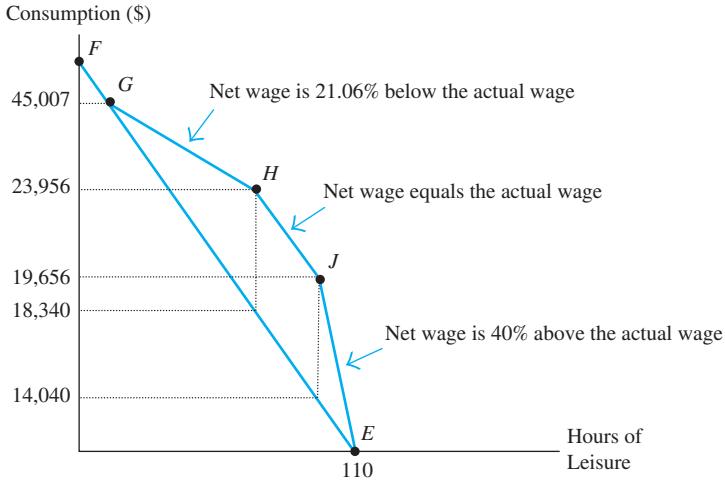
An alternative approach to improving the economic status of low-income persons is given by the Earned Income Tax Credit (EITC). This program began in 1975 and has grown substantially since. By 2015, it was the largest cash-benefit entitlement program in the United States, granting almost $70 billion to low-income households.

To illustrate how the EITC works, consider a household composed of a working mother with two qualifying children. In 2017, this woman could claim a tax credit of up to 40 percent of her earnings as long as she earned less than $14,040 per year, resulting in a maximum credit of $5,616. This maximum credit would be available as long as she earned between $14,040 and $18,340. After reaching the $18,340 threshold, however, the credit would begin to be phased out. Each additional dollar earned reduces the credit by 21.06 cents. This formula implies that the credit completely disappears once the woman earns $45,007.

Figure 2-18 illustrates how the EITC introduces a number of "kinks" into the worker's opportunity set. The figure assumes that the worker does not have any nonlabor income. In the absence of the EITC, the worker faces the straight budget line given by *FE*. The EITC changes the net wage associated with an additional hour of work. As long as the worker earns less than $14,040 per year, the worker can claim a tax credit of up to 40 percent of earnings. Suppose, for instance, that the wage rate is $10 an hour and that the worker decides to work only 1 hour during the entire year. She can then file a tax return that would grant her a $4 tax credit. Therefore, the EITC implies that the worker's net wage is $14, a 40 percent raise. This 40 percent tax credit makes the budget line steeper, as illustrated by the segment *JE* in Figure 2-16.

If the woman earns $14,040, she receives the maximum tax credit, or $5,616. And she is eligible for this maximum credit as long as she earns between $14,040 and $18,340.

FIGURE 2-18 The EITC and the Budget Line (Not Drawn to Scale)
In the absence of the tax credit, the budget line is given by *FE*. The EITC grants the worker a credit of 40 percent on labor earnings as long she earns less than $14,040. The credit is capped at $5,616. The worker receives this amount as long as she earns between $14,040 and $18,340. The tax credit is then phased out gradually. The worker's net wage is 21.06 cents below her actual wage whenever she earns between $18,340 and $45,007.



As long as the worker is in this range, therefore, the EITC does not change the net wage. It simply generates an increase in the worker's income of $5,616—as illustrated by the segment *HJ* in Figure 2-16. Put differently, the EITC generates a pure income effect in this range of the program.

Once the worker's annual earnings exceed $18,340, the EITC is phased out at a rate of 21.06 cents for every dollar earned. Suppose, for example, that the worker earns exactly $18,340 and decides to work an additional hour at $10 an hour. The tax credit is then cut back by about $2.11, implying that the worker's net wage is only $7.89 an hour. The EITC, therefore, acts like a wage cut, flattening out the budget line, as illustrated by segment *GH* in Figure 2-18. Once the worker earns $45,007 during the year, she no longer qualifies for the EITC and her budget line reverts back to the original budget line (as in segment *FG*).

This illustration of how the EITC works illustrates how government programs change the worker's opportunity set, creating strangely shaped budget "lines." These kinks can have important effects on the worker's labor supply decision.

So how does the EITC affect labor supply? The various panels of Figure 2-20 illustrate a number of possibilities. In Figure 2-19*a*, the worker would not be in the labor force in the absence of the EITC program (she maximizes her utility by being at the endowment point *P*). The 40 percent increase in the net wage associated with the EITC draws the woman into the labor force, and she maximizes her utility by moving to point *R.* The reason for the increased propensity to work should be clear. The EITC increases the net wage for nonworkers, making it more likely that the wage can exceed the reservation wage. The theory, therefore, has an important prediction: The EITC should increase the labor force participation rate in the targeted groups.

In Figure 2-19*b*, the person would work even if the EITC were not in effect (at point *P*). This worker's annual income implies that the EITC generates an income effect—without affecting the net wage. The worker maximizes her utility by moving to point *R,* and she works fewer hours.

Finally, in Figure 2-19*c*, the person would work a large number of hours in the absence of the EITC (at point *P*). The EITC cuts her net wage, and she maximizes her utility by cutting hours and moving to the kink at point *R.*

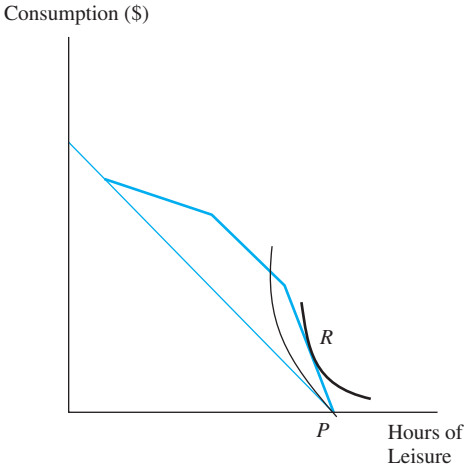
The EITC, therefore, has two distinct effects on labor supply. First, it increases the number of workers. Because the tax credit is granted only to persons who work, more persons will enter the workforce to take advantage of this program. Second, it may change the number of hours worked by persons who would have been in the labor force even in the absence of the program. As drawn in the various panels of Figure 2-19, the EITC motivated workers to work fewer hours—but the change in the net wage generates both income and substitution effects and the impact of the EITC on hours worked will depend on the relative size of these two effects.

The available evidence confirms the theoretical prediction that the EITC draws many new persons into the labor force.[31] Some of this evidence is summarized in Table 2-4. The

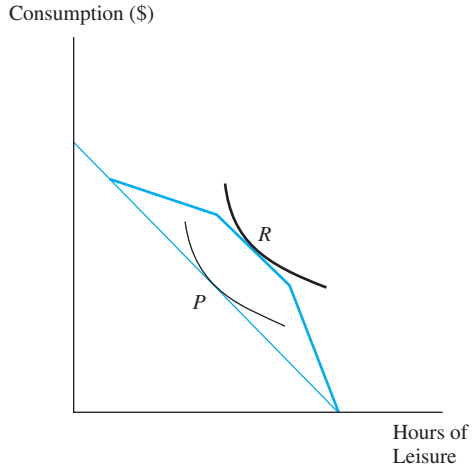
[31] V. Joseph Hotz and John Karl Scholz, "The Earned Income Tax Credit," in Robert A. Moffitt, editor, *Means-Tested Transfer Programs in the United States,* Chicago: University of Chicago Press, 2003; Alexander Gelber and Joshua W. Mitchell, "Taxes and Time Allocation: Evidence from Single Women and Men," *Review of Economic Studies* 79 (July 2012): 863–897; and Jesse Rothstein, "Is the EITC as Good as an NIT? Conditional Cash Transfers and Tax Incidence." *American Economic Journal: Economic Policy* 2 (February 2010): 177–208.

FIGURE 2-19 The Impact of the EITC on Labor Supply

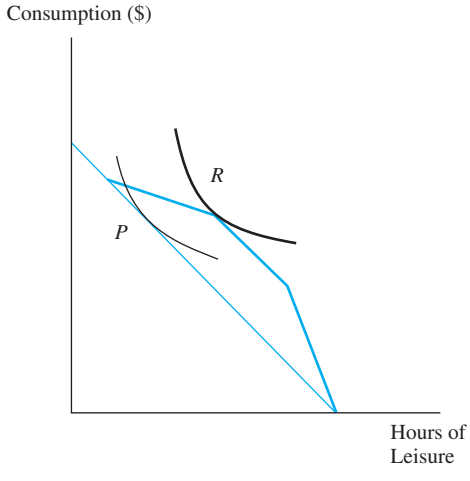
The EITC shifts the budget line, and will draw new workers into the labor market. In (*a*), the person enters the workforce by moving from point *P* to *R*. The impact of the EITC on the labor supply of persons who are already working is less clear. In the shifts illustrated in (*b*) and (*c*), the worker works fewer hours.



(*a*) EITC Draws Worker into Labor Market



(*b*) EITC Reduces Hours of Work



(*c*) EITC Reduces Hours of Work

Tax Reform Act of 1986 substantially expanded the benefits available through the EITC. The theory suggests that this legislative change should have increased the labor force participation rates of the targeted groups. Consider the population of unmarried women. Those who have at least one child potentially qualify for the EITC (depending on how

TABLE 2-4 The Impact of the Earned Income Tax Credit on Labor Force Participation

Source: Nada Eissa and Jeffrey B. Liebman, "Labor Supply Response to the Earned Income Tax Credit," *Quarterly Journal of Economics* 111 (May 1996): 617.

	Participation Rate before Legislation (%)	Participation Rate after Legislation (%)	Difference (%)	Difference-in-Differences (%)
Treatment group—eligible for the EITC:				
Unmarried women with children	72.9	75.3	2.4	
Control group—not eligible for the EITC:				2.4
Unmarried women without children	95.2	95.2	0.0	

much they earn), whereas those without children do not qualify. Table 2-4 shows that the labor force participation rate of the eligible women increased from 72.9 to 75.3 percent after the 1986 tax reform went into effect, an increase of 2.4 percentage points.

Before one can conclude that this change in labor force participation rates can be attributed to the EITC, one must consider the possibility that other factors could account for the 2.4 percentage point increase in labor force participation rates. A booming economy, for instance, could have drawn more women into the labor market even in the absence of the tax reform.

As in the typical experiment conducted in the physical sciences, therefore, we need a "control group"—a group of workers who would have experienced the same type of labor supply change but who were not "injected" with the benefits provided by the EITC. Such a group could perhaps be unmarried women without children. It turns out that their labor force participation rate did not change at all after the Tax Reform Act of 1986—it stood at 95.2 percent both before and after.

The impact of the EITC on labor force participation, therefore, can be calculated by comparing the trend in the "treated group" (the unmarried women with children) with the trend in the "control group" (the unmarried women without children). The labor force participation rate changed by 2.4 percentage points in the treated group and by 0 percentage points in the control group. One can then estimate the net impact of the EITC on labor force participation by taking a "difference-in-differences": 2.4 percentage points minus 0 percentage points, or 2.4 percentage points.

This intuitive methodology for measuring the impact of specific policy changes or economic shocks on labor market outcomes is known as the **difference-in-differences estimator** and has become very popular. It compares the observed change in a group "treated" by a particular policy shift with the observed change in a similar group that was unaffected by that shift. It is crucial to recognize that the validity of the conclusion depends on our having chosen a correct control group—a control group that is indeed similar to the treated group in all other ways so that the difference-in-differences nets out the impact of *all* other factors on the trends that we are interested in.[32]

[32] Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan, "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119 (February 2004): 249–275.

Theory at Work

GAMING THE EITC

The discussion of the Earned Income Tax Credit illustrates one obvious fact: It is very difficult for most persons to figure out how tax policies change the opportunity set. Few people would bother to sit down, work their way through the convoluted rules, and draw how the budget lines have shifted as a result of the policy. Nevertheless, such a complicated exercise is required if we want to respond correctly to the changed opportunities. It would seem, therefore, that many of us might not take full advantage of the tax credits simply because we cannot figure them out.

But knowledge about available opportunities may be "in the air" and could be easily passed on, particularly among workers who are able to manipulate their reported earnings in a way that maximizes the credit granted by the EITC. Self-employed workers obviously have many opportunities to manipulate their reported income, so that they will have an easier time making sure that their reported income "just happens" to maximize the amount of the tax benefit. In other words, a self-employed person would quickly realize that it is in their interest to have a level of disposable income that takes full advantage of the 40 percent tax credit that the EITC program grants to low earners.

An examination of detailed tax records for the entire U.S. population between 1996 and 2009 shows that such gaming of the EITC indeed takes place. Many of the self-employed "choose" to report self-employment income that happens to maximize the tax refund. In 2008, for example, 6.5 percent of EITC claimants in Chicago in 2008 were self-employed and reported earnings *exactly* at the refund-maximizing level.

The data also reveal a lot of variation in information about the EITC parameters across cities. In contrast to the Chicago experience, for example, fewer than one percent of EITC claimants in Rapid City, SD, were self-employed workers bunched at the kink. It turns out that if a self-employed worker moves from a "low-information" neighborhood (that is, a neighborhood where few self-employed workers bunch at the kink) to a "high-information" ZIP code (where many self-employed workers are bunched), the likelihood of obtaining the full EITC benefit increases substantially, suggesting that it pays to live near someone who knows the rules.

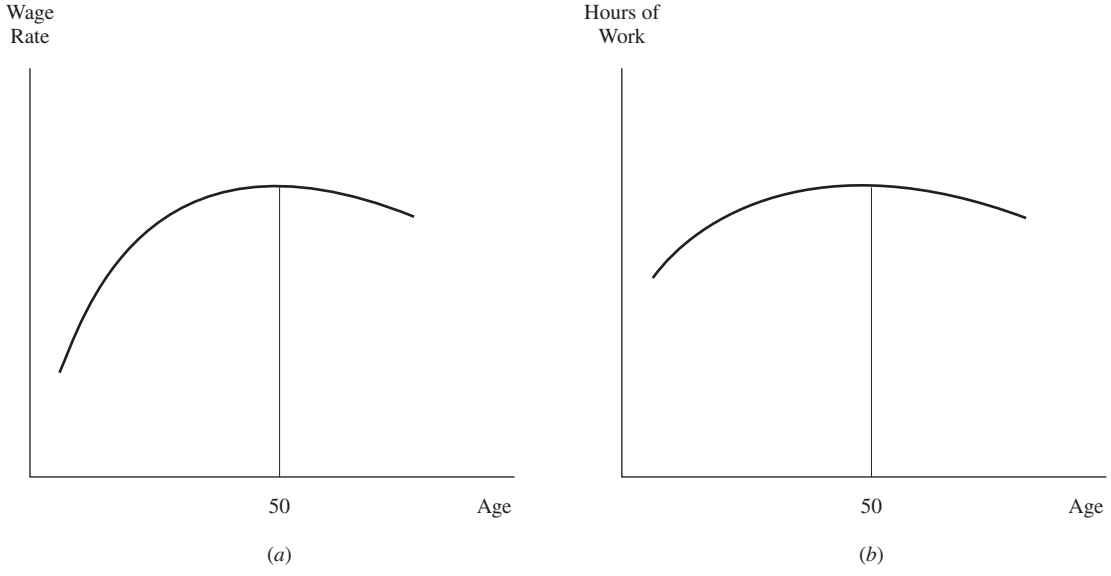
Source: Raj Chetty, John N. Friedman, and Emmanual Saez, "Using Differences in Knowledge Across Neighborhoods to Uncover the Impacts of the EITC on Earnings," *American Economic Review* 103 (December 2013): 2683–2721.

2-13 Labor Supply over the Life Cycle

Up to this point, our model of labor supply analyzes the decision of whether to work and how many hours to work from the point of view of a worker who allocates his time in a single time period and who ignores the fact that he will have to make similar decisions continuously over many years. In fact, because consumption and leisure decisions are made over the entire working life, workers can "trade" some leisure time today in return for additional consumption tomorrow. For instance, a person who devotes a great deal of time to his job today can save some of the earnings and use these savings to increase his consumption of goods in the future.

As we will see in the chapter on human capital, a great deal of evidence suggests that the typical worker's age-earnings profile—the path of a worker's wage over the life cycle—has a predictable pattern: Wages tend to be low when the worker is young; they rise as the worker ages, peaking at about age 50; and the wage rate tends to remain stable or decline slightly after age 50. This path is illustrated in Figure 2-20*a*. The typical age-earnings

FIGURE 2-20 The Life Cycle Path of Wages and Hours for a Typical Worker
(*a*) The age-earnings profile of a typical worker rises rapidly when the worker is young, reaches a peak at around age 50, and then wages either stop growing or decline slightly. (*b*) The changing price of leisure implies that the worker will devote relatively more hours to the labor market when the wage is high and fewer hours when the wage is low.



profile suggests that the price of leisure is low for younger and older workers and is highest for prime-age workers.

Consider how the worker's labor supply should respond to the wage increase that occurs between ages 20 and 30, or to the wage decline that might occur as the worker nears retirement age. It is important to note that these types of wage changes are part of the aging process *for a given worker.* A change in the wage along the worker's wage profile is called an "evolutionary" wage change, for it indicates how the wages of a particular worker evolve over time. It is crucial to note that an evolutionary wage change has no impact whatsoever on the worker's total *lifetime income.* The worker fully expects his wage to go up as he matures and to go down as he gets closer to retirement. As a result, an evolutionary wage change alters the price of leisure—but does not alter the value of the total opportunity set available to the worker over his life. To be more precise, suppose we know that our age—earnings profile takes on the precise shape illustrated in Figure 2-20*a*. The fact that our wage rises slightly from age 37 to 38 or declines slightly from age 57 to 58 does not increase or decrease our lifetime wealth. We already expected these evolutionary wage changes to occur and they have already been incorporated in the calculation of lifetime wealth.

Suppose then that the wage falls as a worker nears retirement age. Would the worker then be better off by working a lot of hours at age 50 and consuming leisure in his sixties, or would the worker be better off by working relatively few hours at age 50 and becoming a workaholic in his sixties?

The worker will clearly find it worthwhile to work more hours at age 50, invest the money, and buy goods and leisure at some point in the future when the wage is lower and leisure is not as expensive. After all, this type of labor supply decision would increase the

worker's lifetime wealth; it gives him a much larger opportunity set than would be available if he were to work many hours in his sixties (when the wage is low) and consume many hours of leisure in his fifties (when the wage is high).

A very young worker faces an analogous situation. His wage is relatively low—and he will find it optimal to consume leisure activities when he is very young, rather than in his thirties and forties, when the price of those leisure activities will be very high. The argument, therefore, suggests that we will generally find it optimal to concentrate on work activities in those years when the wage is high and to concentrate on leisure activities in those years when the wage is low.[33]

This approach to life cycle labor supply decisions implies that hours of work and the wage rate should move together over time *for a particular worker,* as illustrated in Figure 2-20*b*. This implication differs strikingly from our earlier conclusion that a wage increase generates both income and substitution effects, and that there could be a negative relationship between wages and hours of work if income effects dominate. This important difference between the models (that is, the one-period "static" model considered in the previous sections and the life cycle model presented here) arises because the two models mean very different things by a change in the wage. In the 1-period model, an increase in the wage expands the worker's opportunity set and hence creates an income effect that increases the demand for leisure. In the life cycle model, an evolutionary wage change—the wage change that workers expect as they age—does not change the total lifetime income available to a *particular* worker, and leaves the lifetime opportunity set intact.

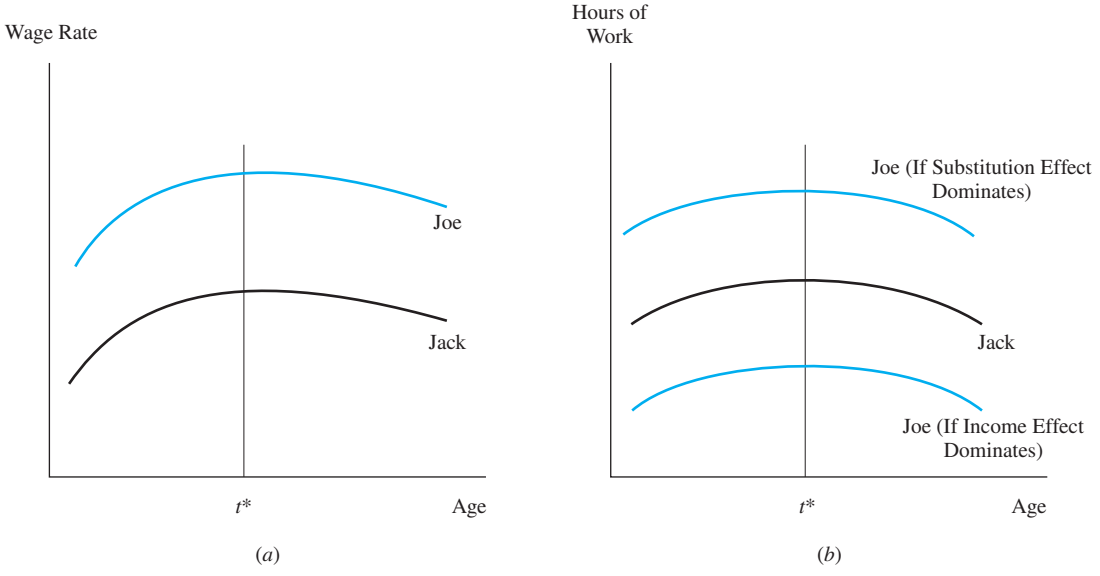
In contrast, if we were to compare two workers, say Joe and Jack, with different age-earnings profiles, the difference in hours of work between these two workers would be affected by both income and substitution effects. As illustrated in Figure 2-21*a*, Joe's wage exceeds Jack's at every age. Both Joe and Jack should work more hours when wages are high. Their life cycle profiles of hours of work are illustrated in Figure 2-21*b*. We do not know, however, which of the two workers will work more hours. Even though Joe has a higher wage and finds leisure to be very expensive, he also has a higher lifetime income and will want to consume more leisure. The difference in the level of the two wage profiles, therefore, generates an income effect. If these income effects are sufficiently strong, Joe's hours-of-work profile will lie below Jack's; if substitution effects dominate, Joe will always work more hours than Jack.

The life cycle approach suggests a link not only between wages and hours of work, but also between wages and labor force participation rates. As we saw earlier in the chapter, the labor force participation decision depends on a comparison of the reservation wage to the market wage. In each year of the life cycle, therefore, the worker will compare the reservation wage to the market wage. Suppose that the reservation wage is roughly constant over time. The person is then more likely to enter the labor market in periods when the wage is high. As a result, participation rates are likely to be low for young workers, high for workers in their prime working years, and low again for older workers.

The participation decision, however, also depends on how reservation wages vary over the life cycle. The reservation wage measures the bribe required to enter the labor market.

[33] James J. Heckman, "Life Cycle Consumption and Labor Supply: An Explanation of the Relationship between Income and Consumption over the Life Cycle," *American Economic Review* 64 (March 1974): 188–194.

FIGURE 2-21 Hours of Work over the Life Cycle for Two Workers with Different Wage Paths
Joe's wage exceeds Jack's at every age. Although both Joe and Jack work more hours when the wage is high, Joe works more hours than Jack only if the substitution effect dominates. If the income effect dominates, Joe works fewer hours than Jack.



For instance, the presence of small children in the household might increase the value of time in the household sector for the person most responsible for child care and, hence, would increase the reservation wage. Therefore, it is not surprising to find that some married women participate in the labor force intermittently. They work prior to the arrival of the first child, withdraw from the labor market when the children are small and need full-time care, and return to the labor market once the children enroll in school.

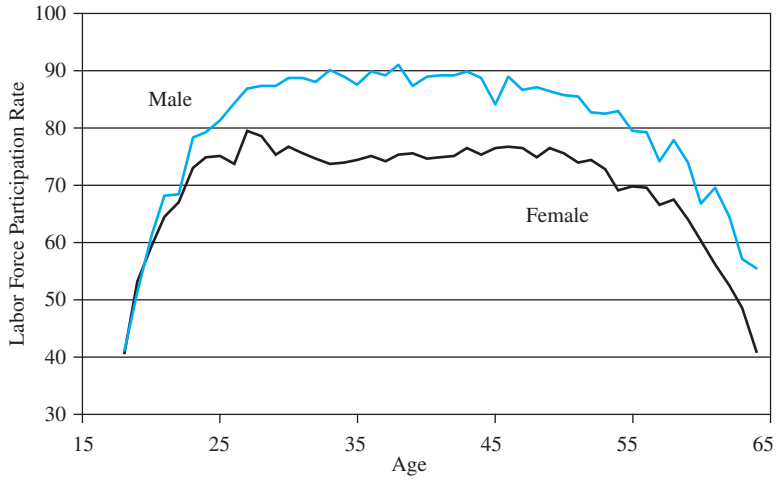
The key implication of the analysis can be easily summarized. The evidence on age-earnings profiles suggests that the wage is relatively low for young workers, increases as the worker matures and accumulates various types of skills, and then may decline for older workers. The model then suggests that the profile of hours of work over the life cycle will have exactly the same shape as the age-earnings profile: Hours of work increase as the wage rises and decline as the wage falls. The theoretical prediction that people allocate their time over the life cycle so as to take advantage of changes in the price of leisure is called the **intertemporal substitution hypothesis.**

Evidence

The available evidence suggests that both labor force participation rates and hours of work respond to evolutionary wage changes. Figure 2-22 illustrates the relationship between labor force participation rates and age in the United States. Male participation rates peak when men are between 25 and 45 years old and begin to decline noticeably after age 45. In contrast, female participation rates, probably because of the impact of child-raising activities on the participation decision, decline slightly for women in their late 20s and early 30s.

FIGURE 2-22 Labor Force Participation Rates over the Life Cycle, 2017

Source: U.S. Bureau of Labor Statistics, *Current Population Surveys,* Annual Social and Economic Supplement, 2017.



Overall, the trends illustrated in the figure are consistent with the theoretical prediction that participation rates should be highest when the wage is high (that is, when workers are in their thirties and forties). The decline in labor force participation rates observed after age 55, however, is much too steep to be explained by the wage decline that is typically observed as workers near retirement age. The rapid decline in participation rates at older ages may be health related and, as we will see later in this chapter, also may be attributable to the work disincentive effects of various retirement and disability insurance programs.

Figure 2-23 illustrates the actual relationship between hours of work and age. As with participation rates, hours of work among working men rise rapidly until about age 30, peak at ages 35–45, and begin to decline at age 50. During the prime working years, men work about 2,100 hours annually. In contrast, hours of work among working women do not peak until age 50 (probably because some younger women work in part-time jobs while they have small children in the household).

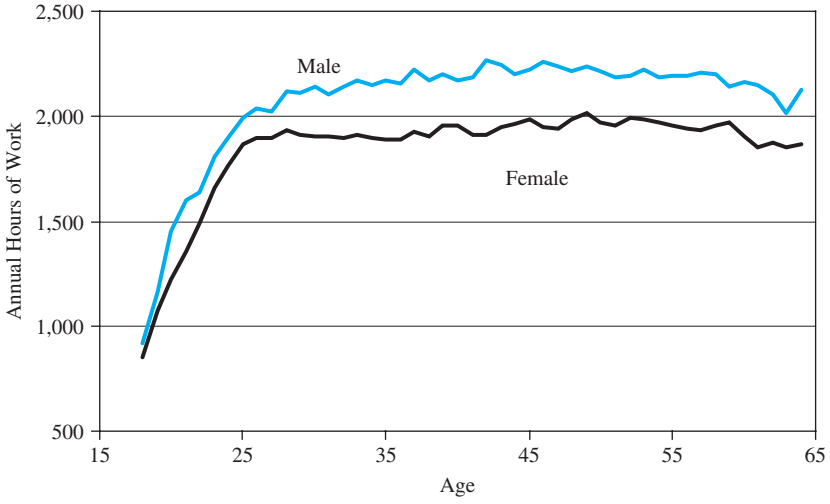
Estimation of Life Cycle Models

The estimation of the intertemporal labor supply elasticity—the crucial parameter that determines how hours of work evolve in the life cycle model of the labor–leisure choice—helped introduce what has become a very useful econometric technique into the labor economics literature.[34] The economic model states that we should be tracking a specific individual over the lifetime so that we can observe how *his* hours of work change from year to year as a response to year-to-year wage changes. Suppose that we have a longitudinal

[34] Thomas E. MaCurdy, "An Empirical Model of Labor Supply in a Life-Cycle Setting," *Journal of Political Economy* 89 (December 1981): 1059–1085. See also Joseph G. Altonji, "Intertemporal Substitution in Labor Supply: Evidence from Micro Data," *Journal of Political Economy* 94 (June 1986, Part 2): S176–S215; Joseph V. Hotz, Kinn Kydland, and Guilherme Sedlacek, "Intertemporal Preferences and Labor Supply," *Econometrica* 56 (March 1988): 335–360; and Casey Mulligan, "Substitution over Time: Another Look at Life Cycle Labor Supply," *NBER Macroeconomics Annual* 13 (1998): 75–134.

FIGURE 2-23 Hours of Work over the Life Cycle, 2017

Source: U.S. Bureau of Labor Statistics, *Current Population Surveys,* Annual Social and Economic Supplement, 2017.



data set that allows us to observe a particular worker i twice, say, at the ages of 40 and 41. Let H_{it} give his hours of work at age t, and w_{it} gives his wage rate at that age. It is easy to see that one can difference the data for each individual and estimate the following regression model across the sample of different workers:

$$\Delta H_{it} = \sigma \Delta w_{it} + \text{other variables} \qquad \textbf{(2-14)}$$

where ΔH_{it} gives the year-to-year change in hours of work and Δw_{it} gives the year-to-year change in the worker's wage. The coefficient σ would be related to the intertemporal labor supply elasticity because it measures the change in hours of work for a given person resulting from a particular change in his wage rate.

The statistically interesting part of the problem arises when one observes the same person for more than two periods. Suppose, for example, that we have a sample containing 1,000 workers and that each worker in our data is observed for 20 years. Although one could imagine differencing the data a number of times, there exists a statistically simpler procedure that effectively does the same thing. In particular, we would stack all 20 observations for a particular worker across all workers. The new regression model, therefore, would have 20,000 observations. We would then estimate the following regression model on this stacked data set:

$$H_{it} = \sigma w_{it} + \alpha_1 F_1 + \alpha_2 F_2 + \cdots + \alpha_{1,000} F_{1,000} + \text{other variables} \qquad \textbf{(2-15)}$$

where F_1 is a "dummy variable" set equal to one if that observation refers to person 1, and zero otherwise; F_2 is another dummy variable set equal to one if that observation refers to person 2, and zero otherwise; and so on. In effect, the regression model in equation (2-15) includes a dummy variable for each person in the data, and there would be 1,000 such dummy variables.

The set of dummy variables ($F_1, \ldots, F_{1,000}$) are called **fixed effects,** because they indicate that hours of work for worker *i,* for whatever reasons, have a fixed factor that determines the person's hours of work on a permanent basis, even apart from year-to-year wage fluctuations. The set of individual-specific fixed effects included in the regression model in equation (2-15) controls for any factors that are specific to persons and lets us measure how wage changes for a particular person affect that person's hours of work. In fact, if each worker in our data were only observed twice, the method of including fixed effects in the regression model yields exactly the same numerical results as the common-sense differencing illustrated in equation (2-14).[35]

The elasticities of intertemporal substitution estimated by the method of fixed effects tend to be positive, but numerically small: A 10 percent evolutionary increase in the wage leads to less than a 1 percent increase in hours of work. This is not too surprising because as Figure 2-21 clearly shows, although hours of work do increase early on in the life cycle and decline as retirement age approaches, hours of work tend to be "sticky" over a long stretch of the working life.[36]

The statistical method of fixed effects has become a commonly used empirical technique in the toolkit of modern labor economics. It is easy to see why: There are obviously many person-specific factors that affect how many hours we work. Some of us are workaholics, and some of us would rather binge-watch the latest hot show on television. Our tastes for work are, to a large extent, fixed; they are a part of who we are. The individual-specific fixed effects help control for these idiosyncratic differences among workers and allow us to focus on what is most important in terms of the economic models: How changes in economic opportunities for a given worker affect that worker's labor supply.

Labor Supply over the Business Cycle

Not only does labor supply respond to changes in economic opportunities over a worker's life cycle, but the worker also may adjust his labor supply to take advantage of changes in opportunities induced by business cycles. Do recessions motivate many persons to enter the labor market in order to "make up" the income of family members who have lost their jobs? Or do the unemployed give up hope of finding work in a depressed market and leave the labor force altogether?

The **added worker effect** provides one possible mechanism. Under this hypothesis, so-called secondary workers who are currently out of the labor market (such as mothers with small children) are affected by the recession because the main breadwinner becomes unemployed or faces a wage cut. As a result, family income falls and the secondary workers get jobs to make up the loss. The added worker effect implies that the labor force participation rate of secondary workers has a countercyclical trend (that is, it moves in a direction opposite to the business cycle): It rises during recessions and falls during expansions.

[35] More precisely, the equivalence result requires that the "other variables" included in the regression model in equation (2-15) also be differenced.

[36] There is a debate over this conclusion. The magnitude of the labor supply response to life cycle changes in the wage has important implications in macroeconomics. Some macroeconomic models require sizable intertemporal elasticities to explain the behavior of employment over the business cycle.

There is also a **discouraged worker effect.** The discouraged worker effect argues that many unemployed workers find it almost impossible to find jobs during a recession and simply give up. Rather than continue a fruitless search for a job, these workers decide to wait out the recession and drop out of the labor force. The labor force participation rate would then have a procyclical trend: It falls during recessions and increases during expansions.

The business cycle will typically create both added workers and discouraged workers. But which effect dominates? This question is typically addressed by correlating the labor force participation rate of a particular group with the aggregate unemployment rate, a summary measure of economic activity. If the added worker effect dominates, the correlation should be positive because worsening economic conditions encourages more persons to enter the labor market. If the discouraged worker effect dominates, the correlation should be negative because the high level of unemployment convinces many workers to give up and drop out of the labor market. There is evidence that the correlation between the labor force participation rate of many groups and the aggregate unemployment rate is negative, so the discouraged worker effect dominates.[37]

Because the discouraged worker effect dominates, the official unemployment rate might be too low. Recall that the BLS defines the unemployment rate as the ratio of persons who are unemployed to persons who are in the labor force (that is, the employed plus the unemployed). If an unemployed person becomes discouraged and leaves the labor force, he or she is no longer counted among the unemployed. As a result, the official unemployment rate may understate the unemployment problem during severe recessions.

However, the argument that the discouraged workers should be included in the unemployment statistics has been questioned.[38] Some of the discouraged workers may be "taking advantage" of the relatively poor labor market conditions during a recession to engage in leisure activities. The life cycle model of labor supply suggests that workers should allocate more of their time to the labor market in those years when the wage is higher. The real wage typically rises during expansions (when the demand for labor is high) and declines during recessions (when the demand for labor slackens). We would then expect the labor force participation rate to be high at the peak of economic activity and to decline as economic conditions worsen. The procyclical trend in the labor force participation rate then arises not because workers give up hope of finding jobs during recessions but because it is not worthwhile to work in those years when the real wage is low. In an important sense, the so-called discouraged workers are doing precisely what the life cycle model of labor supply suggests they should do: Allocate their time optimally by consuming leisure when it is cheap to do so. As a result, the pool of hidden unemployed perhaps should not be part of the unemployment statistics.

[37] Jacob Mincer, "Labor Force Participation and Unemployment: A Review of Recent Evidence," in R. A. Gordon and M. S. Gordon, editors, *Prosperity and Unemployment,* New York: Wiley, 1966, pp. 73–112; and Shelly Lundberg, "The Added Worker Effect," *Journal of Labor Economics* 3 (January 1985): 11–37.

[38] This argument is developed at length in the influential article by Robert E. Lucas and Leonard Rapping, "Real Wages, Employment, and Inflation," *Journal of Political Economy* 77 (October 1969): 721–754.

2-14 Policy Application: Disability Benefits and Labor Force Participation

There has been a marked drop in labor force participation among older men. It has been argued that the availability of generous benefits from the Social Security Disability Program can explain part of the trend. Workers who become disabled are eligible to receive disability payments for as long as the disability lasts. The monthly disability benefit equals the Social Security retirement benefits that the worker would have received had he or she continued working until age 65, *regardless of the worker's age at the time the disability occurred.*

Some persons would obviously like to claim that they are disabled in order to enjoy the leisure activities associated with early retirement. As a result, the eligibility requirements for the disability program are harsh and strictly enforced. Workers applying for disability benefits must often be certified as being disabled by government-picked physicians; there is a waiting period of five months before the worker can apply for disability benefits; and the worker cannot be employed in "gainful activities" (where the worker earns more than $1,180 per month in 2018).

A correlation between a person's employment status and whether that person is enrolled in the disability program would not provide any information about whether the existence of the program creates work disincentives. After all, there may be an underlying health impairment that makes the person both more likely to qualify for benefits and reduces the person's ability to enter the workforce.

To get around this problem, researchers have constructed creative ways of documenting whether more generous benefits have a direct impact on a person's labor force participation. The goal is to answer a simple question: Do the benefits encourage some persons who otherwise would have worked to try to enroll in the program and drop out of the labor force?

One early study, for example, tracked the labor supply decisions of the applicants whose disability application was rejected by the government.[39] If the rejected claims were mainly attempts to misuse the program, one might expect that the rejected applicants would return to the labor force once they learned that they cannot "get away" with this early retirement strategy. Around 40 percent of the rejected applicants indeed go back to work after the final determination of their case.

Some of the most convincing evidence is provided by a study of the Canadian experience.[40] The disability program is a federal program in the United States, which implies that eligibility and benefit levels are the same throughout the country. In Canada, however, there are two programs: The Quebec Pension Program (QPP), which covers persons residing in Quebec, and the Canada Pension Program (CPP), which covers persons residing in the rest of Canada. Although these two systems are similar in many ways, the QPP was substantially more generous than the CPP before 1987. In January 1987, however, the CPP raised its benefit levels to bring the two programs to parity.

[39] John Bound, "The Health and Earnings of Rejected Disability Insurance Applicants," *American-Economic Review* 79 (June 1989): 482–503.

[40] Jonathan Gruber, "Disability Insurance Benefits and Labor Supply," *Journal of Political Economy* 108 (December 2000): 1162–1183.

TABLE 2-5 The Impact of Disability Benefits on Labor Supply in Canada

Source: Jonathan Gruber, "Disability Insurance Benefits and Labor Supply," *Journal of Political Economy* 108 (December 2000): 1175.

	Before	After	Difference	Difference-in-Differences
Annual benefits:				
Canada Pension Program	$5,134	$7,776	$2,642	$1,666
Quebec Pension Program	6,876	7,852	976	
Percent of men aged 45 to 59 not employed last week:				
Treatment group: CPP	20.0%	21.7%	1.7%	2.7%
Control group: QPP	25.6	24.6	−1.0	

Table 2-5 provides a difference-in-differences analysis of the impact of this change on the labor supply of the affected population. The top rows of the table show that benefit levels in the rest of Canada increased by $2,642 (Canadian dollars) between 1986 and 1987, as compared to an increase of only $976 in Quebec.

The bottom rows of the table document how the increased generosity affected labor supply. The fraction of men aged 45 to 59 who did not work fell from 25.6 to 24.6 in Quebec (a decrease of 1.0 percentage point), likely reflecting changes in aggregate economic activity over the period. In contrast, the proportion of men residing outside Quebec who did not work *rose* from 20.0 to 21.7 percent, an increase of 1.7 percentage points. The difference-in-differences estimator (or $1.7 - (-1.0)$) implies that the increased generosity of the disability program increased the proportion of men who did not work by 2.7 percentage points.

Recent studies of the American experience provide equally strong evidence of a link between disability benefits and labor force participation.[41] In the United States, a person's application for disability status is reviewed by a randomly chosen disability examiner, a professional who looks at the medical record and determines whether a person is truly disabled and unable "to engage in substantial gainful activity." Inevitably, it's easier to "pass the test" with some examiners than with others. It turns out that even 4 years after the application, and after controlling for observable health impairments, the employment rate of those applicants who happened to land a tougher examiner (making it more likely that their application was rejected) was far higher than the employment rate of the applicants who landed an easier examiner.

[41] Nicole Maestas, Kathleen J. Mullen and Alexander Strand, "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt," *American Economic Review* 103 (August 2013): 1797–1829. See also David H. Autor, Mark Duggan, Kyle Greenberg, and David S. Lyle, "The Impact of Disability Benefits on Labor Supply: Evidence from the VA's Disability Compensation Program," *American Economic Journal: Applied Economics* 8 (July 2016): 31–68.

Summary

- The reservation wage is the wage that makes a person indifferent between working and not working. A person enters the labor market when the market wage rate exceeds the reservation wage.
- Utility-maximizing workers allocate their time so that the last dollar spent on leisure activities yields the same utility as the last dollar spent on goods.
- An increase in nonlabor income reduces hours of work of workers.
- An increase in the wage generates both an income and a substitution effect among persons who work. The income effect reduces hours of work; the substitution effect increases hours of work. The labor supply curve, therefore, is upward sloping if substitution effects dominate and downward sloping if income effects dominate.
- An increase in nonlabor income reduces the likelihood that a person enters the labor force. An increase in the wage increases the likelihood that a person enters the labor force.
- The labor supply elasticity is on the order of −0.1 for men and +0.2 for women.
- Welfare programs create work disincentives because they provide cash grants to participants as well as tax those recipients who enter the labor market. In contrast, credits on earned income create work incentives and draw many nonworkers into the labor force.

Key Concepts

added worker effect, *67*
budget constraint, *29*
budget line, *29*
discouraged worker effect, *68*
difference-in-differences estimator, *60*
employment rate, *20*
fixed effects, *67*
hidden unemployed, *21*
household production function, *45*
income effect, *32*
indifference curve, *25*
intertemporal substitution hypothesis, *64*
labor force, *20*
labor force participation rate, *20*
labor supply curve, *39*
labor supply elasticity, *40*
marginal rate of substitution (MRS) in consumption, *27*
marginal utility, *26*
neoclassical model of labor–leisure choice, *23*
opportunity set, *29*
reservation wage, *37*
substitution effect, *36*
unemployment rate, *20*
utility function, *24*

Review Questions

1. What happens to the reservation wage if nonlabor income increases, and why?
2. What economic factors determine whether a person participates in the labor force?
3. How does a typical worker decide how many hours to allocate to the labor market?
4. What happens to hours of work when nonlabor income decreases?
5. What happens to hours of work when the wage rate falls? Decompose the change in hours of work into income and substitution effects.
6. What happens to the probability that a particular person works when the wage rises? Does such a wage increase generate an income effect?

7. Does a correlation between hours of work and nonlabor income measure the income effect?
8. Why do welfare programs create work disincentives?
9. Why does the earned income tax credit increase the labor force participation rate of targeted groups?
10. Why did the labor force participation rate of women increase so much in the past century?
11. Why does a worker allocate his or her time over the life cycle so as to work more hours in those periods when the wage is highest? Why does the worker not experience an income effect during those periods?
12. What is the added worker effect? What is the discouraged worker effect?

Problems

2-1. The table below reports the unemployment rate, labor force participation rate, and (working-age) population for the United States in January 2008, 2011, and 2016. Using the data, answer the following questions.

a. What was the size of the labor force at the start of each year?

b. How many people were officially unemployed at the start of each year?

c. What about these numbers may cause some concern even though the unemployment rate to start 2016 was a notch below the unemployment rate in 2008 as the economy was entering the Great Recession?

	2008	2011	2016
Unemployment Rate	5.0%	9.1%	4.9%
Labor Force Participation Rate	66.2%	64.2%	62.7%
Working-age Population	234m	238m	251m

2-2. Charlie and Larry both face the same budget line for consumption and leisure. At every possible consumption-leisure bundle on the budget line, Charlie always requires marginally more leisure than does Larry in order to be equally happy when asked to forego a dollar of consumption. Using a standard budget line, graph several indifference curves and the optimal consumption-leisure bundle for both people. Which person optimally chooses more consumption? Which feature of indifference curves guarantees this result?

2-3. Tom earns $15 per hour for up to 40 hours of work each week and $30 per hour for every hour in excess of 40. Tom also faces a 20 percent tax rate, pays $4 per hour in child care expenses for each hour he works, and receives $80 in child support payments each week. There are 110 (non-sleeping) hours in the week. Graph Tom's weekly budget line.

2-4. Cindy gains utility from consumption C and leisure L. The most leisure she can consume in any given week is 110 hours. Her utility function is $U(C, L) = C \times L$. This functional form implies that Cindy's marginal rate of substitution is C/L. Cindy receives $660 each week from her great-grandmother–regardless of how much Cindy works. What is Cindy's reservation wage?

2-5. Currently a firm pays 10% of each employee's salary into a retirement account, regardless of whether the employee also contributes to the account. The firm is considering changing this system to a 10% match meaning that the firm will match the employee's contribution into the account up to 10% of each employee's salary. Some people at the firm think this change will lead employees to save more and therefore be more able to afford to retire at a younger age, while others believe this change will lead employees to have less retirement savings and therefore be less able to afford to retire. Explain why either point of view could be correct.

2-6. Shelly's preferences for consumption and leisure can be expressed as

$$U(C, L) = (C - 100) \times (L - 40).$$

This utility function implies that Shelly's marginal utility of leisure is $C - 100$ and her marginal utility of consumption is $L - 40$. There are 110 hours in the week available to split between work and leisure. Shelly earns $10 per hour after taxes. She also receives $320 worth of assistance benefits each week regardless of how much she works.

(a) Graph Shelly's budget line.

(b) What is Shelly's marginal rate of substitution when $L = 100$ and she is on her budget line?

(c) What is Shelly's reservation wage?

(d) Find Shelly's optimal amount of consumption and leisure.

2-7. Explain why receiving a cash grant from the government can entice some workers to stop working (and entices no one to start working) while the earned income tax credit can entice some people who otherwise would not work to start working (and entices no one to stop working).

2-8. In 1999, 4,860 TANF recipients were asked how many hours they worked in the previous week. In 2000, 4,392 of these recipients were again subject to the same TANF rules and were again asked their hours of work during the previous week. The remaining 468 individuals were randomly assigned to a "Negative Income Tax" (NIT) experiment which gave out financial incentives for welfare recipients to work and were subject to its rules. Like the other group, they were asked about their hours of work during the previous week. The data from the experiment are contained in the table below.

	Number Of Recipients	Number of Recipients Who Worked At Some Time in the Survey Week		Total Hours of Work By All Recipients in the Survey Week	
		1999	2000	1999	2000
TANF	4,392	1,217	1,568	15,578	20,698
NIT	468	131	213	1,638	2,535
Total	4,860	1,348	1,781	17,216	23,233

(a) What effect did the NIT experiment have on the employment rate of public assistance recipients? Develop a standard difference-in-differences table to support your answer.

(b) What effect did the NIT experiment have on the weekly hours worked of public assistance recipients who worked positive hours during the survey week? Develop a standard difference-in-differences table to support your answer.

2-9. Consider two workers with identical preferences, Phil and Bill. Both workers have the same life cycle wage path in that they face the same wage at every age, and they know what their future wages will be. Leisure and consumption are both normal goods.

(a) Compare the life cycle path of hours of work between the two workers if Bill receives a one-time, unexpected inheritance at the age of 35.

(b) Compare the life cycle path of hours of work between the two workers if Bill had always known he would receive (and, in fact, does receive) a one-time inheritance at the age of 35.

2-10. Under current law, most Social Security recipients do not pay federal or state income taxes on their Social Security benefits. Suppose the government proposes to tax these benefits at the same rate as other types of income. What is the impact of the proposed tax on the optimal retirement age?

2-11. A worker plans to retire at the age of 65, at which time he will start collecting his retirement benefits. Then there is a sudden change in the forecast of inflation when the worker is 63 years old. In particular, inflation is now predicted to be higher than it had been expected so that the average price level of market goods and wages is now expected to be higher. What effect does this announcement have on the person's preferred retirement age:

(a) if retirement benefits are fully adjusted for inflation?

(b) if retirement benefits are not fully adjusted for inflation?

2-12. Presently, there is a minimum and maximum social security benefit paid to retirees. Between these two bounds, a retiree's benefit level depends on how much she contributed to the system over her work life. Suppose Social Security was changed so that everyone aged 65 or older was paid $12,000 per year regardless of how much she earned over her working life or whether she continued to work after the age of 65. How would this likely affect the number of hours worked by retirees?

2-13. Over the last 100 years, real household income and standards of living have increased substantially in the United States. At the same time, the total fertility rate, the average number of children born to a woman during her lifetime, has fallen in the United States from about three children per woman in the early twentieth century to about two children per woman in the early twenty-first century. Does this suggest that children are inferior goods?

2-14. Consider a person who can work up to 80 hours each week at a pre-tax wage of $20 per hour but faces a constant 20% payroll tax. Under these conditions, the worker maximizes her utility by choosing to work 50 hours each week. The government proposes a negative income tax whereby everyone is given $300 each week and anyone can supplement her income further by working. To pay for the negative income tax, the payroll tax rate will be increased to 50%.

(a) On a single graph, draw the worker's original budget line and her budget line under the negative income tax.

(b) Show that the worker will choose to work fewer hours if the negative income tax is adopted.

(c) Will the worker's utility be greater under the negative income tax?

2-15. The absolute value of the slope of the consumption-leisure budget line is the after-tax wage, w. Suppose some workers earn w for up to 40 hours of work each week, and then earn $2w$ for any hours worked thereafter (called overtime). Other workers earn w for up to 40 hours of work each week, and then only earn $0.5w$ thereafter as working more than 40 hours requires getting a second job which pays an hourly wage less than their primary job. Both types of workers experience a "kink" in their consumption-leisure budget line.

(a) Graph in general terms the budget line for each type of worker.

(b) Which type of worker is likely to work up to the point of the kink, and which type of worker is likely to choose a consumption-leisure bundle far away from the kink?

Selected Readings

Gary S. Becker, "A Theory of the Allocation of Time," *Economic Journal* 75 (September 1965): 493–517.

Raj Chetty, John N. Friedman, and Emmanual Saez, "Using Differences in Knowledge Across Neighborhoods to Uncover the Impacts of the EITC on Earnings," *American Economic Review* 103 (December 2013): 2683–2721.

Nada Eissa and Jeffrey B. Liebman, "Labor Supply Response to the Earned Income Tax Credit," *Quarterly Journal of Economics* 111 (May 1996): 605–637.

Jeffrey Grogger and Charles Michalopoulos, "Welfare Dynamics under Time Limits," *Journal of Political Economy* 111 (June 2003): 530–554.

James J. Heckman, "Life Cycle Consumption and Labor Supply: An Explanation of the Relationship between Income and Consumption over the Life Cycle," *American $$$-Economic Review* 64 (March 1974): 188–194.

James J. Heckman, "Sample Selection Bias as a Specification Error with an Application to the Estimation of Labor Supply Functions," in James P. Smith, editor, *Female Labor Supply: Theory and Estimation.* Princeton, NJ: Princeton University Press, 1980, pp. 206–248.

Nicole Maestas, Kathleen J. Mullen, and Alexander Strand, "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt," *American Economic Review* 103 (August 2013): 1797–1829.

Thomas E. MaCurdy, "An Empirical Model of Labor Supply in a Life-Cycle Setting," *Journal of Political Economy* 89 (December 1981): 1059–1085.

Jacob Mincer, "Labor Force Participation of Married Women," in H. Gregg Lewis, editor, *Aspects of Labor Economics.* Princeton, NJ: Princeton University Press, 1962, pp. 63–97.

Robert A. Moffitt, "Welfare Programs and Labor Supply," in Alan J. Auerbach and Martin Feldstein, editors, *Handbook of Public Economics,* vol. 4. Amsterdam: Elsevier, 2006.

Chapter 3

Labor Demand

The laborer is worthy of his hire.

—The Gospel of St. Luke

The last chapter analyzed the factors that determine how many persons enter the labor market and how many hours those workers are willing to supply to employers. Labor market outcomes, however, depend not only on our willingness to work, but also on the firms' willingness to hire us.

The hiring and firing decisions made by firms simultaneously create and destroy many jobs. In 2008, for example, just prior to the Great Recession, nearly 6.6 percent of jobs in the United States were newly created and 6.4 percent of existing jobs vanished.[1]

Our analysis of labor demand begins by recognizing that firms do not hire workers simply because employers want to see "bodies" filling in various positions in the firm. Rather, firms hire workers because consumers want to purchase goods and services. Firms are the middlemen that hire workers to produce those goods and services. The firm's labor demand—just like the firm's demand for other inputs in the production process such as land, buildings, and machines—is a "derived demand," derived from the wants and desires of consumers.

Despite the apparent similarity between the determinants of the firm's demand for labor and the firm's demand for other inputs, economists devote a lot of effort to the separate study of labor demand. After all, workers *do* differ from other inputs in a number of important ways. All of us are keenly interested in the behavior of the firms that rent our services for 8 hours a day. Some firms provide working conditions and social opportunities that are quite amenable, whereas working conditions in other firms may be appalling. The determinants of the demand for labor also have important social and political implications. In fact, many of the central questions in economic policy involve the number of workers that firms employ and the wage that they offer those workers. Such diverse policies as minimum wages and restrictions on an employer's ability to fire or lay off workers are attempts to regulate various aspects of the firm's labor demand.

[1] Steven J. Davis, R. Jason Faberman, and John Haltiwanger, "Labor Market Flows in the Cross Section and Over Time," *Journal of Monetary Economics* 59 (January 2012): 1–18.

3-1 The Production Function

The firm's **production function** describes the technology that the firm uses to produce goods. We initially assume that there are only two factors of production: The number of employee-hours hired by the firm (E) and capital (K), which includes land, machines, and other physical infrastructure. We write the production function as

$$q = f(E, K) \qquad \textbf{(3-1)}$$

where q is the firm's output. The production function tells us how much output is produced by any combination of labor and capital.

The definition of the labor input makes two restrictive assumptions. First, the number of employee-hours E is given by the product of the number of workers times the average number of hours worked per person. By focusing on the product E, rather than on its two separate components, we are assuming that the firm gets the same output when it hires 10 workers for an 8-hours day as when it hires 20 workers for a 4-hours shift. To simplify the presentation, we will often ignore the employee-hours distinction, and refer to E as the number of workers hired.

Second, the production function overlooks that there are different types of workers, and that these different types will make different contributions to production. Some workers are college graduates, others are high school dropouts; some have a lot of work experience, others have little.

It is useful, however, to start by ignoring these complications. The simpler model provides a solid understanding of how firms make hiring decisions. We can then build upon this foundation to allow for a more general specification of the production technology.

Marginal Product and Average Product

The key concept associated with the firm's production function is that of marginal product. The **marginal product of labor** (which we denote by MP_E) gives the change in output resulting from hiring an additional worker, holding constant the quantities of all other inputs. Similarly, the **marginal product of capital** (or MP_K) gives the change in output resulting from a one-unit increase in the capital stock, holding constant the quantities of all other inputs. We assume that the marginal products of both labor and capital are positive numbers: Hiring either more workers or more capital leads to more output.

It is easy to understand how we calculate the marginal product of labor by using a numerical example. Table 3-1 summarizes the firm's production when it hires different numbers of workers, *holding capital constant.* If the firm hires one worker, it produces 11 units of output. The marginal product of the first worker hired, therefore, is 11. If the firm hires two workers, production rises to 27 units of output, and the marginal product of the second worker is 16.

Figure 3-1 graphs the data in our example to illustrate the assumptions that are typically made about the production function. Figure 3-1*a* shows the total product curve. This curve describes what happens to output as the firm hires more workers. The total product curve is obviously upward sloping.

The marginal product of labor curve, illustrated in Figure 3-1*b,* gives the slope of the total product curve—that is, the rate of change in output as more workers are hired. In our

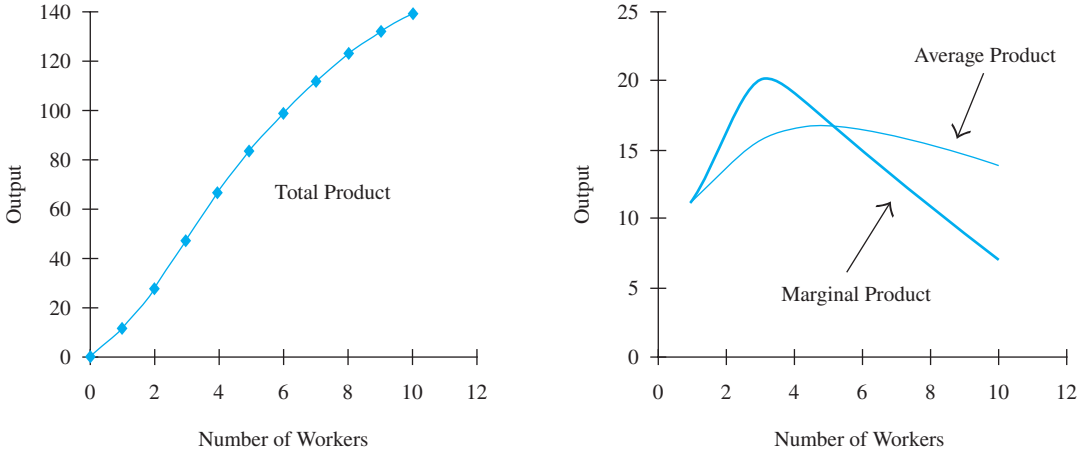
TABLE 3-1
Calculating the Marginal and Average Product of Labor (Holding Capital Constant)

Note: The calculations for the value of marginal product and the value of average product assume that the price of the output is $2.

Number of Workers Employed	Output (Units)	Marginal Product (Units)	Average Product (Units)	Value of Marginal Product ($)	Value of Average Product ($)
0	0	—	—	—	—
1	11	11	11.0	22	22.0
2	27	16	13.5	32	27.0
3	47	20	15.7	40	31.3
4	66	19	16.5	38	33.0
5	83	17	16.6	34	33.2
6	98	15	16.3	30	32.7
7	111	13	15.9	26	31.7
8	122	11	15.3	22	30.5
9	131	9	14.6	18	29.1
10	138	7	13.8	14	27.6

numerical example, output first rises at an increasing rate as more workers are hired. This implies that the marginal product of labor is rising, perhaps because of the initial gains resulting from assigning workers to specific tasks. Eventually, output increases at a decreasing rate. In other words, the marginal product of labor begins to decline, so that the next worker hired adds less to the firm's output than a previously hired worker. In our

FIGURE 3-1 **The Total Product, Marginal Product, and Average Product Curves**
(*a*) The total product curve gives the relationship between output and the number of workers hired by the firm, holding capital fixed. (*b*) The marginal product curve gives the output produced by each additional worker and the average product curve gives the output per worker.



(*a*) (*b*)

example, the marginal product of the third worker hired is 20, but the marginal product of the fourth worker is 19, and that of the fifth worker declines further to 17.

The assumption that the marginal product of labor eventually declines follows from the **law of diminishing returns.** Recall that MP_E is defined in terms of a *fixed* level of capital. The first few workers may increase output substantially because the workers can specialize in narrowly defined tasks. As more and more workers are added to a fixed capital stock, the gains from specialization decline and the marginal product of labor declines. We will assume that the law of diminishing returns operates over some range of employment.

We define the **average product** of labor (or AP_E) as the amount of output produced by the average worker. This quantity is defined by $AP_E = q/E$. In our numerical example, the firm produces 66 units of output when it hires four workers, so the average product is 16.5 units.

Figure 3-1*b* shows the geometric relationship between the marginal product and the average product curves. An easy-to-remember rule is: *The marginal curve lies above the average curve when the average curve is rising, and the marginal curve lies below the average curve when the average curve is falling.* This implies that the marginal curve intersects the average curve at the point where the average curve peaks. It should be clear that the assumption of diminishing returns also implies that the average product of labor curve must eventually decline.

Profit Maximization

To analyze the firm's hiring decisions, we make a very simple assumption about the firm's behavior. The firm wants to maximize profits. The firm's profits are given by

$$\text{Profits} = pq - wE - rK \qquad \textbf{(3-2)}$$

where p is the price at which the firm can sell its output, w is the wage rate (that is, the cost of hiring an additional employee-hour), and r is the price of capital.

In this chapter, we assume that the firm is a small player in the industry. As a result, the price of the output p is unaffected by how much output this firm produces and sells, and the prices of labor (w) and capital (r) are also unaffected by how much labor and capital the firm hires. From the firm's point of view, all of these prices are constants, beyond its control. A firm that cannot influence prices is said to be a **perfectly competitive firm.** Because a perfectly competitive firm cannot influence prices, it maximizes profits by hiring the "right" amount of labor and capital.

3-2 The Short Run

Define the *short run* as a time span that is sufficiently brief that the firm cannot increase or reduce the size of its plant or purchase or sell physical equipment. In the short run, therefore, the firm's capital stock is fixed at some level K_0.

The firm can then determine the output produced by each additional worker by reading the numbers off the marginal product curve. For example, Figure 3-1 indicates that the eighth worker hired increases the firm's output by 11. To obtain the dollar value of what each additional worker produces, we multiply the marginal product of labor times the price of the output. This quantity is called the **value of marginal product** of labor and is given by

$$VMP_E = p \times MP_E \qquad (3\text{-}3)$$

The value of marginal product of labor is the dollar increase in revenue generated by an additional worker, holding capital constant. Suppose the price of the output equals $2. The eighth worker hired would then contribute $22 to the firm's revenue.

The value of marginal product curve is illustrated in Figure 3-2 (and the underlying data are reported in Table 3-1). Because the value of marginal product equals the marginal product of labor times the (constant) price of the output, the value of marginal product curve is simply a "blown-up" version of the marginal product curve.

We define the **value of average product** of labor as

$$VAP_E = p \times AP_E \qquad (3\text{-}4)$$

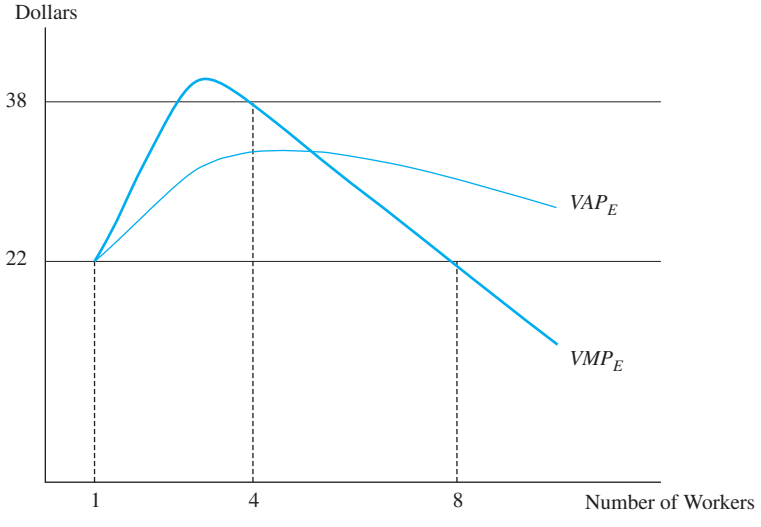
The value of average product gives the dollar value of output per worker. Because both the value of marginal product and the value of average product curves are blown-up versions of the underlying marginal product and average product curves, the geometric relationship in Figure 3-2 is identical to the relationship discussed earlier.

How Many Workers Should the Firm Hire?

The competitive firm can hire all the labor it wants at a constant wage of w dollars. Suppose the wage in the labor market is $22. As shown in Figure 3-2, a profit-maximizing firm will then hire eight workers. At this level of employment, the value of marginal product of labor equals the wage rate *and* the value of marginal product curve is downward sloping, or

$$VMP_E = w \quad \text{and} \quad VMP_E \text{ is declining} \qquad (3\text{-}5)$$

FIGURE 3-2 The Firm's Hiring Decision in the Short Run
A firm hires workers up to the point where the wage rate equals the value of marginal product of labor. If the wage is $22, the firm hires eight workers.



At the profit-maximizing solution, the dollar gain from hiring an additional worker equals the cost of that hire, and it does not pay to further expand because the value of hiring more workers is falling.

The intuition for this result is as follows: Suppose the firm decides to hire only six workers. If the firm hired the seventh worker, it would get more in additional revenues than it would pay out to that worker (the value of marginal product of the seventh worker is $26 and the wage is only $22). A profit-maximizing firm will want to expand and hire more labor. If the firm were to hire more than eight workers, however, the value of marginal product would be lower than the cost of the hire. Suppose, for instance, that the firm wants to hire the ninth worker. It would cost $22 to hire that worker, even though the value of marginal product is only $18.

Note that Figure 3-2 also indicates that the wage would equal the value of marginal product if the firm hired just one worker. At that point, however, the value of marginal product curve is upward sloping. It is easy to see why hiring just one worker does not maximize profits. If the firm hired another worker, that second worker would contribute even more to the firm's revenue than the first *and* cost just as much.

This is why the law of diminishing returns plays such an important role in the theory. If VMP_E kept rising, the firm would maximize profits by expanding indefinitely. It would then be difficult to assume that the firm's decisions do not affect the price of output or the price of labor and capital. In effect, the law of diminishing returns limits the size of a competitive firm.

It is worth emphasizing that the profit-maximizing condition requiring that the wage equal the value of marginal product of labor *does not say* that the firm should set the wage equal to the value of marginal product. The competitive firm has no influence over the wage, and, hence, the firm cannot "set" the wage equal to anything. All the firm can do is set its employment level so that the value of marginal product of labor equals the predetermined wage.

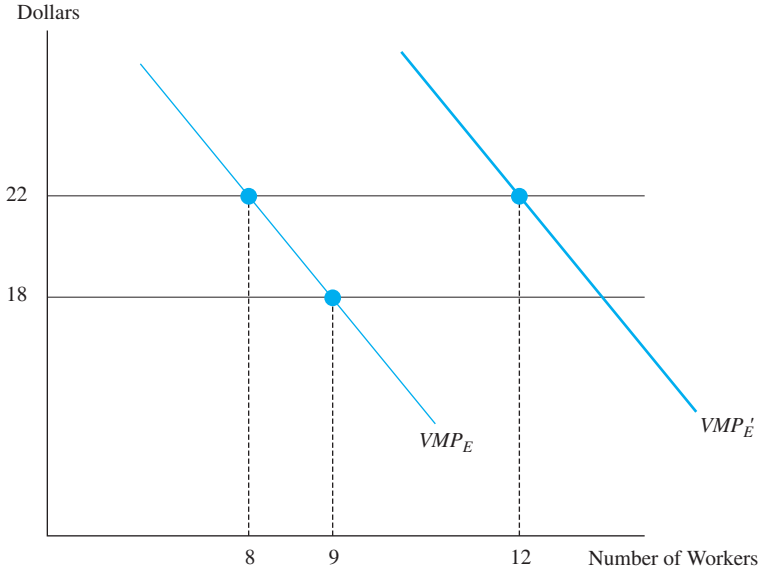
Finally, consider the firm's hiring decision if the competitive wage was very high, such as $38 in Figure 3-2. At this wage, it would seem that the firm should hire four workers. But if the firm hired four workers, the value of the average product of labor ($33) is less than the wage. Because the per-worker contribution to the firm is less than the wage, the firm loses money and exits the market. The only points on the value of marginal product curve that are relevant are the ones that lie on the downward-sloping portion of the curve *below* the point where the VAP_E curve intersects the VMP_E curve. For convenience, we will restrict our attention to that particular segment of the VMP_E curve.

The Short-Run Labor Demand Curve for a Firm

We can now derive the short-run **demand curve for labor.** This demand curve tells us what happens to the firm's employment as the wage changes, holding capital constant. The construction of the short-run labor demand curve is presented in Figure 3-3, which draws the relevant downward-sloping portion of the firm's VMP_E curve. Initially, the wage is $22 and the firm hires eight workers. If the wage falls to $18, the firm hires nine workers. The short-run demand curve for labor, therefore, is given by the value of marginal product curve. Because the value of marginal product of labor declines as more workers are hired, it must be the case that a fall in the wage increases the number of workers hired.

The height of the labor demand curve depends on the price of the output. The short-run demand curve will shift up if the output becomes more expensive. For example, suppose

FIGURE 3-3 The Short-Run Demand Curve for Labor
Because marginal product declines, the short-run demand curve for labor is downward sloping. A drop in the wage from \$22 to \$18 increases employment. An increase in the price of the output shifts the value of marginal product curve upward and increases employment.



that the output price increases, shifting the value of marginal product curve in Figure 3-3 from VMP_E to VMP_E'. If the wage were \$22, the increase in output price raises the firm's employment from 8 to 12 workers. There is, therefore, a positive relation between employment and output price.

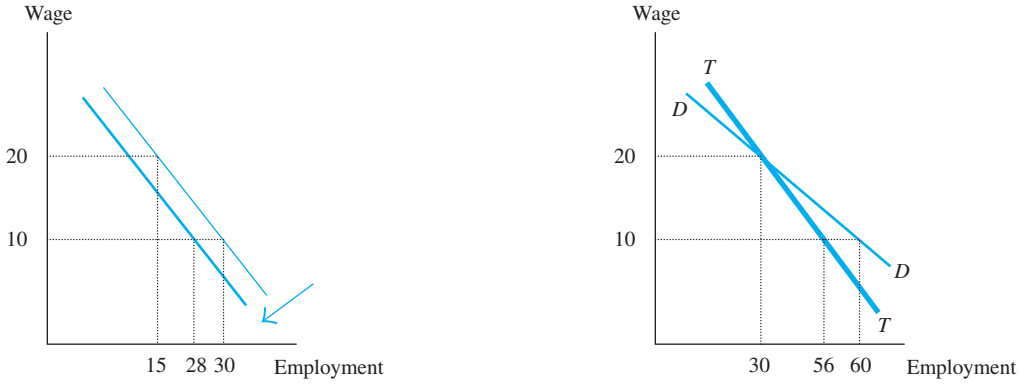
Finally, recall that the short-run demand curve holds capital constant at K_0. We would have derived a different short-run labor demand curve if we had held the capital stock constant at a different level K_1. The relationship between the VMP_E curve and the level of the capital stock is discussed below.

The Short-Run Labor Demand Curve in the Industry

We derived the short-run labor demand curve for a single firm. We can apply the same approach to derive the short-run labor demand curve for every firm in the industry, the group of firms that produce the same output. It would seem that the industry's labor demand curve can be obtained by adding up the demand curves of the individual firms. Suppose that every firm hires 15 workers when the wage is \$20, but hires 30 workers when the wage falls to \$10. We could perhaps get the industry demand curve by summing up the employment across firms. If there were two firms in the industry, this would imply that the industry hires 30 workers when the wage is \$20 and 60 workers when the wage is \$10.

This approach, however, is incorrect because it ignores that the labor demand curve for a firm takes the price of the output as given. Each firm in a perfectly competitive industry

FIGURE 3-4 The Short-Run Demand Curve for the Industry
Each firm in the industry hires 15 workers when the wage is $20. If the wage falls to $10, each firm hires 30 workers. If all firms expand, the output of the industry increases, reducing the price of the output and reducing the value of marginal product, so the labor demand curve of each individual firm shifts slightly to the left. At the lower price of $10, each firm hires only 28 workers. The industry demand curve is not given by the horizontal sum of the firms' demand curves (*DD*), but takes into account the impact of the industry's expansion on output price (*TT*).



(*a*) Individual Firms (*b*) Industry

is small enough that it cannot influence prices. But if all firms in the industry take advantage of the lower wage by hiring more workers, there would be a lot more output and this would imply that the price of the output would fall. As a result, if all firms expand their employment, the value of marginal product (or output price times marginal product) would fall, and the labor demand curve of each individual firm shifts slightly to the left. Industry employment would then expand less than would have been the case if we just added up the demand curves of individual firms.

Figure 3-4 illustrates this point for an industry with two identical firms. As shown in Figure 3-4*a,* each firm hires 15 workers when the wage is $20 and 30 workers when the wage falls to $10. The sum of these two demand curves is given by the curve *DD* in Figure 3-4*b.* It is impossible, however, for every firm in the industry to expand its employment without lowering the price of the output. As a result, the demand curve for each firm shifts back slightly, so that at the lower wage of $10, each firm hires only 28 workers. The industry, therefore, employs 56 workers at the lower wage. The "true" industry labor demand curve is then given by *TT.* This curve is steeper than the industry demand curve one would obtain by just summing horizontally the demand curves of individual firms.

We use an elasticity to measure the responsiveness of employment in the industry to changes in the wage rate. The short-run **elasticity of labor demand** is defined as the percentage change in short-run employment (E_{SR}) resulting from a one percent change in the wage:

$$\delta_{SR} = \frac{\Delta E_{SR}/E_{SR}}{\Delta w/w} = \frac{\Delta E_{SR}}{\Delta w} \cdot \frac{w}{E_{SR}} \qquad (3\text{-}6)$$

Because the short-run demand curve is downward sloping, it must be the case that the elasticity is negative. In our example, the industry hires 30 workers when the wage is \$20 and 56 workers when the wage fall is \$10. The short-run elasticity is:

$$\delta_{SR} = \frac{\%\ \Delta E_{SR}}{\%\ \Delta w} = \frac{(56-30)/30}{(10-20)/20} = -1.733 \qquad \textbf{(3-7)}$$

Labor demand is elastic if the absolute value of the elasticity is greater than one. Labor demand is inelastic if the absolute value of the elasticity is less than one.

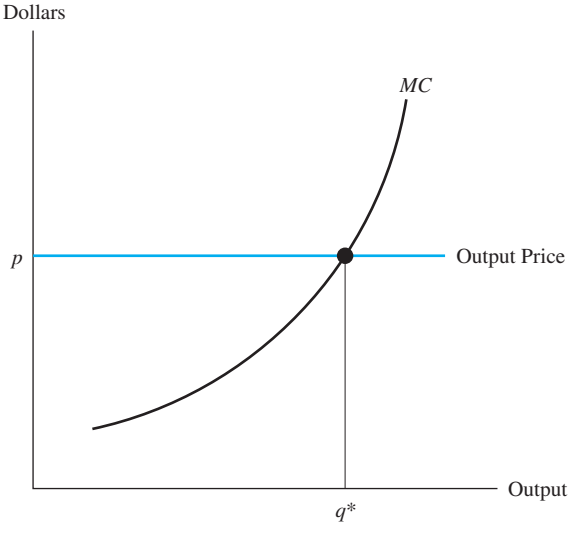
An Alternative Interpretation of the Marginal Productivity Condition

The requirement that firms hire workers up to the point where the value of marginal product of labor equals the wage gives the firm's "stopping rule" in its hiring decision—that is, the rule that tells the firm when to stop hiring. This rule is also known as the **marginal productivity condition.** An alternative and more familiar way of describing profit-maximizing behavior gives the stopping rule for the firm's output: A profit-maximizing firm produces up to the point where the cost of producing an additional unit of output (or **marginal cost**) equals the revenue obtained from selling that unit (or **marginal revenue**).

Figure 3-5 illustrates this condition. The marginal cost (*MC*) curve is upward sloping—as the firm expands, costs increase at an increasing rate. For a competitive firm, the revenue from selling an additional unit of output is given by the constant price *p*. The equality

FIGURE 3-5 The Firm's Output Decision

A profit-maximizing firm produces up to the point where the output price equals the marginal cost of production. This profit-maximizing condition is identical to the one stating that firms hire workers up to the point where the wage equals the value of marginal product.



of price and marginal cost occurs at output q^*. If the firm were to produce fewer than q^* units of output, it would increase its profits by expanding production; the revenue from selling an extra unit of output exceeds the cost of producing that unit. But if the firm were to produce more than q^* units, it would increase its profits by shrinking. The marginal cost of producing those extra units exceeds the marginal revenue.

The profit-maximizing condition equating price and marginal cost (which gives the optimal level of output) is identical to the profit-maximizing condition equating the wage and the value of marginal product of labor (which gives the optimal number of workers). Recall that MP_E tells us how many units of output an additional worker produces. Suppose that $MP_E = 5$. This implies that it takes one-fifth of a worker to produce one extra unit of output. More generally, if one additional worker produces MP_E units of output, then $1/MP_E$ worker produces one unit of output. Each of these workers gets paid a wage of w dollars. Hence, the cost of producing an extra unit of output is equal to

$$MC = w \times \frac{1}{MP_E} \tag{3-8}$$

The profit-maximizing condition that marginal cost equals price can then be written as

$$w \times \frac{1}{MP_E} = p \tag{3-9}$$

By rearranging terms in equation (3-9), we obtain the marginal productivity condition $w = p \times MP_E$. The condition telling the profit-maximizing firm how much output to produce is identical to the one telling the firm how many workers to hire.

3-3 The Long Run

In the long run, the firm's capital stock is not fixed. The firm can expand or shrink its plant size and equipment. Therefore, in the long run, the competitive firm maximizes profits by choosing both how many workers to hire and how much plant and equipment to invest in.

Isoquants

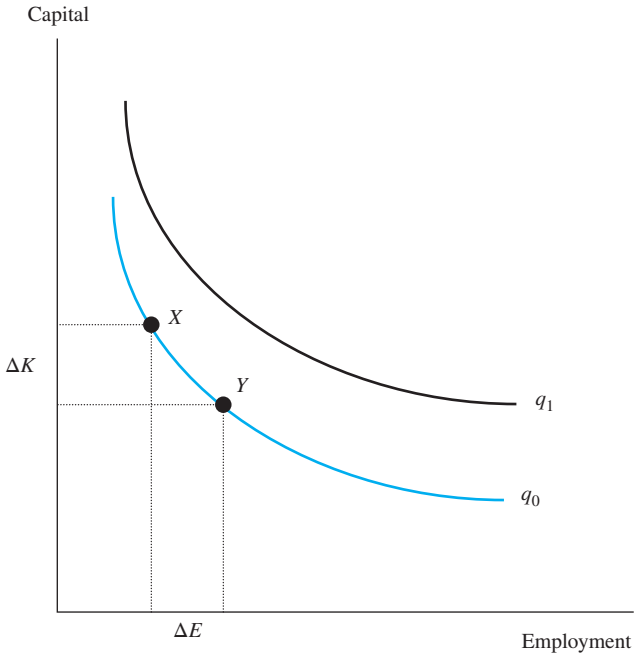
An **isoquant** gives the combinations of labor and capital that produce the same level of output. Isoquants, therefore, describe the production function in exactly the same way that indifference curves describe a utility function. Figure 3-6 shows the isoquants associated with the production function $q = f(E, K)$. The isoquant labeled q_0 gives all the capital–labor combinations that produce q_0 units of output, and the isoquant labeled q_1 gives all the combinations producing q_1 units.

Figure 3-6 illustrates the properties of these constant-output curves:

1. Isoquants must be downward sloping.
2. Isoquants do not intersect.
3. Higher isoquants are associated with higher levels of output.
4. Isoquants are convex to the origin.

These properties correspond exactly to the properties of indifference curves. And, just as the slope of an indifference curve is given by the negative of the ratio of marginal utilities,

FIGURE 3-6 Isoquant Curves
All capital–labor combinations along a single isoquant produce the same level of output. The input combinations at points X and Y produce q_0 units of output. Input combinations that lie on higher isoquants produce more output.



the slope of an isoquant is given by the negative of the ratio of marginal products. In particular,[2]

$$\frac{\Delta K}{\Delta E} = -\frac{MP_E}{MP_K} \qquad (3\text{-}10)$$

The absolute value of this slope is called the **marginal rate of technical substitution.** Convex isoquants imply a *diminishing* marginal rate of technical substitution (or a flatter isoquant) as the firm substitutes more labor for capital.

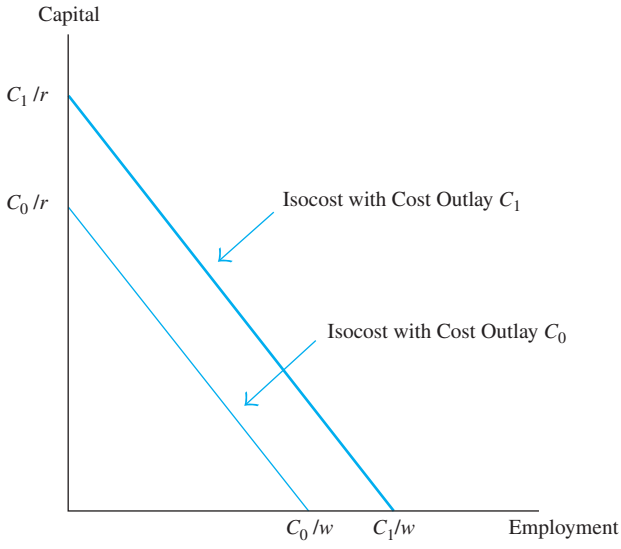
Isocosts

The firm's costs of production, C, are given by

$$C = wE + rK \qquad (3\text{-}11)$$

[2] Let's calculate the slope of the isoquant between points X and Y in Figure 3-6 (assuming that points X and Y are very close to each other). In going from X to Y, the firm hires ΔE more workers, and each of these workers produces MP_E units of output. Hence, the gain in output is given by the product $\Delta E \times MP_E$. In going from X to Y, however, the firm is also getting rid of ΔK units of capital. Each of these units has a marginal product of MP_K. The decrease in output is then given by $\Delta K \times MP_K$. Because output is the same at all points along the isoquant, the gain in output resulting from hiring more workers must equal the reduction in output resulting from cutting the capital stock, so that $(\Delta E \times MP_E) + (\Delta K \times MP_K) = 0$. Equation (3-10) follows by rearranging terms.

FIGURE 3-7 Isocost Lines
All capital–labor combinations along a single isocost curve are equally costly. Capital–labor combinations on a higher isocost curve are costlier. The slope of an isocost equals the ratio of input prices ($-w/r$).



Let's consider how the firm can spend a particular amount of money, say C_0. The firm could decide to hire only capital, in which case it could hire C_0/r units of capital (where r is the price of capital), or it could hire only labor, in which case it could hire C_0/w workers. The line connecting all the various combinations of labor and capital that the firm could hire with an outlay of C_0 dollars is called an **isocost** line, and is illustrated in Figure 3-7.

The isocost line gives the combinations of labor and capital that are equally costly, with higher isocost lines implying higher costs. Figure 3-7 illustrates the isocost lines associated with outlays C_0 and C_1, where $C_1 > C_0$. We can derive the slope of an isocost line by rewriting equation (3-11) as

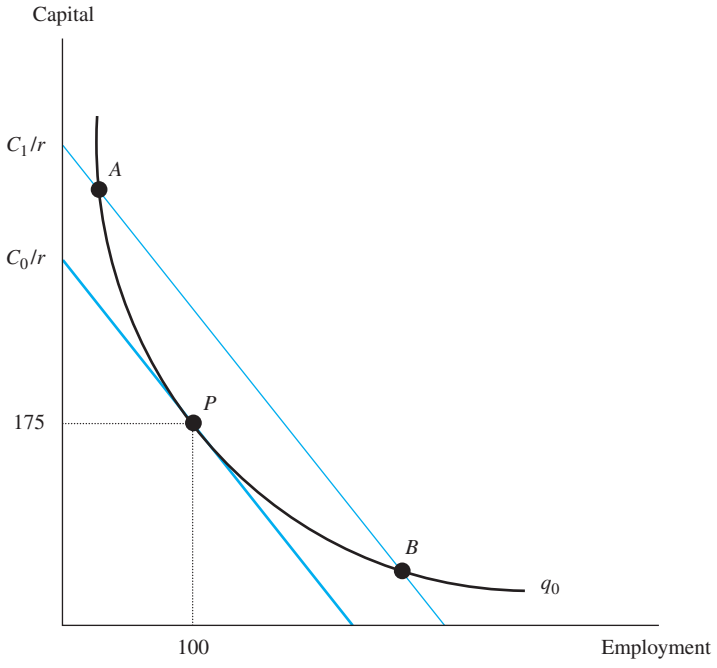
$$K = \frac{C}{r} - \frac{w}{r}E \qquad \textbf{(3-12)}$$

This equation is of the form $y = a + bx$, with intercept C/r and slope $-w/r$. The slope of the isocost line, therefore, is the negative of the ratio of input prices.

Cost Minimization

A profit-maximizing firm producing q_0 units of output obviously wants to produce these units at the lowest possible cost. Figure 3-8 illustrates the solution to this cost-minimization problem. The firm chooses the combination of labor and capital (100 workers and 175 machines) at point P, where the isocost is tangent to the isoquant. At P, the firm produces q_0 units of output at the lowest cost because it uses a capital–labor combination that lies on the lowest isocost. The firm could produce q_0 using other capital–labor combinations, such as points A or B on the isoquant. Those choices, however, would be costlier because they place the firm on a higher isocost line (with a cost outlay of C_1 dollars).

FIGURE 3-8 The Firm's Optimal Combination of Inputs
A firm minimizes the cost of producing q_0 by using the capital–labor combination at point *P,* where the isoquant is tangent to the isocost. All other capital–labor combinations (such as those in points *A* and *B*) lie on a higher isocost.



At the cost-minimizing solution *P,* the slope of the isocost equals the slope of the isoquant, or

$$\frac{MP_E}{MP_K} = \frac{w}{r} \qquad \textbf{(3-13)}$$

Cost minimization, therefore, requires that the marginal rate of technical substitution equal the ratio of input prices. The intuition is easily grasped if we rewrite equation (3-13) as

$$\frac{MP_E}{w} = \frac{MP_K}{r} \qquad \textbf{(3-14)}$$

The last worker hired produces MP_E units of output at a cost of w dollars. If the marginal product of labor is 20 units and the wage is \$10, the ratio MP_E/w implies that the last dollar spent on labor yields two units of output. Similarly, the ratio MP_K/r gives the output yield of the last dollar spent on capital. Cost-minimization requires that the last dollar spent on labor yield as much output as the last dollar spent on capital. In other words, the last dollar spent on each input must give the same "bang for the buck."

The hypothesis that firms minimize the cost of producing a *particular* level of output is often confused with the hypothesis that firms maximize profits. It should be clear that if we constrain the firm to produce q_0 units of output, a profit-maximizing firm must produce

this level of output in the cheapest way possible. Profit-maximizing firms, therefore, will *always* use the combination of labor and capital that equates the ratio of marginal products to the ratio of input prices.

But equation (3-13) was derived by *assuming* that the firm was going to produce q_0 units of output, regardless of any other considerations. A profit-maximizing firm will not choose to produce just any level of output; it will choose to produce the *optimal* level of output, where the marginal cost of production equals the price of the output (or q^* units in Figure 3-5).

Therefore, the condition that the ratio of marginal products equals the ratio of prices does not tell us everything we need to know about profit-maximizing firms in the long run. We saw earlier that for a given level of capital—*including the optimal level of capital*—the firm's employment is determined by equating the wage with the value of marginal product of labor. By analogy, the profit-maximizing condition that tells the firm how much capital to hire is obtained by equating the price of capital with the value of marginal product of capital VMP_K. Therefore, long-run profit maximization requires that labor and capital be hired up to the point where

$$w = p \times MP_E \quad \text{and} \quad r = p \times MP_K \qquad \textbf{(3-15)}$$

The ratio of the two conditions in equation (3-15) implies that the ratio of input prices equals the ratio of marginal products. Put differently, profit maximization implies cost minimization.

3-4 The Long-Run Demand Curve for Labor

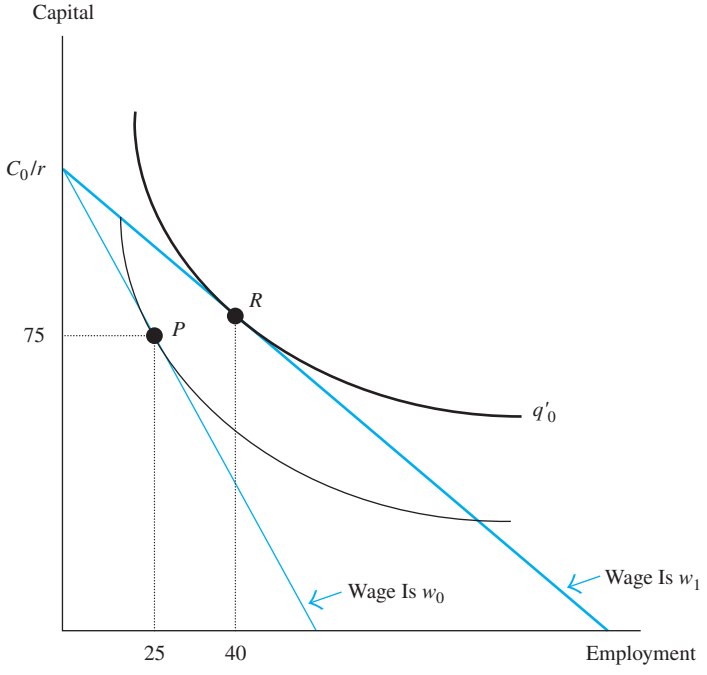
What happens to the firm's long-run demand for labor when the wage changes? We initially consider a firm producing q_0 units of output. Assume that this output happens to be *the* profit-maximizing level of output; at that level of production, output price equals marginal cost. A profit-maximizing firm produces q_0 in the least costly way, using a mix of labor and capital where the ratio of marginal products equals the ratio of input prices. The wage is initially w_0. Figure 3-9 illustrates the optimal combination of inputs at point P, with the firm using 75 units of capital and employing 25 workers. The cost outlay equals C_0 dollars.

Suppose the market wage falls to w_1. How will the firm respond? The absolute value of the slope of the isocost line is equal to the ratio of input prices (or w/r), so the isocost line will be flattened by the wage cut. Because of the resemblance between the wage change in Figure 3-9 and the wage change in the neoclassical model of labor-leisure choice that we discussed in the chapter on labor supply, there is a strong inclination to duplicate the various steps of our earlier geometric analysis.

We have to be extremely careful when drawing the new isocost line, however, because *the obvious way of shifting the isocost line is also the wrong way of shifting it.* As illustrated in Figure 3-9, we might want to shift the isocost by rotating it around the original intercept C_0/r. If this rotation were "legal," the firm would move from P to R. The wage reduction increases the firm's employment from 25 to 40 workers and increases output from q_0 to q_0' units.

Although we are tempted to draw Figure 3-9, the analysis is wrong. The rotation of the isocost around the original intercept C_0/r implies that the firm's cost outlay is being held constant, at C_0 dollars. *Nothing in the theory of profit maximization requires that the firm*

FIGURE 3-9 The Impact of a Wage Reduction, Holding Constant Initial Cost Outlay at C_0
A wage reduction flattens the isocost curve. If the firm were to hold the initial cost outlay constant at C_0 dollars, the isocost would rotate around C_0 and the firm would move from point P to point R. A profit-maximizing firm, however, will not generally want to hold the cost outlay constant when the wage changes.



incur the same costs before and after a wage change. The long-run constraints of the firm are given by the technology (as summarized by the production function) and by the constant price of the output and other inputs (p and r). In general, the firm will not maximize profits by constraining itself to have the same cost outlay before and after a wage change.

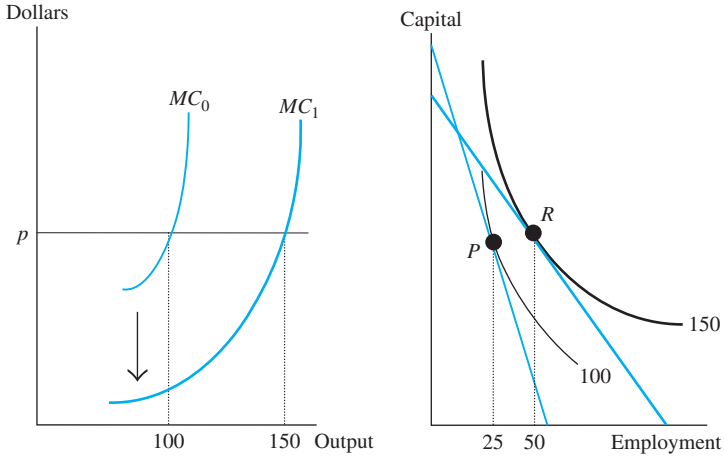
Will the Firm Expand if the Wage Falls?

The wage reduction will typically cut the marginal cost of production.[3] It is cheaper to produce an additional unit of output when labor is cheap than when labor is expensive. The lower wage would then encourage the firm to expand production. Figure 3-10*a* shows the impact of this reduction in marginal cost on the firm's scale. Because the marginal cost curve drops from MC_0 to MC_1, the wage cut encourages the firm to produce 150 units of output rather than 100 units.

The firm, therefore, will "jump" to a higher isoquant, as shown in Figure 3-10*b*. As noted earlier, the cost of producing 150 units of output need not be the same as the cost of

[3] It can be shown that the marginal cost of production falls when the inputs used in the production process are "normal" inputs—in the sense that the firm uses more labor and more capital as it expands, holding the prices of labor and capital constant. The key result of the theory—that the long-run labor demand curve is downward sloping—also holds even if labor were an inferior input.

FIGURE 3-10 Impact of Wage Reduction on Output and Employment of a Profit-Maximizing Firm
(*a*) A wage cut reduces the marginal cost of production and encourages the firm to expand (from producing 100 to 150 units). (*b*) The firm moves from point *P* to point *R,* increasing the number of workers hired from 25 to 50.



(*a*) Firm's Output Decision (*b*) Firm's Hiring Decision

producing only 100 units. As a result, the new isocost need not originate from the same point in the vertical axis as the old isocost. However, a profit-maximizing firm will produce the 150 units of output in the least costly way. The optimal mix of inputs is given by the point on the higher isoquant where the isoquant is tangent to a new isocost, which has a slope equal to w_1/r (and hence is flatter than the original isocost). This solution is given by point *R* in Figure 3-10*b.*

As drawn, the firm's employment increases from 25 to 50 workers. We will see below that the firm will *always* hire more workers when the wage falls, so that the long-run labor demand curve must be downward sloping. Point *R* also implies that the firm uses more capital. We will see below that this need not always be the case. In general, a wage cut can either increase or decrease the amount of capital demanded.

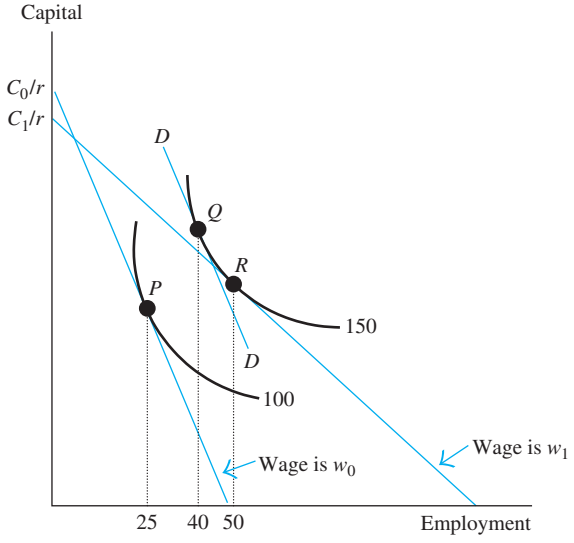
Substitution and Scale Effects

In our derivation of a worker's labor supply curve, we decomposed the impact of a wage change on hours of work into income and substitution effects. This section uses a similar decomposition to assess the impact of a wage change on the firm's employment. In particular, the wage cut reduces the price of labor relative to that of capital, encouraging the firm to adjust its input mix so that it is more labor intensive. In addition, the wage cut reduces the marginal cost of production and encourages the firm to expand. As the firm expands, it wants to hire even more workers.

Figure 3-11 illustrates these two effects. The firm is initially at point *P,* facing a wage of w_0, producing 100 units of output, and hiring 25 workers. When the wage falls to w_1, the firm moves to point *R,* producing 150 units of output and hiring 50 workers.

It is useful to view the move from *P* to *R* as a two-stage move. To conduct the decomposition, Figure 3-11 introduces a new isocost line, labeled *DD.* This isocost is tangent to

FIGURE 3-11 Substitution and Scale Effects
A wage cut generates substitution and scale effects. The scale effect (the move from point *P* to point *Q*) encourages the firm to expand, increasing the firm's employment. The substitution effect (from *Q* to *R*) encourages the firm to use a more labor-intensive method of production, further increasing employment.



the new isoquant (which produces 150 units of output), but parallel to the isocost that the firm faced before the wage reduction. In other words, the absolute value of the slope of the *DD* isocost is equal to w_0/r, the original price ratio. The tangency point between this new isocost and the new isoquant is given by point *Q*.

We define the move from *P* to *Q* as the **scale effect.** The scale effect indicates what happens to the demand for the firm's inputs as the firm expands production, holding input prices constant. As long as capital and labor are "normal inputs," the scale effect increases both employment (from 25 to 40 workers) and the capital stock.

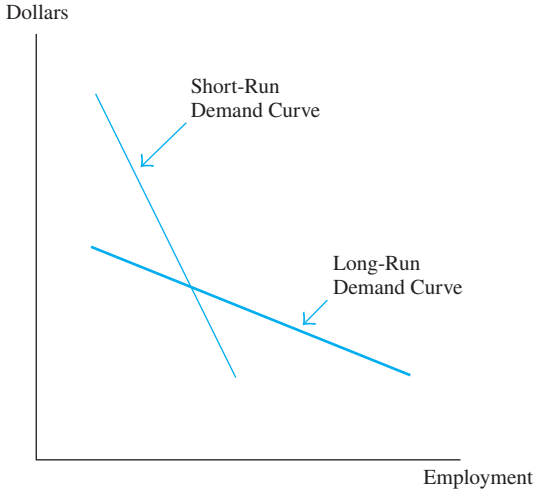
The wage cut also encourages the firm to adopt a different method of production, one that is more labor intensive. The **substitution effect,** given by the move from *Q* to *R*, shows what happens to the firm's employment as the wage changes, holding output constant. As drawn, the substitution effect raises the firm's employment from 40 to 50 workers. Note that the substitution effect must decrease the firm's demand for capital.

Both the substitution and scale effects induce the firm to employ more workers as the wage falls. As drawn in Figure 3-12, the firm hires more capital when the wage falls, so that the scale effect (which increases the demand for capital) outweighs the substitution effect (which reduces the demand for capital). The firm would use less capital if the substitution effect dominated the scale effect.

We use the concept of an elasticity to measure the responsiveness of changes in long-run employment (E_{LR}) to changes in the wage. The long-run elasticity of labor demand is given by

$$\delta_{LR} = \frac{\Delta E_{LR}/E_{LR}}{\Delta w/w} = \frac{\Delta E_{LR}}{\Delta w} \cdot \frac{w}{E_{LR}} \qquad \textbf{(3-16)}$$

FIGURE 3-12 The Short- and Long-Run Labor Demand Curves
In the long run, the firm can take full advantage of the economic opportunities introduced by a change in the wage. The long-run demand curve is more elastic than the short-run demand curve.



Because the long-run labor demand curve is downward sloping, the long-run elasticity of labor demand must be negative.

An important principle in economics states that consumers and firms can respond more easily to changes in the economic environment when they face fewer constraints. Put differently, extraneous constraints prevent us from taking full advantage of the opportunities presented by changing prices. This principle suggests that the long-run demand curve for labor is more elastic than the short-run demand curve, as illustrated in Figure 3-12. In the short run, the firm is "stuck" with a fixed capital stock and finds it difficult to change its scale. In the long run, firms are more responsive, can adjust both labor and capital, and can fully take advantage of changes in the price of labor.

Estimates of the Labor Demand Elasticity

Many studies attempt to estimate the elasticity of labor demand, typically by correlating observed changes in employment at a firm with wage changes.[4] Given our discussion of the problems encountered when estimating the labor supply elasticity, it should not be surprising that there is also a lot of variation in estimates of the labor demand elasticity. Although most of the estimates indicate that the labor demand curve is downward sloping, the range of the estimates is considerable.

Despite this dispersion, there is some consensus that the short-run elasticity lies between −0.4 and −0.5. In other words, a 10 percent increase in the wage reduces employment by perhaps 4 to 5 percentage points in the short run. The evidence also suggests that

[4] An encyclopedic survey of this literature is given by Daniel S. Hamermesh, *Labor Demand,* Princeton, NJ: Princeton University Press, 1993.

Theory at Work

CALIFORNIA'S OVERTIME REGULATIONS

The Fair Labor Standards Act of 1938 requires that covered workers be paid 1.5 times the wage for any hours worked in excess of 40 hours per week. Unlike most states, California also requires that workers be paid 1.5 times the wage for any hours worked in excess of 8 hours per *day*—even if they work fewer than 40 hours during the week. Before 1974, this legislation applied only to women. After 1980, the legislation covers both men and women.

The theory of labor demand makes a clear prediction about how the legislation should affect the probability that men in California work more than 8 hours per day. That probability should have declined between the 1970s and the 1980s—as the overtime-per-day regulation was extended to cover men and employers switched to cheaper methods of production.

The table below shows that 17.1 percent of California's working men worked more than 8 hours per day in 1973. By 1985, only 16.9 percent of working men worked more than 8 h per day. Before we can attribute this slight reduction in the length of the workday to the overtime legislation, we need to know what would have happened to the length of the workday in the absence of the legislation. In other words, we need a control group.

One possible control group could be working men in other states—men whose workday was unaffected by the change in California's policies. The fraction of men in other states working more than 8 hours per day rose during the same period, from 20.1 to 22.8 percent. The difference-in-differences estimate of the impact of California's legislation implies a substantial reduction of 2.9 percentage points on the probability of men working more than 8 hours per day. Alternatively, the control group could be California's working women—who were always covered by the legislation. The probability that their workday lasted more than 8 hours also rose during the period. The data, therefore, imply that the higher labor cost substantially reduced the probability that working men worked more than 8 hours per day.

EMPLOYMENT EFFECTS OF OVERTIME REGULATION IN CALIFORNIA

Source: Daniel S. Hamermesh and Stephen J. Trejo, "The Demand for Hours of Labor: Direct Estimates from California," *Review of Economics and Statistics* 82 (February 2000): 38–47.

	Treatment Group	Control Group	
	Men in California (%)	Men in Other States (%)	Women in California (%)
Workers working more than 8 hours per day in			
1973	17.1	20.1	4.0
1985	16.9	22.8	7.2
Difference	−0.2	2.7	3.2
Difference-in-differences	—	**−2.9**	**−3.4**

estimates of the long-run labor demand elasticity cluster around −1, so the long-run labor demand curve is indeed more elastic than the short-run curve. About one-third of the long-run elasticity is typically attributed to the substitution effect and about two-thirds to the scale effect.

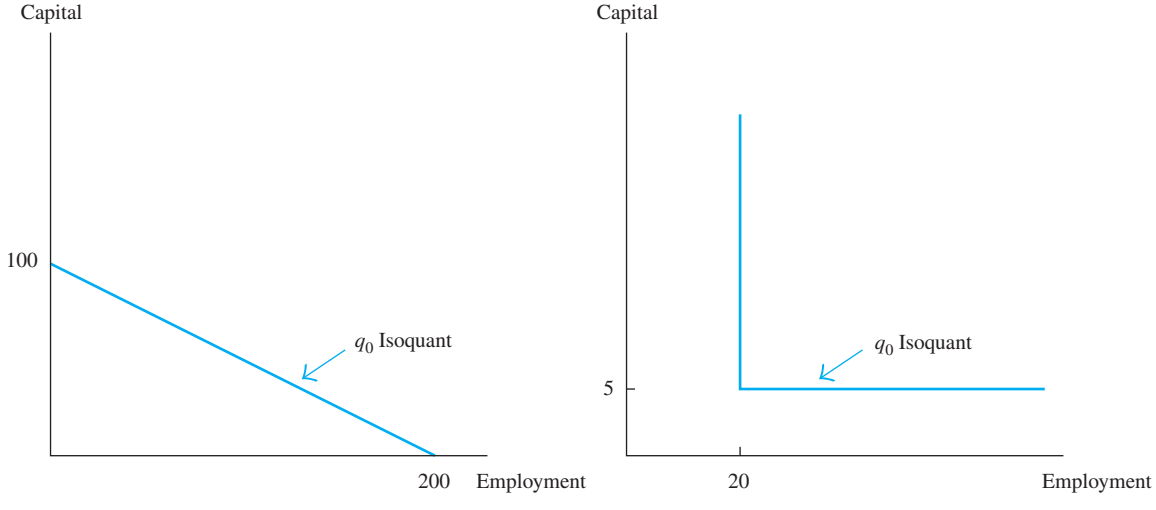
3-5 The Elasticity of Substitution

The size of the firm's substitution effect depends on the curvature of the isoquant. Figure 3-13 illustrates two extreme situations. In Figure 3-13*a,* the isoquant is a straight line, with a slope equal to −0.5. In other words, output remains constant whenever the firm lays off two workers and replaces them with one machine. The "rate of exchange" between labor and capital is the same regardless of how many workers or how much capital the firm already has. The marginal rate of technical substitution is constant when the isoquant is a line. Whenever any two inputs can be substituted at a constant rate, the two inputs are called **perfect substitutes.**[5]

Figure 3-13*b* illustrates the other extreme. The right-angled isoquant implies that using 20 workers and 5 machines yields q_0 units of output. If we hold capital constant at five units, adding more workers has no impact on output. Similarly, if we hold labor constant at 20 workers, adding more machines has no impact on output. A firm that does not wish to throw away money has only one recipe for producing q_0: Use 20 workers and 5 machines! When the isoquant between any two inputs is right-angled, the inputs are called **perfect complements.**

The substitution effect is very large when labor and capital are perfect substitutes. When the isoquant is linear, the firm minimizes the cost of producing output q_0 by using either

FIGURE 3-13 Isoquants When Inputs Are Either Perfect Substitutes or Perfect Complements
Capital and labor are perfect substitutes if the isoquant is linear (so that two workers can always be substituted for one machine). The two inputs are perfect complements if the isoquant is right-angled. The firm then gets the same output when it hires 5 machines and 20 workers as when it hires 5 machines and 25 workers.



(*a*) Perfect Substitutes (*b*) Perfect Complements

[5] The definition of perfect substitutes does *not* imply that the two inputs have to be exchanged on a one-to-one basis; that is, one machine hired for each worker laid off. Our definition only requires that the rate at which capital can be exchanged for labor is constant.

100 machines or 200 workers, depending on which alternative is cheaper. If input prices change sufficiently, the firm will jump from one extreme to the other.

In contrast, there is no substitution effect when the two inputs are perfect complements. Because there is only one recipe for producing q_0, a change in the wage does not alter the input mix at all. The firm must always use 20 workers and 5 machines to produce q_0, regardless of the prices of labor and capital.

There are many substitution possibilities in between these two extremes, depending on the curvature of the isoquant. The closer the isoquant is to a line, the larger the size of the substitution effect. To measure the curvature of the isoquant, we use the **elasticity of substitution.** This elasticity is defined by

$$\sigma = \frac{\text{Percent change in } K/L}{\text{Percent change in } w/r} \qquad \textbf{(3-17)}$$

The elasticity of substitution gives the percentage change in the capital/labor ratio resulting from a 1 percent change in the relative price of labor, holding output constant. As the relative price of labor increases, the substitution effect tells us that the capital/labor ratio increases (that is, the firm gets rid of labor and replaces it with capital). At the extremes, the elasticity of substitution is zero if the isoquant is right-angled (as in Figure 3-14*b*), and infinite if the isoquant is linear (as in Figure 3-14*a*). The elasticity will be a positive number for isoquants that have the usual convex shape. The size of the substitution effect depends directly on the size of the elasticity of substitution.

3-6 What Makes Labor Demand Elastic?

The famous **Marshall's rules of derived demand** describe the situations that are likely to generate elastic labor demand curves in the industry. In particular:

- *Labor demand is more elastic the greater the elasticity of substitution.* This rule follows from the fact that the size of the substitution effect depends on the curvature of the isoquant. The greater the elasticity of substitution, the more the isoquant looks like a straight line, and the more "similar" labor and capital are in the production process. The firm can then easily replace labor for machines as the wage rises.
- *Labor demand is more elastic the greater the elasticity of demand for the output.* An increase in the wage will increase the marginal cost of production, raising output price and reducing consumer demand for the product. Because less output is being sold, firms cut employment. The more consumers respond to the higher price (that is, the more elastic the demand curve for the output), the larger the cut in employment and the more elastic the industry's labor demand curve.
- *Labor demand is more elastic the greater labor's share in total costs.* Suppose labor is an "important" input in production, in the sense that labor's share of total costs is large. This situation might occur when production is very labor intensive, as with a firm using highly skilled workers to produce handmade furniture. Even a small increase in the wage would substantially increase the marginal cost of production. This increase in marginal cost raises output price, cutting consumer demand for the expensive furniture. Firms, in response, cut back on employment. But if labor is "unimportant," so that labor

makes up only a small share of total costs, a wage increase has only a small impact on marginal cost, on the price of the output, and on consumer demand. There is little need for the firm's employment to shrink.[6]

- *The demand for labor is more elastic the greater the supply elasticity of other factors of production.* We have assumed that firms can hire as much capital as they want at the constant price *r.* Suppose there is a wage increase and firms want to substitute from labor to capital. If the supply curve of capital is inelastic, so that the price of capital increases substantially as the firm acquires more and more capital, the economic incentives for moving along an isoquant are reduced. In other words, it is not quite as profitable to get rid of labor and use capital instead. The demand curve for labor, therefore, is more elastic the easier it is to increase the capital stock (that is, the more elastic the supply curve of capital).

Union Behavior

The behavior of labor unions shows how Marshall's rules help us understand various aspects of the labor market. Consider a competitive industry that is not unionized. A union wants to organize the industry's workforce, and promises the workers that collective bargaining will increase their wage substantially. Because the labor demand curve is downward sloping, firms might respond to the higher wage by moving up the demand curve and laying off some workers. The union's organizing drive will have a greater chance of being successful when the demand curve for labor is inelastic. An inelastic demand curve ensures that few workers would be laid off even if they get a huge union-mandated wage increase. It is in the union's interests, therefore, to take actions that make the industry's labor demand curve more inelastic.

It is then not surprising that unions resist technological advances that increase the possibility of substituting between labor and capital. The typesetters' unions, for example, long objected to the introduction of computerized typesetting equipment in the newspaper industry. This behavior was an obvious attempt to reduce the value of the elasticity of substitution. A smaller elasticity of substitution reduces the size of the substitution effect and makes the demand curve for labor more inelastic.

Unions will also want to limit the availability of goods that compete with the output of unionized firms. The United Auto Workers (UAW) was a strong supporter of policies that made it difficult for Japanese cars to crack the U.S. market. If the UAW obtained a huge wage increase for its workers, the price of American-made cars would rise substantially. This price increase would drive potential buyers toward foreign imports. But if the union could prevent the importation of Toyotas, Nissans, and Hondas, consumers would have

[6] Marshall's third rule holds only when the absolute value of the elasticity of product demand exceeds the elasticity of substitution. The reason for this exception follows from the fact that we can arbitrarily make the labor input ever less important by redefining it in seemingly irrelevant ways. For example, we can subdivide the labor input of workers producing handmade furniture into Irish workers, Italian workers, Mexican workers, and so on. Each of these new labor inputs would obviously make up a very small fraction of total costs, but it is incorrect to say that the demand curve for Irish workers is less elastic than the demand curve for all workers. See Saul D. Hoffman, "Revisiting Marshall's Third Law: Why Does Labor's Share Interact with the Elasticity of Substitution to Decrease the Elasticity of Labor Demand," *Journal of Economic Education,* 40, no. 4 (2009): 437–445

few alternatives to buying a high-priced American-made car. It is in the union's interests, therefore, to reduce the elasticity of product demand by limiting the variety of goods that are available to consumers.

Marshall's rules also imply that unions are more likely to be successful when the share of labor costs is small. Unions can then make high wage demands without raising the marginal cost (and hence the price) of the output very much. Unions that organize small groups of workers such as electricians or carpenters tend to be very successful in getting sizable wage increases.[7] Because these specialized occupations make up a small fraction of total labor costs, the demand curve for these workers is relatively inelastic.

Finally, unions often attempt to raise the price of other inputs, particularly nonunion labor. The Davis–Bacon Act requires that contractors involved in publicly financed projects pay the "prevailing wage" to construction workers. Not surprisingly, the prevailing wage is typically defined as the union wage, even if the contractor hires nonunion labor. This type of regulation raises the cost of switching from union labor to other inputs. Union support of prevailing wage laws, therefore, can be interpreted as an attempt to make the supply of other factors of production more inelastic and reduce the elasticity of demand for union labor.

3-7 Factor Demand with Many Inputs

Although we have assumed that the production function has only two inputs, labor and capital, we can easily extend the theory to account for more realistic production processes. There are clearly many different types of workers (such as skilled and unskilled) and many different types of capital (such as old machines and new machines). The technology is then summarized by the production function:

$$q = f(x_1, x_2, x_3, \ldots, x_n) \qquad \textbf{(3-18)}$$

where x_i denotes the quantity of the i^{th} input. Define the marginal product of the i^{th} input, or MP_i, as the change in output resulting from a one-unit increase in that input, holding constant the quantities of all other inputs.

We can use this production function to derive the short- and long-run demand curves for a particular input. It will still be true that a profit-maximizing firm hires the i^{th} input up to the point where its price (or w_i) equals the value of marginal product:

$$w_i = p \times MP_i \qquad \textbf{(3-19)}$$

As long as we assume that the law of diminishing returns holds, all of the key results derived in the case of a two-factor production function continue to hold. The short-run and long-run demand curves for each input are downward sloping; the long-run demand curve is more elastic than the short-run demand curve; and a wage change generates both a substitution effect and a scale effect.

[7] These unions are called "craft unions," in contrast to the "industrial unions" that unionize all workers in a given industry (like the UAW).

The presence of many inputs raises the possibility that the demand for input i might increase when the price of input j increases, but might fall when the price of input k increases. To measure how the demand for input i responds to changes in the price of other inputs, we define the **cross-elasticity of factor demand** as

$$\delta_{ij} = \frac{\% \Delta x_i}{\% \Delta w_j} \qquad \textbf{(3-20)}$$

The cross-elasticity gives the percentage change in the demand for input i resulting from a one percent change in the wage of input j.

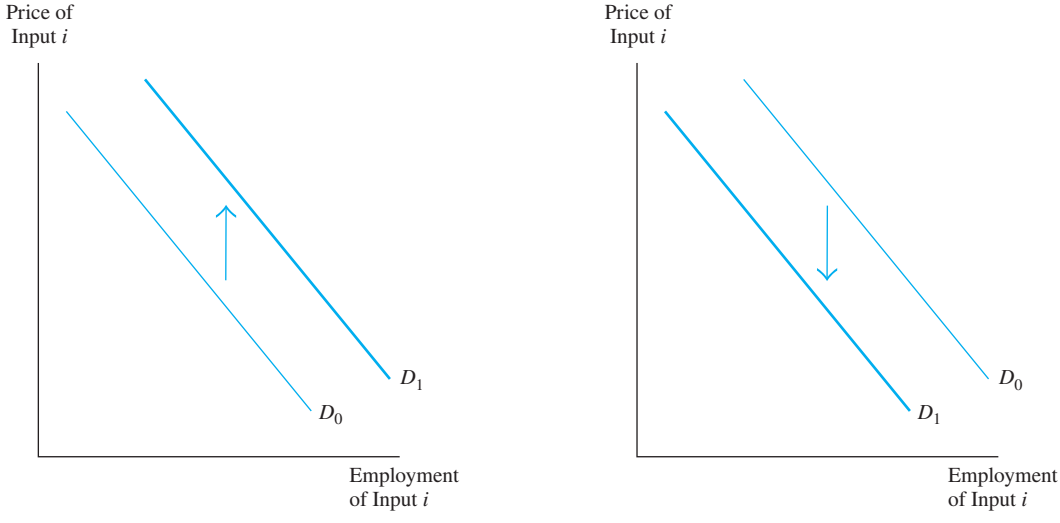
The sign of the cross-elasticity provides one definition of whether any two inputs are substitutes or complements in production. If the cross-elasticity is positive, so that the demand for input i increases when the wage of input j rises, inputs i and j are said to be substitutes. After all, the increase in w_j increases the demand for input i at the same time that it reduces the demand for input j. The two inputs are substitutes because the firm is getting rid of the more expensive input and replacing it with the relatively cheaper input.

If the cross-elasticity of factor demand is negative, the demand for input i falls as a result of the increase in w_j, and inputs i and j are said to be complements. The inputs are complements because they both respond in exactly the same way to a rise in w_j—the firm uses less of both inputs. Put differently, the two inputs "go together."

Figure 3-14 illustrates this definition of substitutes and complements in terms of shifting demand curves. In Figure 3-14*a,* the two inputs are substitutes. As input j became more expensive, employers substituted toward input i and the demand curve for input i shifts up.

FIGURE 3-14 The Demand Curve for Input *i* Responds to a Price Increase in Input *j*

(*a*) The demand curve for input i shifts up if the two inputs are substitutes. (*b*) The demand curve for input i shifts down if the two inputs are complements.



(*a*) inputs are i and j are substitutes

(*b*) inputs i and j are complements

In Figure 3-14*b,* the two inputs are complements. The demand curve for input *i* shifts down when the price of input *j* rose. In other words, there is less demand for both inputs when input *j* becomes more expensive.

The available evidence suggests that unskilled labor and capital are substitutes, but that skilled labor and capital are complements.[8] In other words, as the price of machines falls, employers hire fewer unskilled workers. At the same time, however, the demand for skilled workers rises because skilled workers and capital equipment "go together."

This result is known as the **capital-skill complementarity hypothesis.** The hypothesis has important policy implications. For example, government subsidies to investments in physical capital (such as an investment tax credit) affects different groups of workers differently. Because the tax credit lowers the price of capital, it increases the demand for capital, reduces the demand for unskilled workers, and increases the demand for skilled workers. An investment tax credit, therefore, might spur investment in the economy, but also worsens the relative economic conditions of less-skilled workers.[9]

3-8 Overview of Labor Market Equilibrium

We have examined the factors that encourage many of us to enter the workforce and that encourage firms to demand a particular number of workers. The labor market is the place where workers looking for jobs and firms looking for workers finally meet and compare what they have to offer. This interaction between workers and firms determines the **equilibrium** wage and employment levels: the levels that "balance" the number of hours that workers wish to work with the number of hours that firms wish to employ. This section briefly describes the equilibrium. The subsequent chapter on labor market equilibrium analyzes the properties of this solution in greater detail.

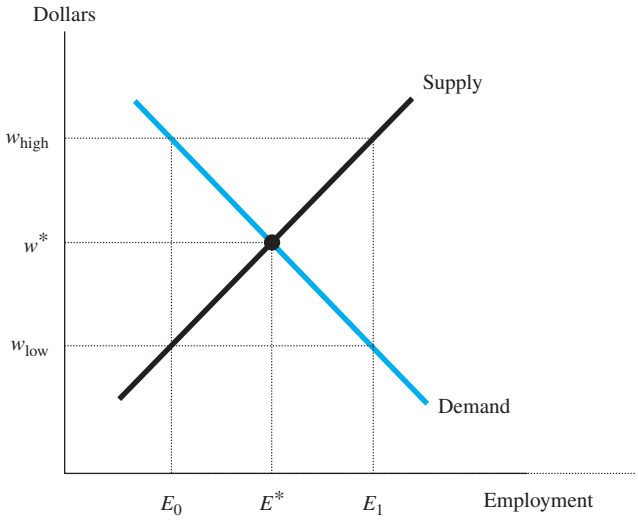
Figure 3-15 illustrates the labor demand and labor supply curves in a particular labor market. As drawn, the supply curve slopes up, so we are assuming that substitution effects dominate income effects. The demand curve is negatively sloped. The equilibrium wage and employment levels are given by the point where the supply and demand curves intersect. A total of E^* workers are employed and each receives the market wage w^*.

To see why the intersection represents a labor market equilibrium, suppose that workers were getting paid a wage of w_{high}, above the equilibrium wage. At this wage, the demand curve indicates that firms are only willing to hire E_0 workers, and the supply curve indicates that E_1 persons are looking for work. A wage above the equilibrium level, therefore, generates a surplus of persons competing for the few available jobs, putting downward pressure on the wage. If firms offered a wage below equilibrium, such as w_{low}, the situation would be exactly reversed. Employers want to hire many workers, but few persons are willing to work at the going wage, putting upward pressure on the wage.

[8] Zvi Griliches, "Capital-Skill Complementarity," *Review of Economics and Statistics* 51 (November 1969): 465–468; and Claudia Goldin and Lawrence F. Katz, "The Origins of Technology-Skill Complementarity," *Quarterly Journal of Economics* 113 (August 1998): 693–732.

[9] The increasing "polarization" in labor market outcomes between low- and high-skill workers is documented by David H. Autor and David Dorn, "The Growth of Low-Skill Service Jobs and the Polarization of the U.S. Labor Market," *American Economic Review* 103 (August 2013): 1553–1597.

FIGURE 3-15 Wage and Employment Determination in a Competitive Market
In a competitive labor market, equilibrium is attained where supply equals demand. The equilibrium wage is w^* and E^* workers are employed.



Once the equilibrium wage is attained, the conflicting desires of employers and workers have been balanced. The number of workers who are looking for work exactly equals the number of workers that employers want to hire. In the absence of any other economic shocks, the equilibrium level of the wage and employment can persist indefinitely.

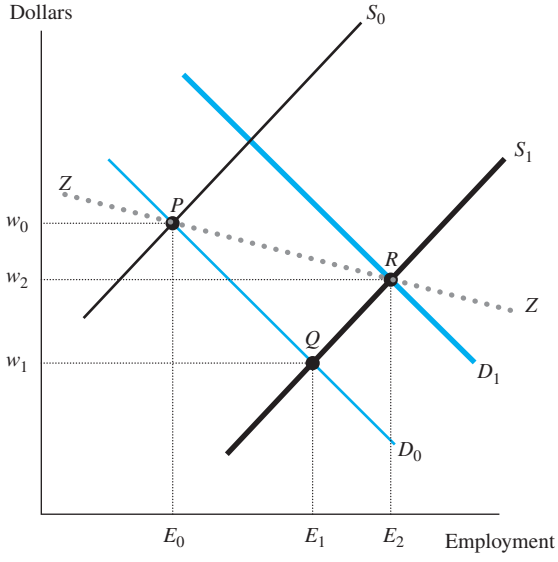
3-9 Rosie the Riveter as an Instrumental Variable

Much of the state-of-the-art research in labor economics involves trying to estimate labor demand and labor supply curves for particular groups. The findings reached by these studies are often used to predict how particular labor market shocks or policy changes will alter earnings and employment opportunities.

The typical attempt to estimate a labor demand curve starts by observing data on employment and wages in a particular labor market—for example, the employment and wages of women. Figure 3-16 shows how the observed employment and wage data are generated by our theory. Initially, the labor market is in equilibrium at point *P,* with wage w_0 and employment E_0. Suppose that the supply curve of women shifts to the right. The new equilibrium would be at *Q,* with wage w_1 and employment E_1. The observed data consist of the pair of wages (w_0 and w_1) and the pair of employment statistics (E_0 and E_1). Figure 3-16 shows that these data can be used to trace out (or *identify*) the labor demand curve D_0. In other words, if we could observe a real-world situation where the only curve that shifted was the supply curve, the resulting data on wages and employment would allow us to estimate the labor demand elasticity.

In most real-world situations, however, both the supply curve and the demand curve are probably shifting at the same time. When both curves shift, the new equilibrium would be

FIGURE 3-16 Shifts in Supply and Demand Curves Generate Observed Wage and Employment Data
The market is initially in equilibrium at *P,* and we observe wage w_0 and employment E_0. If only the supply curve shifts, we can observe w_1 and E_1, and the available data lets us to trace out the labor demand curve. But if both the supply and demand curves shift, we observe w_2 and E_2, and the available data trace out the curve *ZZ,* which does not provide any information about either the supply or the demand curve.



at a point like *R,* with wage w_2 and employment E_2. The data we now observe consist of the pair of wages (w_0 and w_2) and the pair of employment statistics (E_0 and E_2). These data would allow us to trace out the curve *ZZ,* a curve that provides no information whatsoever about either the labor supply elasticity or the labor demand elasticity. When the two curves are moving at the same time, therefore, the resulting data on wages and employment do not help us identify the underlying structure of the labor market. Put differently, the resulting data (that is, the line *ZZ*) could not be used to predict how a particular policy shift (such as a tax credit for child care expenses) would affect female wages and employment.

The "trick" for estimating the labor demand elasticity, therefore, lies in finding a situation where some factor shifts the supply curve but leaves the demand curve untouched. We call a variable that shifts one of the curves and not the other an **instrument** or an **instrumental variable.** The availability of an instrument for supply lets us then use the **method of instrumental variables** to estimate the labor demand elasticity.[10]

A recent study provides a simple illustration of how particular historical events generate instruments that can be used to estimate the labor demand curve.[11] Nearly 16 million men were mobilized to serve in the Armed Forces during World War II, and most of them were

[10] Analogously, an instrument that shifted only the demand curve would allow us to estimate the labor supply elasticity.

[11] Daron Acemoglu, David H. Autor, and David Lyle, "Women, War and Wages: The Effect of Female Labor Supply on the Wage Structure at Midcentury," *Journal of Political Economy* 112 (June 2004): 497–551.

sent overseas. This shrinking in the number of male workers drew many women into the civilian workforce, giving rise to the stereotype of Rosie the Riveter, a woman who aided the war effort by performing "man's work." In 1940, only 28 percent of women over the age of 15 participated in the labor force. By 1945, the female participation rate was over 34 percent. Although some of these women left the workforce after the war, many of them stayed, permanently increasing the number of working women.

To understand how the method of instrumental variables works, it is crucial to get a better sense of the historical circumstances. In October 1940, the Selective Service Act began a mandatory draft registration for all men aged 21–35. By 1947, when the draft finally ended, six separate registrations had been mandated, eventually requiring all men aged 18–64 to register. After each registration, the local draft boards used lotteries to determine the order in which registrants were called to active duty.

The draft boards were authorized to grant deferments to some men, typically based on a man's marital and parental status and on whether he had skills that were essential to civilian production. Farmers, for instance, were deferred because food was needed to support the war effort. Because of these deferments, men living in farm states were less likely to be drafted than men living in more urban states like New York or Massachusetts. Similarly, because most military units were segregated, relatively few blacks were drafted, and the geographic distribution of the black population created even more geographic dispersion in mobilization rates. Table 3-2 reports the mobilization rate for the states, defined as the proportion of registered men aged 18–44 who served in the military between 1940 and 1945. The interstate variation is substantial; 41 percent in Georgia, 50 percent in California, and 55 percent in Massachusetts.

The mobilization rate provides the instrument that shifts the supply curve of female labor differently in different states. After all, Rosie would be more likely to become a riveter in those states where draft boards sent a larger fraction of men into active duty. As Figure 3-17*a* shows, there is indeed a positive correlation between the 1939–1949 growth in female employment and the state's mobilization rate. The regression line (with standard errors in parentheses) is

$$\begin{aligned}&\text{Percent change in female employment} = \\ &\underset{(31.88)}{-94.56} + \underset{(0.67)}{2.62}\,\text{Mobilization rate}\end{aligned} \qquad \textbf{(3-21)}$$

This regression implies that a 1-point increase in the state's mobilization rate increased female labor supply by 2.62 percent.

It turns out that mobilization rates were also correlated with the wage growth of female workers. Figure 3-17*b* shows the negative relation between the 1939–1949 percent change in the female wage and the mobilization rate across states. Female wages grew least in the states where more men went off to war. The regression line relating the two variables is

$$\begin{aligned}&\text{Percent change in female wage} = \\ &\underset{(21.45)}{171.69} - \underset{(0.45)}{2.58}\,\text{Mobilization rate}\end{aligned} \qquad \textbf{(3-22)}$$

A 1-point increase in the mobilization rate led to a 2.58 percent drop in the female wage.

The regressions reported in equations (3-23) and (3-24) can now be used to estimate the labor demand elasticity. For every 1-point increase in the mobilization rate, female

TABLE 3-2 Mobilization Rate of Men and Changes in Female Wages and Employment, 1939–1949

Source: Daron Acemoglu, David H. Autor, and David Lyle, "Women, War and Wages: The Effect of Female Labor Supply on the Wage Structure at Midcentury," *Journal of Political Economy* 112 (June 2004): 497–551. The mobilization rate gives the proportion of men aged 18–44 who served in the military between 1940 and 1945; the percent change in female employment gives the log change in the total number of nonfarm weeks worked by women aged 14–64; the percent change in the female wage gives the (deflated) change in the log weekly wage of women employed full time multiplied by 100.

State	Mobilization Rate (%)	Change in Female Employment (%)	Change in Female Wage (%)	State	Mobilization Rate (%)	Change in Female Employment (%)	Change in Female Wage (%)
Alabama	43.6	20.3	81.0	North Carolina	42.1	23.3	51.6
Arkansas	43.6	19.2	79.5	North Dakota	41.8	−12.5	51.8
Arizona	49.4	70.2	38.4	Nebraska	46.3	30.4	49.0
California	50.0	65.7	31.3	New Hampshire	53.0	20.1	41.8
Colorado	49.7	54.5	50.2	New Jersey	49.7	24.3	35.7
Connecticut	49.4	27.9	34.5	New Mexico	47.8	51.1	50.6
Delaware	46.9	39.4	24.6	New York	48.4	24.9	33.7
Florida	47.7	35.2	69.9	Ohio	47.8	32.4	41.1
Georgia	41.2	16.7	65.2	Oklahoma	49.0	25.9	55.1
Iowa	45.3	2.9	51.2	Oregon	53.1	66.5	42.3
Idaho	49.8	53.3	58.1	Pennsylvania	52.6	31.9	37.9
Illinois	47.6	26.2	42.0	Rhode Island	54.1	27.8	28.6
Indiana	45.3	31.6	48.3	South Carolina	42.7	31.1	80.0
Kansas	49.0	18.8	55.6	South Dakota	42.2	6.5	52.5
Kentucky	45.2	15.1	51.1	Tennessee	44.9	19.5	52.4
Louisiana	43.5	19.5	69.4	Texas	46.0	48.5	66.8
Massachusetts	54.5	24.8	26.9	Utah	52.8	56.9	35.3
Maryland	46.9	22.1	48.9	Virginia	44.7	34.5	56.1
Maine	50.3	19.1	38.4	Vermont	47.3	21.9	62.6
Michigan	45.3	39.1	48.6	Washington	52.4	72.8	39.2
Minnesota	46.8	23.9	47.5	Wisconsin	43.3	27.3	44.4
Missouri	45.5	13.2	48.2	West Virginia	48.4	27.3	47.5
Mississippi	43.7	2.2	73.0	Wyoming	48.9	36.2	39.6
Montana	49.4	10.1	44.2				

employment (F) increased by 2.62 percent and female wages (w_F) fell by 2.58 percent. Phrased differently, a historical shock that reduced the female wage by 2.58 percent increased female employment by 2.62. Therefore, the labor demand elasticity is given by the ratio of these two numbers, or

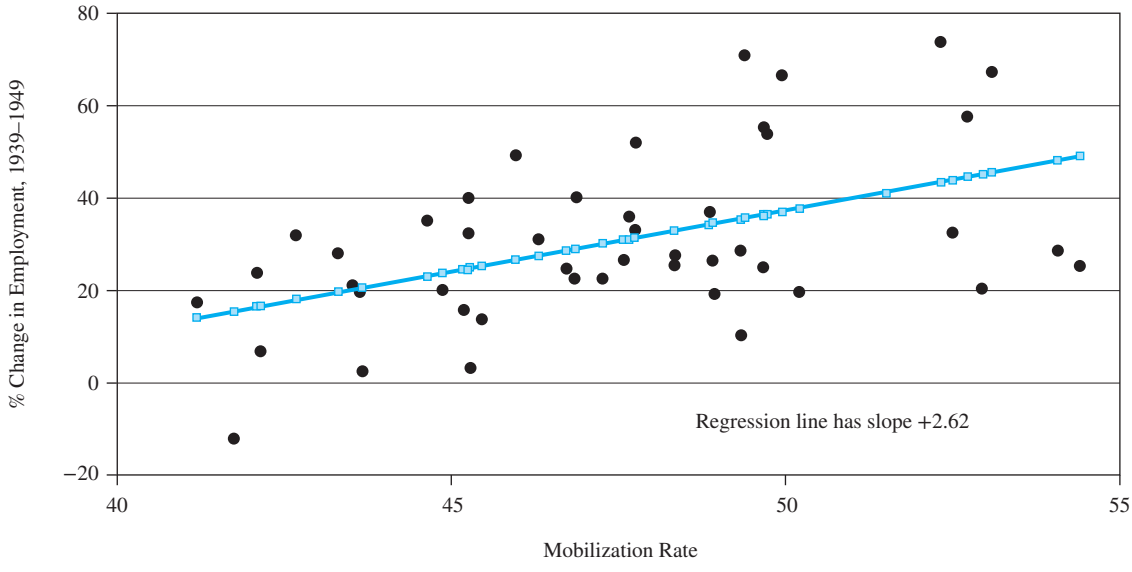
$$\delta = \frac{\%\,\Delta F}{\%\,\Delta w_F} = \frac{2.62}{-2.58} = -1.07 \qquad \textbf{(3-23)}$$

The trend in female wages and employment during World War II suggests that the labor demand elasticity for women is around −1.0.

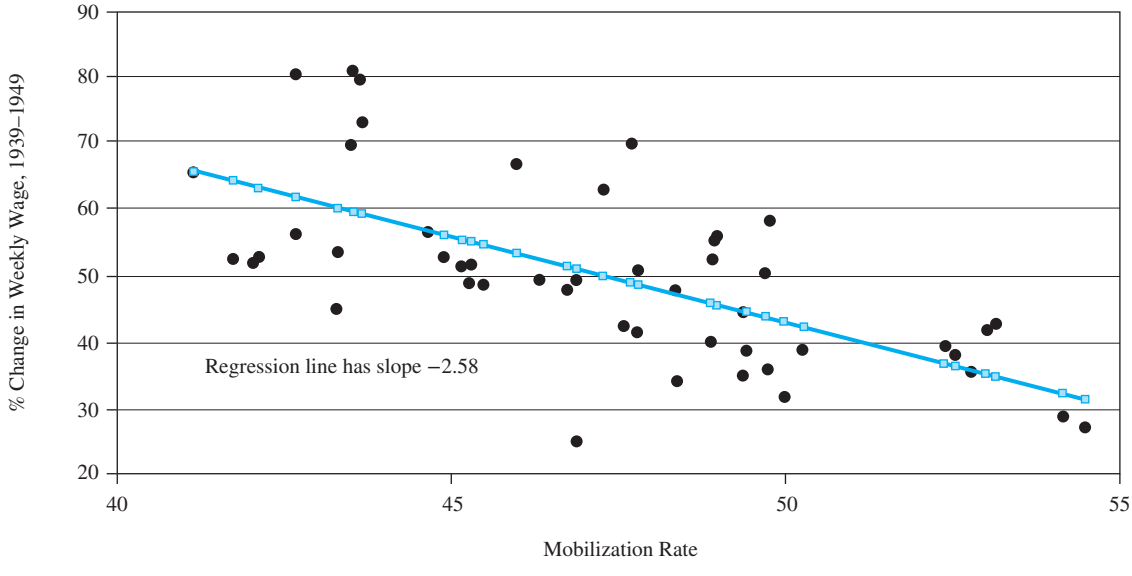
The approach summarized visually in Figure 3-17*a* and *b* can be expanded to control for other factors that might shift the labor supply or labor demand curves differently in different states, such as the educational attainment or age distribution of women. Although this multivariate approach cannot be easily graphed, the method of instrumental variables

FIGURE 3-17 The Impact of Wartime Mobilization of Men on Female Labor Supply and Wages

(*a*) Mobilization rate and changes in female employment, by state.



(*b*) Mobilization rate and changes in female wages, by state.



relies on the same basic logic: The availability of an instrument that shifts only the labor supply curve lets us use the resulting data on wages and employment to trace out the labor demand curve.

The discussion also shows the main weakness of using instrumental variables: The legitimacy of the entire exercise hinges on finding a *valid* instrument, a variable that shifts only one of the curves in the supply–demand framework. Much of the disagreement over the interpretation of empirical evidence in labor economics revolves around whether the researcher is using a valid instrument that permits the identification of either the labor supply or the labor demand curve. The ratio in equation (3-25) is a labor demand elasticity only if interstate differences in the mobilization rate generated interstate differences in female labor supply *but did not generate interstate differences in female labor demand.* Because the labor demand curve is given by the value of marginal product curve, the mobilization rate is a valid instrument only if it is uncorrelated with both interstate differences in the price level and interstate differences in female productivity.

3-10 Policy Application: The Minimum Wage

The U.S. federal government introduced mandatory minimum wages in 1938 as one of the provisions of the Fair Labor Standards Act (FLSA). The nominal minimum wage was initially set at 25 cents an hour, and only 43 percent of nonsupervisory workers were covered by the legislation. Workers in such industries as agriculture and intrastate retail services were exempt. As Figure 3-18 shows, the nominal minimum wage has been adjusted at irregular intervals in the past six decades. It now stands at $7.25 an hour. The coverage of the minimum wage also has been greatly expanded. Most workers who are not employed by state or local governments are now covered by the legislation.

Figure 3-18 shows an important characteristic of the federal minimum wage in the United States: It is not indexed to inflation or productivity growth. As a result, the *real* minimum wage declines between the time that the nominal floor is set and the next time that Congress raises it. For instance, the minimum wage was set at $3.35 per hour in 1981, or 42 percent of the average wage in manufacturing. In 1989, the nominal minimum wage was still $3.35 per hour, but this wage was only 32 percent of the average wage in manufacturing. The "ratcheting" in the real minimum suggests that the economic impact of minimum wages declines the longer it has been since it was last raised.

Figure 3-19 presents the standard model used to analyze the impact of the minimum wage.[12] The competitive labor market is in equilibrium at wage w^* and employment E^*. The government then imposes a minimum wage of $\overline{w}$. Suppose initially that the minimum wage has universal coverage, so all workers are affected by the legislation. And suppose that the penalties associated with paying less than the minimum wage are sufficiently stiff that employers actually comply with the legislation.

Once the government sets the wage floor at $\overline{w}$, firms move up the labor demand curve and employment falls to $\overline{E}$. As a result of the minimum wage, therefore, some workers ($E^* - \overline{E}$) are displaced from their current jobs and become unemployed. But the higher

[12] The model was first derived in George J. Stigler, "The Economics of Minimum Wage Legislation," *American Economic Review* 36 (June 1946): 358–365.

FIGURE 3-18 Minimum Wages in the United States, 1938–2017

Source: U.S. Bureau of the Census, *Statistical Abstract of the United States,* Washington, DC: Government Printing Office, various issues; U.S. Bureau of the Census, *Historical Statistics of the United States, Colonial Times to 1970,* Washington, DC: Government Printing Office, 1975; and U.S. Bureau of Labor Statistics, Employment, Hours, and Earnings from the Current Employment Statistics Survey, Washington, DC.

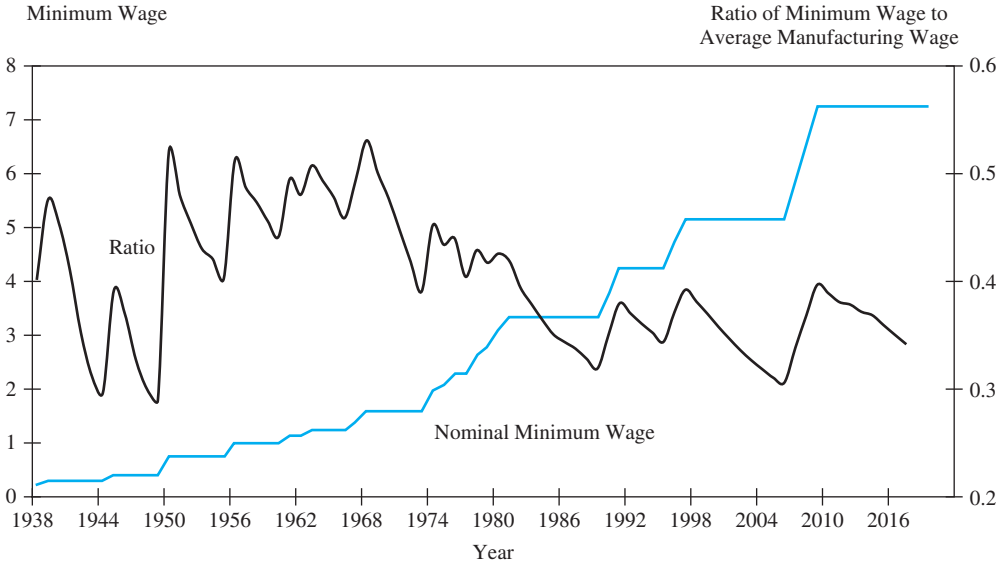
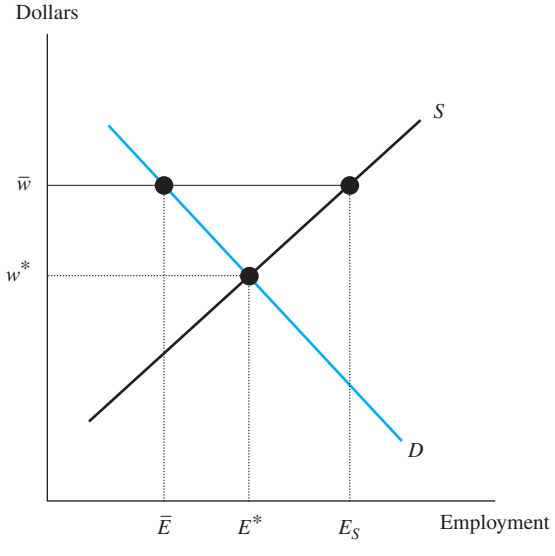


FIGURE 3-19 The Impact of the Minimum Wage on Employment

A minimum wage set at $\overline{w}$ forces employees to cut employment (from E^* to $-\overline{E}$). The higher wage also encourages ($E_S - E^*$) additional workers to enter the market. The minimum wage, therefore, creates unemployment.



wage also encourages some persons to enter the labor market. In fact, E_S workers would now like to be employed. An additional $E_S - E^*$ workers enter the labor market, cannot find jobs, and are added to the unemployment rolls.

Therefore, a minimum wage creates unemployment both because some workers lose their jobs and because some persons who did not find it worthwhile to work at the competitive wage find it worthwhile to work at the higher minimum wage. The unemployment rate, or the ratio of unemployed workers to the size of the labor force, is given by $(E_S - \bar{E})/E_S$. This unemployment persists because firms do not wish to hire more workers and because the unemployed workers want to work at the minimum wage. The unemployment rate clearly depends on the level of the minimum wage, as well as on the elasticities of labor supply and labor demand. It is easy to verify that the unemployment rate is larger the higher the minimum wage and the more elastic the demand and supply curves.

The minimum wage is presumably supposed to raise the income of the least-skilled workers in the economy, for whom the competitive wage would be low. As a result of the minimum wage, however, these workers now become particularly vulnerable to layoffs. The workers who are lucky enough to retain their jobs benefit from the legislation, but the minimum wage provides little consolation to those who lose their jobs.

Compliance with the Minimum Wage Law

This standard model in Figure (3-19) assumes that all firms comply with the legislation. But there seem to be many employers who do not comply. In 2010, for example, when the minimum wage stood at $7.25 an hour, 2.5 million workers (or 3.5 percent of all workers paid by the hour) were paid less than $7.25.[13]

The reason for the high rate of noncompliance is that firms caught breaking the law face only trivial penalties. When a minimum wage violation is detected, the Department of Labor typically attempts to negotiate a settlement between the firm and the affected workers. As part of the settlement, the firm agrees to pay the workers the difference between the minimum wage and the actual wage for the last two years of work. Punitive damages are rare.

In effect, firms that break the law and are caught by the government received an interest-free loan. They delayed paying a portion of their payroll for up to 2 years. And firms that break the law and are not caught, which probably comprise the vast majority of cases, can continue hiring workers at the lower competitive wage. Obviously, the greater the degree of noncompliance, the smaller the employment cut resulting from the minimum wage.

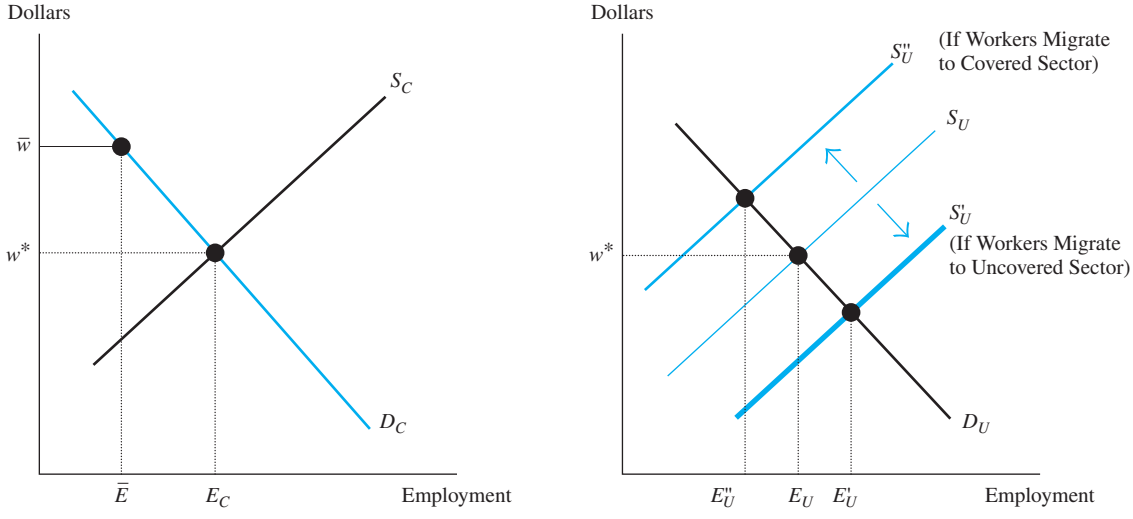
The Covered and Uncovered Sectors

The model in Figure 3-19 also assumed that all workers are covered by the legislation. To see how the adverse employment effects of minimum wages may be moderated by less-than-universal coverage, consider a labor market with two sectors: The covered sector in Figure 3-20*a* and the uncovered sector in Figure 3-20*b*.[14] Prior to the imposition of a

[13] U.S. Bureau of the Census, *Statistical Abstract of the United States, 2008,* Washington, DC: Government Printing Office, 2012, Table 653. See Orley Ashenfelter and Robert S. Smith, "Compliance with the Minimum Wage Law," *Journal of Political Economy* 87 (April 1979): 333–350.

[14] Jacob Mincer, "Unemployment Effects of Minimum Wages," *Journal of Political Economy* 84 (August 1976): S87–S104.

FIGURE 3-20 The Impact of the Minimum Wage on the Covered and Uncovered Sectors
If the minimum wage applies only to jobs in the covered sector, the displaced workers might move to the uncovered sector, shifting the supply curve to the right and reducing the uncovered sector's wage. If it is easy to get a minimum-wage job, workers in the uncovered sector might quit their jobs and wait in the covered sector until a higher-paying job opens up, shifting the supply curve in the uncovered sector to the left and raising the uncovered sector's wage.



minimum wage, there exists a single equilibrium wage, w^*, in both sectors (determined by the intersection of the supply curve S_C and the demand curve D_C in the covered sector, and the intersection of S_U and D_U in the uncovered sector). The minimum wage is imposed only on the industries that comprise the covered sector. Workers employed in the uncovered sector are left to the mercy of the market and receive the competitive wage.

Once the minimum wage is imposed, the wage rises to $\bar{w}$ in the covered sector and some workers lose their jobs. Covered sector employment falls to $\bar{E}$ and there are $(E_C - \bar{E})$ displaced workers. Many of the displaced workers, however, can migrate to the uncovered sector and find work there. If some of this migration takes place, the supply curve in the uncovered sector might shift to S'_U (as illustrated in Figure 3-20*b*). The uncovered sector wage then declines and the number of workers in the uncovered sector increases from E_U to E'_U.

However, this is not the only possible type of migration. After all, some workers initially employed in the uncovered sector might decide that it is worthwhile to quit their low-paying jobs and hang around in the covered sector until a minimum-wage job opens up. If many workers in the uncovered sector take this course of action, the direction of migration might then be from the uncovered to the covered sector. The supply curve in the uncovered sector would shift to S''_U in Figure 3-20*b, raising* the uncovered sector wage.

The analysis in Figure 3-20 shows how the free movement of workers in and out of labor markets can help equilibrate real wages in an economy *despite* the intentions of

policy makers. If workers could freely migrate from one sector to the other, one would expect that the flow would continue as long as workers expected one of the sectors to offer a higher wage. The migration would stop when the *expected* wage was exactly the same in the covered and uncovered sectors.

Let's calculate how much income a worker who enters the covered sector can expect to take home. Let π be the probability that a worker who enters the covered sector gets a job there, so that $1 - \pi$ is the probability that a worker in the covered sector is unemployed. If the worker lands a minimum-wage job, he gets wage $\overline{w}$; if he does not land a job, he has no income (ignoring any unemployment compensation). The wage that a person in the covered sector can expect is then given by

$$\begin{aligned}\text{Expected wage in covered sector} = \\ [\pi \times \overline{w}] + [(1 - \pi) \times 0] = \pi\overline{w}\end{aligned} \qquad \textbf{(3-24)}$$

or a weighted average of the minimum wage $\overline{w}$ and zero.

The worker's alternative is to stay in the uncovered sector. The uncovered sector wage is set by competitive forces and equals w_U. Because there is no unemployment in the uncovered sector, this wage is a "sure thing" for workers in that sector.

Workers move to whichever sector pays the higher expected wage. If the covered sector pays a higher expected wage, the flow of workers to minimum-wage jobs will lower the probability of getting a job and reduce the expected wage. If the wage is higher in the uncovered sector, the migration of workers shifts the supply curve in the uncovered sector outward and lowers the competitive wage w_U. The free migration of workers across sectors should lead to

$$\pi\overline{w} = w_U \qquad \textbf{(3-25)}$$

The expected wage in the covered sector equals the for-sure wage in the uncovered sector.

Factors that influence the probability of landing a minimum-wage job help determine the direction of migration between the two sectors. Suppose workers who get a minimum-wage job keep it for a long time. It is then difficult for a person who has just entered the covered sector to find a job. An unemployed worker will quickly recognize that he is better off working in the uncovered sector where wages are lower, but jobs are available. If the persons who hold minimum-wage jobs are footloose (so that there is a lot of turnover), there is a high chance of getting a minimum-wage job, encouraging many workers to queue up for job openings in the covered sector.

Evidence

The simplest economic model of the minimum wage predicts that as long as the demand curve for labor is downward sloping, an increase in the minimum wage must reduce employment of the affected groups. The size of the employment effect depends on the elasticity of labor demand. Note that an increase in the minimum wage moves the firm up the short-run labor demand curve. Legislated changes in the minimum wage, therefore, generate wage-employment data points that help to identify the labor demand elasticity.

A large literature attempts to determine if, in fact, minimum wages reduce employment. Many of the empirical studies focus on the impact of minimum wages on teenagers,

Theory at Work

THE MINIMUM WAGE AND DRUNK DRIVING

Much of the impact of the minimum wage in the United States falls on teenagers, both in terms of their employment opportunities and in terms of their disposable income. This demographic group, of course, is more likely to be financially dependent on their parents, who typically cover expenses for necessities like rent and food. Many teenagers can then target the increased income resulting from a minimum wage increase to consuming goods that some would consider to be "nonnecessities," including video games, music, movies, and alcohol.

Motor vehicle accidents are the leading cause of death for persons aged 16–20, with nearly one-third of the accidents involving the consumption of alcohol. Although it is illegal for teenagers to purchase or publicly possess alcohol, almost 20 percent of teenagers have, in fact, driven under the influence.

The interplay between the increase in disposable income resulting from a minimum wage increase and the consumption habits of teenagers raises a disturbing scenario. An increase in the minimum wage will increase the disposable income of employed teenagers, which can then lead to increased consumption of nonnecessities, including alcohol. Put bluntly, an increase in the minimum wage might increase the number of alcohol-related traffic fatalities.

A recent study examined state-level increases in the minimum wage to document before-and-after trends in the number of alcohol-related traffic fatalities involving a teenage driver. The figure below shows the trend before and after the enactment of a higher state minimum wage in the sample of states that increased the minimum sometime between 2003 and 2006. Although there is a downward drift in the fatality rate prior to the increase in the state minimum wage, there is a sharp increase afterwards. The estimated impact is not trivial: A 10 percent increase in the state minimum wage (from, say, $10 to $11) is predicted to increase the number of teen-related drunk-driving fatalities by around 8 percent, or roughly an additional 127 fatalities per year. Yet another tradeoff for policy makers to worry about.

Source: Scott Adams, McKinley L. Blackburn, and Chad D. Cotti, "Minimum Wages and Alcohol-Related Traffic Fatalities among Teens," *Review of Economics and Statistics* 83 (August 2012): 828–840.

a group clearly affected by the legislation.[15] In 2010, about 25 percent of teenage workers paid hourly rates earned the minimum wage or less, as compared to only 3.8 percent of workers over the age of 25.[16]

Many of the studies estimate the impact by essentially correlating changes in teenage employment with some measure of the real minimum wage, after adjusting for other variables that could potentially affect teenage employment. These studies often find that the elasticity of teenage employment with respect to the minimum wage is probably between −0.1 and −0.3 so that a 10 percent increase in the minimum wage reduces teenage employment by between 1 and 3 percent.[17]

Beginning in the 1990s, as labor economists began to more fully appreciate that these correlations might not be estimating the labor demand elasticity, researchers began to look at case studies that trace out the impact of specific minimum-wage increases on particular industries or sectors. Surprisingly, some of these studies concluded that a minimum wage increase might not have *any* adverse employment effects.

The best-known case study analyzes the impact of the minimum wage in New Jersey and Pennsylvania.[18] On April 1, 1992, New Jersey increased its minimum wage to $5.05 per hour, the highest minimum wage in the United States, but the neighboring state of Pennsylvania did not follow suit and kept the minimum wage at $4.25, the federally mandated minimum. The New Jersey–Pennsylvania comparison provides a "natural experiment" that can be used to assess the employment impacts of minimum wage legislation.

Suppose, for example, that one contacts a large number of fast-food establishments (such as Wendy's, Burger King, and KFC) on *both* sides of the New Jersey–Pennsylvania state line prior to and after the New Jersey minimum went into effect. The restaurants in Pennsylvania were unaffected by the New Jersey minimum wage, so employment in these restaurants should have changed only because of changes in economic conditions such as seasonal shifts in consumer demand for fried chicken and hamburgers. Employment in restaurants in New Jersey were affected both by the increase in the legislated minimum as well as by changes in economic conditions. By comparing the employment change in

[15] See Alison Wellington, "Effects of the Minimum Wage on the Employment Status of Youths: An Update," *Journal of Human Resources* 26 (Winter 1991): 27–47; Laura Giuliano, "Minimum Wage Effects on Employment, Substitution, and the Teenage Labor Supply: Evidence from Personnel Data," *Journal of Labor Economics* 31 (January 2013): 155–194; and David Neumark, J. M. Ian Salas, and William Wascher, "Revisiting the Minimum Wage-Employment Debate: Throwing Out the Baby with the Bathwater?" *Industrial and Labor Relations Review* 78 (May 2014 Supplement): 608–648.

[16] U.S. Bureau of Labor Statistics, "Characteristics of Minimum Wage Workers: 2010," February 2011.

[17] Charles Brown, "Minimum Wages, Employment, and the Distribution of Income," in Orley C. Ashenfelter and David Card, editors, *Handbook of Labor Economics,* vol. 3B, Amsterdam: Elsevier, 1999, pp. 2101–2163; David Neumark and William Wascher, "Minimum Wages and Employment," *Foundations and Trends in Microeconomics 3* (2007): 1–182;

[18] David Card and Alan B. Krueger, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review* 84 (September 1994): 772–793. Related studies, which also report evidence that the minimum wage does not have an adverse employment effect, include Lawrence F. Katz and Alan B. Krueger, "The Effect of the Minimum Wage on the Fast-Food Industry," *Industrial and Labor Relations Review* 46 (October 1992): 6–2, and David Card, "Do Minimum Wages Reduce Employment? A Case Study of California, 1987–89," *Industrial and Labor Relations Review* 46 (October 1992): 38–54.

TABLE 3-3 The Employment Effect of Minimum Wages in New Jersey and Pennsylvania

Source: David Card and Alan B. Krueger, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review* 84 (September 1994), Table 3.

	Employment in Typical Fast-Food Restaurant (in full-time equivalents)	
	New Jersey	Pennsylvania
Before New Jersey increased the minimum wage	20.4	23.3
After New Jersey increased the minimum wage	21.0	21.2
Difference	0.6	−2.1
Difference-in-differences	**2.7**	

the restaurants on both sides of the state line, one can "net out" the effect of changes in economic conditions and isolate the impact of the minimum wage on employment. In short, we can use the difference-in-differences method to calculate the employment effect of minimum wages.

Table 3-3 summarizes the results of this influential study. The fast-food restaurants on the New Jersey side of the border did not experience a decline in employment relative to the restaurants on the Pennsylvania side of the border. The typical fast-food restaurant in New Jersey hired 0.6 more workers after the minimum wage increase than it did before the increase. At the same time, however, the macroeconomic trends in the fast-food industry led to a decline in employment of about 2.1 workers in neighboring Pennsylvania—a state that was unaffected by the minimum wage increase. The difference-in-differences estimate of the impact of the minimum wage on employment, therefore, was an *increase* of about 2.7 workers in the typical fast-food restaurant. The implied labor demand elasticity is +0.7. Put bluntly, the minimum wage does not reduce employment because the labor demand curve is upward sloping. Needless to say, this line of research, if correct, raises fundamental questions about the validity of the models we use to understand the labor market.[19]

Many studies have tried to examine why this evidence differs so sharply from the earlier time-series results, and why the implications of our simple—and seemingly sensible—supply-and-demand framework seem to be soundly rejected by the data.

One reason may be that the employment effect of the minimum wage is indeed negative, but small. It might then be hard to detect the reduction in employment in a rapidly changing economic environment with very noisy data. Sampling errors could lead researchers to find near-zero or even positive effects.

In fact, it has been documented that the survey data used in the New Jersey–Pennsylvania study had a lot of measurement error and the noise in the data generated correspondingly noisy estimates of the labor demand elasticity.[20] The study used employment data collected

[19] A potential explanation for the upward-sloping demand curve is that fast-food restaurants have some degree of market power when hiring workers, so that the labor market is not competitive. This explanation will be discussed in more detail in the chapter on labor market equilibrium.

[20] David Neumark and William Wascher, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania, Comment," *American Economic Review* 90 (December 2000): 1362–1396.

via a telephone survey. Each of the sampled fast-food restaurants was called in February–March 1992 and in November–December 1992 and asked about the number of employees. If we ranked all the sampled restaurants in New Jersey by their reported change in employment, the restaurant at the 10^{th} percentile reported a decline in employment of 7 workers, while the restaurant at the 90^{th} percentile reported an increase in employment of 11 workers. Given that the typical fast-food establishment only employed about 20 workers, these changes seem remarkably large.

It turns out that the employment changes implied by the telephone survey are way off from what we get if we used the actual administrative employment records kept by the central headquarters of Wendy's, KFC, and Burger King. In the administrative data, the 10^{th} percentile change in employment in the New Jersey establishments was −4 workers and the 90^{th} percentile change was +4 workers. In fact, the estimated elasticity turns negative, with a value of −0.2, if the administrative data is used to measure the employment impact of the minimum wage increase.

There may also be a conceptual problem with measuring the impact of the minimum wage by examining employment trends in fast-food establishments. This sector may not be representative of the low-wage labor market, giving a myopic and misleading picture of the employment effects. For example, fast-food restaurants might use a production technology where the number of workers is relatively fixed (one worker per grill, one worker per cash register, and so on). As a result, the minimum wage might not reduce employment in existing restaurants, but might instead discourage the national chain from opening additional restaurants or accelerate the closing of the marginally profitable ones. Moreover, the aggregate profitability of the entire chain might "shelter" specific fast-food establishments from the minimum wage. The minimum wage would instead accelerate the decline of the smaller and less-competitive "mom-and-pop" restaurants.

The before-and-after comparisons of employment in affected firms are also affected by the *timing* of the comparisons. Employers may not change their employment exactly on the date that the law goes into effect, but may instead adjust their employment slowly as they assess the impact of the increase in labor costs. A study of the impact of minimum wages in Canada documents that the employment effects of the minimum wage are smallest when one looks at employment immediately after the increase in the minimum wage takes effect, and becomes more negative over time.[21]

The Seattle Minimum Wage Debate

Since the 1990s, the "living wage" movement has argued for a minimum wage set at a level where full-time workers can support themselves without relying on social assistance programs. Many American cities have indeed enacted living wage ordinances. These laws typically set local minimum wages far above the federal minimum.

Prior to 2015, the effective minimum wage in the city of Seattle was $9.47 an hour, which was the state-level minimum wage set by the State of Washington. On July 1, 2014, Seattle enacted an ordinance increasing the minimum wage within the city limits to $11 on April 1, 2015, to $13 on January 1, 2016, and to $15 on January 1, 2017. The mandated

[21] See Michael Baker, Dwayne Benjamin, and Shuchita Stanger, "The Highs and Lows of the Minimum Wage Effect: A Time-Series Cross-Section Study of the Canadian Law," *Journal of Labor Economics* 17 (April 1999): 318–350.

wage increases were sizable: A 37.3 percent wage increase between early 2015 and January 1, 2016, and a 58.4 percent increase between early 2015 and January 1, 2017.

The release of a recent study examining the impact of the first two mandated wage increases (to $11 and to $13) rekindled the academic debate over the minimum wage, and raised important questions about the role of empirical research in a very contentious political environment.[22] It can be plausibly argued that this study, conducted by a team of economists at the University of Washington (UW), used some of the best available data that can be used to analyze the problem. Rather than rely on a survey of fast-food restaurants, or on firm-level data in a particular industry, the study used administrative employment data maintained by the State of Washington. These data contain payroll records for *all* workers who received wages in Washington and who were covered by unemployment insurance. Moreover, the administrative data report information not only on what each worker actually got paid, but also on how many workers were employed by a firm and how many hours each of those workers worked.

The UW study used these data to document what happened to the low-wage labor market in Seattle, where the low-wage labor market comprises workers earning below $19 an hour. Table 3-4 summarizes the findings. The table reports the percent change in both the number of low-wage workers in Seattle and in the number of hours worked, relative to what happened in a control group.

The study employed a newly developed statistical technique, called the **synthetic control method,** to construct the control group.[23] This method aggregates all other cities in Washington State in a way that best resemble what Seattle looked like prior to the imposition of the higher minimum wage. For example, the city of Seattle may best resemble some combination of Olympia, Spokane, and Takoma. The algorithm underlying the synthetic control method essentially searches through all possible combinations across all cities and finds the mix that best matches pre-2015 Seattle. The analysis then uses a standard difference-in-differences framework, reporting what happened in Seattle before and after the minimum wage increase relative to what happened in the synthetic control.

It is evident that Seattle's low-wage labor market reacted to the minimum wage, particularly after it went up to $13: the number of hours worked by low-wage workers fell by about 10 percent and the number of low-wage jobs fell by 5 percent.

And this is where things get interesting. Exactly one week *before* the UW study was released publicly (although it had been circulated privately for some time before that), a research team at the University of California, Berkeley released its own study of the Seattle experience.[24] In contrast to the UW study, the Berkeley team gave a far more glowing appraisal of the Seattle minimum wage. The Berkeley team, however, did not use the

[22] Ekaterina Jardim, Mark C. Long, Robert Plotnick, Emma van Inwegen, Jacob Vigdor, and Hilary Wethin, "Minimum Wage Increases, Wages, and Low-Wage Employment: Evidence from Seattle," National Bureau of Economic Research Working Paper No. 23532, June 2017.

[23] The synthetic control method was developed by Alberto Abadie, Alexis Diamond, and Jens Hainmueller, "Synthetic control methods for comparative case studies: Estimating the effect of California's Tobacco Control Program," *Journal of the American Statistical Association* 105 (June 2010): 493–505.

[24] Michael Reich, Sylvia Allegretto, and Anna Godoey, "Seattle's Minimum Wage Experience 2015–16," Center on Wage and Employment Dynamics, University of California, Berkeley, June 2017.

TABLE 3-4 Difference-in-Differences Impact of the Seattle Minimum Wage: Percent Change Relative to Conditions in Mid-2014

Source: Adapted from Ekaterina Jardim, Mark C. Long, Robert Plotnick, Emma van Inwegen, Jacob Vigdor, and Hilary Wethin, "Minimum Wage Increases, Wages, and Low-Wage Employment: Evidence from Seattle," National Bureau of Economic Research Working Paper No. 23532, June 2017, Table 5.

Year-Quarter	Hours Worked	Number of Jobs
Ordinance enacted on July 1, 2014		
2014.3	+0.8	+0.4
2014.4	+0.3	−1.0
2015.1	−2.3	0.0
Minimum wage increased to $11 on April 1, 2015		
2015.2	−1.3	−1.4
2015.3	−3.4	−1.9
2015.4	−2.1	−4.5
Minimum wage increased to $13 on January 1, 2016		
2016.1	−10.6	−5.1
2016.2	−8.7	−5.2
2016.3	−10.2	−6.3

administrative individual-level data for all low-wage workers in the State of Washington. Instead, they used county- and city-level data, and focused only on employment in the food services industry. The Berkeley study found no employment effects despite the large increase in the minimum wage.

The conflicting evidence led to a public "battle of the experts" in a highly politicized media environment. Through a Freedom of Information Act request, an enterprising reporter for the *Seattle Weekly* obtained the internal emails exchanged between the Seattle city government and the Berkeley team, and the picture that these emails paint of "weaponized research" is disturbing.

The *Seattle Weekly* summarizes the timeline: "The UW shares with City Hall an early draft of its study showing the minimum wage law is hurting the workers it was meant to help; the mayor's office shares the study with researchers known to be sympathetic toward minimum wage laws, asking for feedback; those researchers release a report that's high on Seattle's minimum wage law just a week before the negative report comes out."[25]

The internal emails show coordination between the staff of the mayor's office in Seattle, who strongly supported the minimum wage increase, and the Berkeley researchers. The staff, for example, advised the Berkeley team not to mention the UW study when making the public announcement of their findings: "Tomorrow's release will just highlight your study, correct?. . . Don't want your positive news to serve as a teaser for the UW study." And, indeed, the press release of the Berkeley findings made no mention whatsoever of the UW study.

[25] Daniel, Person, "The City Knew the Bad Minimum Wage Report Was Coming Out, So It Called Up Berkeley," *Seattle Weekly,* June 26, 2017. See also Person, Daniel, "Inside the Minimum Wage Data War," *Seattle Weekly,* July 26, 2017; and Person, Daniel, "Emails Show Mayor's Office and Berkeley Economist Coordinated Release of Favorable Minimum Wage Study," *Seattle Weekly,* July 26, 2017.

So what do we learn? The debate over the employment impact of the minimum wage is, at its core, a debate over the value of the elasticity of labor demand. Those who argue that the minimum wage has no employment effects are really only arguing that the labor demand elasticity is zero. But that technical question happens to be intimately linked to a heavily politicized issue. There is demand by all sides of the political debate for research studies that report an elasticity fitting their policy priors.

The conflicting findings of the UW and Berkeley teams teach a very important lesson. Empirical research in labor economics is "elastic." There are different data sources we can use; different ways of looking at the data; different sample selections. More often than not, these choices influence the bottom line.

Most of the time, the difference in the results is not sufficiently interesting to attract attention by anyone outside the small club of economists who follow the latest twists and turns of a particular issue. But sometimes, as in the minimum wage debate, the researcher's choices spill over into the public arena. It then becomes crucial to carefully dissect what researchers actually do. What methodology was used? How were the data manipulated? Which sample was chosen and why? The dissection would provide the information that might help determine which set of conflicting results seems more sensible.

Summary

- In the short run, a profit-maximizing firm hires workers up to the point where the wage equals the value of marginal product of labor.
- In the long run, a profit-maximizing firm hires each input up to the point where the price of the input equals the value of marginal product of the input. This condition implies that the optimal input mix is one in which the ratio of marginal products of labor and capital equals the ratio of input prices.
- In the long run, a decrease in the wage generates both substitution and scale effects. Both of these effects spur the firm to hire more workers.
- Both the short-run and long-run demand curves for labor are downward sloping, but the long-run demand curve is more elastic than the short-run curve.
- The short-run labor demand elasticity may be on the order of -0.4 to -0.5. The long-run elasticity is on the order of -1.
- Capital and skilled workers are complements in the sense that an increase in the price of capital reduces the demand for skilled workers. Capital and unskilled workers are substitutes in the sense that an increase in the price of capital increases the demand for unskilled workers.
- An instrument is a variable that shifts either the supply or the demand curve. The variation caused by this shock can be used to estimate the labor demand or labor supply elasticity.
- The imposition of a minimum wage on a competitive labor market creates unemployment because some workers are displaced from their jobs and because new workers enter the labor market hoping to find one of the high-paying (but scarce) jobs.
- The elasticity of teenage employment with respect to the minimum wage may be on the order of -0.1 to -0.3.

Key Concepts

average product, *79*
capital-skill complementarity hypothesis, *100*
cross-elasticity of factor demand, *99*
demand curve for labor, *81*
elasticity of labor demand, *83*
elasticity of substitution, *96*
equilibrium, *100*
instrument, *102*
instrumental variable, *102*
isocost, *87*
isoquant, *85*
law of diminishing returns, *79*
marginal cost, *84*
marginal product of capital, *77*
marginal product of labor, *77*
marginal productivity condition, *84*
marginal rate of technical substitution, *86*
marginal revenue, *84*
Marshall's rules of derived demand, *96*
method of instrumental variables, *102*
perfect complements, *95*
perfect substitutes, *95*
perfectly competitive firm, *79*
production function, *77*
scale effect, *92*
substitution effect, *92*
synthetic control method, *115*
value of average product, *80*
value of marginal product, *79*

Review Questions

1. Why does a profit-maximizing firm hire workers up to the point where the wage equals the value of marginal product? Show that this condition is identical to the one that requires a profit-maximizing firm to produce the level of output where the price of the output equals the marginal cost of production.
2. Why is the short-run demand curve for labor downward sloping?
3. What mix of inputs should be used to produce a given level of output?
4. Suppose the firm is hiring labor and capital and that the ratio of marginal products of the two inputs equals the ratio of input prices. Does this imply that the firm is maximizing profits? Why or why not?
5. Suppose the wage increases. Show that in the long run the firm will hire fewer workers. Decompose the employment change into substitution and scale effects.
6. What factors determine the elasticity of the industry's labor demand curve?
7. What is the capital-skill complementarity hypothesis?
8. Explain how and why the method of instrumental variables allows us to estimate the labor demand elasticity.
9. Show how the minimum wage creates unemployment in a competitive market.
10. Discuss the impact of the minimum wage when there are two sectors in the economy: the covered sector (which is subject to the minimum wage) and the uncovered sector (which is not).
11. Summarize the evidence regarding the impact of the minimum wage on employment.

Problems

3-1. Suppose there are two inputs in the production function, labor and capital, and these two inputs are perfect substitutes. The existing technology permits one machine to do the work of three workers. The firm wants to produce 100 units of output. Suppose the price of capital is $750 per machine per week. What combination of inputs will the firm use if the weekly salary of each worker is $300? What combination of inputs

will the firm use if the weekly salary of each worker is $225? What is the elasticity of labor demand as the wage falls from $300 to $225?

3-2. Figure 3-18 in the text shows the ratio of the federal minimum wage to the average hourly manufacturing wage.
   (a) Describe how this ratio has changed from the 1950s to the 1990s. What might have caused this apparent shift in fundamental economic behavior in the United States?
   (b) This ratio fell steadily from 1968 to 1974 and again from 1980 to 1990, but the underlying dynamics of the minimum wage and the average manufacturing wage were different during the two time periods. Explain.
   (c) What has been happening to the ratio of the federal minimum wage (nominal) to the average hourly manufacturing wage from 1990 to today?

3-3. Firm would hire 20,000 workers if the wage rate is $12 but will hire 10,000 workers if the wage rate is $15. Firm B will hire 30,000 workers if the wage is $20 but will hire 33,000 workers if the wage is $15. The workers in which firm are more likely to organize and form a union?

3-4. Consider a firm for which production depends on two normal inputs, labor and capital, with prices w and r, respectively. Initially the firm faces market prices of $w = 6$ and $r = 4$. These prices then shift to $w = 4$ and $r = 2$.
   (a) In which direction will the substitution effect change the firm's employment and capital stock?
   (b) In which direction will the scale effect change the firm's employment and capital stock?
   (c) Can we say conclusively whether the firm will use more or less labor? More or less capital?

3-5. What happens to employment in a competitive firm that experiences a technology shock such that at every level of employment its output is 200 units/hour greater than before?

3-6. Consider each of the following and explain why it is or is not a valid instrument for estimating the labor supply elasticity and/or labor demand elasticity in the United States. (1) Variation in state income tax rates. (2) Variation in state corporate tax rates. (3) Changes in federal income tax rates over time.

3-7. Suppose a firm purchases labor in a competitive labor market and sells its product in a competitive product market. The firm's elasticity of demand for labor is −0.4. Suppose the wage increases by 5 percent. What will happen to the amount of labor hired by the firm? What will happen to the marginal productivity of the last worker hired by the firm?

3-8. A firm's technology requires it to combine 5 person-hours of labor with 3 machine-hours to produce 1 unit of output. The firm has 15 machines in place and the wage rate rises from $10 per hour to $20 per hour. What is the firm's short-run elasticity of labor demand?

3-9. In a particular industry, labor supply is $E_S = 10 + w$ and labor demand is $E_D = 40 - 4w$, where E is the level of employment and w is the hourly wage.
   (a) What is the equilibrium wage and employment if the labor market is competitive? What is the unemployment rate?

(b) Suppose the government sets a minimum hourly wage of $8. How many workers would lose their jobs? How many additional workers would want a job at the minimum wage? What is the unemployment rate?

3-10. Suppose the hourly wage is $10 and the price of each unit of capital is $25. The price of output is constant at $50 per unit. The production function is

$$f(E, K) = E^{1/2} K^{1/2}$$

so that the marginal product of labor is

$$MP_E = (1/2)(K/E)^{1/2}$$

If the current capital stock is fixed at 1,600 units, how much labor should the firm employ in the short run? How much profit will the firm earn?

3-11. Several states set their own minimum hourly wage above the federal minimum wage. To offset higher minimum wages, many of these states offer firms tax incentives that lower the cost of borrowing and/or lower the firm's tax liability on profits. In general, how do these kinds of state policies (that is, higher minimum wages and lower taxes) distort the firm's profit-maximization decisions? Why might we expect to see such policies attract firms in "high tech" industries?

3-12. How does the amount of unemployment created by an increase in the minimum wage depend on the elasticity of labor demand? Do you think an increase in the minimum wage will have a greater unemployment effect in the fast food industry or in the lawn care/landscaping industry?

3-13. Which one of Marshall's rules suggests why labor demand should be relatively inelastic for public school teachers and nurses? Explain.

3-14. Many large cities have recently enacted living wage ordinances that require paying a minimum wage that is higher than the state or federal minimum wage. Moreover, sometimes living wage ordinances state two different minimum wages – one for workers who also receive employer-paid health insurance and one for workers who do not receive health insurance.

(a) Why would living wages distinguish between workers based on their health insurance? In particular, what "problem" might the local government be trying to solve?

(b) Sometimes living wage ordinances apply only to the city government, meaning that the city is required to pay all city workers a high minimum wage while private firms are only subject to state or federal minimum wages. In this case, the living wage creates a covered sector and an uncovered sector. Which workers are in the covered sector? Which workers are in the uncovered sector? What might city officials, who have to manage to a budget, do in response to a living wage ordinance that only applies to city workers?

3-15. Consider a production model with two inputs–domestic labor (E_{Dom}) and foreign labor (E_{For}). The market is originally in equilibrium in that

$$\frac{MP_{Edom}}{w_{dom}} = \frac{MP_{Efor}}{w_{for}}.$$

(a) Suppose a shock occurs that increases the marginal product of foreign labor. Assuming no changes in domestic or foreign wages, explain what will happen to domestic and foreign labor in order to restore the above condition.

(b) In the years following the shock, what are three (significantly different) policies that the domestic country could employ if it wanted to reverse the outflow of labor?

Selected Readings

Daron Acemoglu, David H. Autor, and David Lyle, "Women, War and Wages: The Effect of Female Labor Supply on the Wage Structure at Midcentury," *Journal of Political Economy* 112 (June 2004): 497–551.

David H. Autor and David Dorn, "The Growth of Low-Skill Service Jobs and the Polarization of the U.S. Labor Market," *American Economic Review* 103 (August 2013): 1553–1597.

David Card and Alan B. Krueger, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review* 84 (September 1994): 772–793.

Bruno Crépon and Francis Kramarz, "Employed 40 Hours or Not Employed 39: Lessons from the 1982 Mandatory Reduction of the Workweek," *Journal of Political Economy* 110 (December 2002): 1355–1389.

Claudia Goldin and Lawrence F. Katz, "The Origins of Technology-Skill Complementarity," *Quarterly Journal of Economics* 113 (August 1998): 693–732.

Daniel S. Hamermesh and Stephen J. Trejo, "The Demand for Hours of Labor: Direct Estimates from California," *Review of Economics and Statistics* 82 (February 2000): 38–47.

Andrea Ichino and Regina T. Riphahn, "The Effect of Employment Protection on Worker Effort: Absenteeism during and after Probation," *Journal of the European Economic Association* 3 (March 2005): 120–143.

Ekaterina Jardim, Mark C. Long, Robert Plotnick, Emma van Inwegen, Jacob Vigdor, and Hilary Wethin, "Minimum Wage Increases, Wages, and Low-Wage Employment: Evidence from Seattle," National Bureau of Economic Research Working Paper No. 23532, June 2017.

David Neumark and William Wascher, "Minimum Wages and Employment," *Foundations and Trends in Microeconomics* 3 (2007): 1–182.

Chapter 4

Labor Market Equilibrium

Order is not pressure which is imposed on society from without, but an equilibrium which is set up from within.

—José Ortega y Gasset

Workers prefer to work when the wage is high, but firms prefer to hire when the wage is low. A labor market equilibrium resolves the conflicting desires of workers and firms, and determines the wage and employment levels. By understanding how an equilibrium is reached, we can address what is perhaps the most interesting question in labor economics: Why do wages and employment go up and down?

This chapter analyzes the properties of equilibrium. We will see that if markets are competitive and if firms and workers are free to enter and leave these markets, the equilibrium allocation of workers to firms is efficient; it maximizes the total gains that workers and firms accumulate by trading with each other.

This result is an example of Adam Smith's **invisible hand theorem,** wherein labor market participants in search of their own selfish goals attain an outcome that no one consciously sought to achieve. The prediction that competitive labor markets are efficient plays an important role in the framing of public policy. The evaluation of many government programs often focuses on whether the particular policy leads to a more efficient allocation of resources or whether the efficiency costs are substantial.

We also examine the properties of labor market equilibrium under alternative market structures, such as a monopsony (where there is only one buyer of labor). Each market structure generates an equilibrium with its own unique features. Monopsonists, for instance, hire fewer workers and pay less than competitive firms.

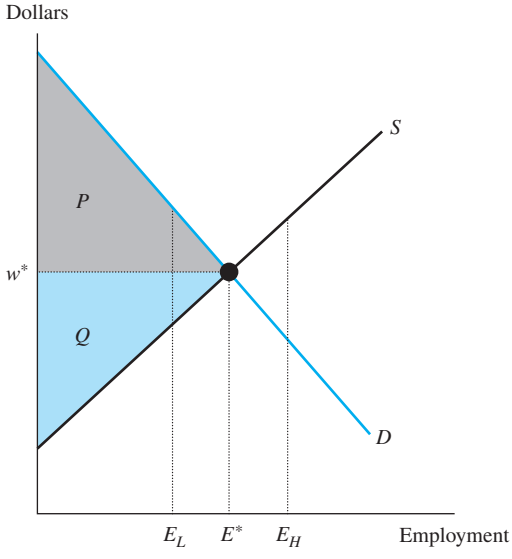
Finally, the chapter uses a number of government policies—including taxes, subsidies, and immigration—to illustrate how these interventions move the labor market to a different equilibrium, altering economic opportunities for both workers and firms.

4-1 Equilibrium in a Single Labor Market

We have already seen how a competitive labor market reaches an equilibrium. We now provide a more detailed discussion of the properties of this equilibrium.

Figure 4-1 illustrates the familiar intersection of labor supply (S) and labor demand (D) curves. The supply curve gives the total number of employee-hours that agents in the

FIGURE 4-1 Equilibrium in a Competitive Labor Market
The labor market is in equilibrium when supply equals demand; E^* workers are employed at a wage of w^*. The triangle P gives the producer surplus; the triangle Q gives the worker surplus. A competitive market maximizes the gains from trade, or the sum $P + Q$.



economy allocate to the market at any given wage; the demand curve gives the total number of employee-hours that firms will want to hire at that wage. Equilibrium occurs when supply equals demand, generating the competitive wage w^* and employment E^*.

Once the competitive wage is established, each firm in the industry hires workers up to the point where that wage equals the value of marginal product of labor. The total number of workers hired by all the firms in the industry must equal the equilibrium employment E^*.

As Figure 4-1 shows, there is no unemployment in a competitive labor market. At the market wage w^*, the number of persons who want to work equals the number of workers firms want to hire. Persons who are not working are also not looking for work *at the going wage*. Of course, many of these persons would enter the labor market if the wage rose (and many workers would withdraw if the wage fell).

Real-world labor markets are continually subjected to many shocks that shift both the supply and demand curves. It is unlikely, therefore, that the labor market ever reaches a stable equilibrium—with wages and employment remaining constant for a very long time. Nevertheless, the concept of an equilibrium is useful because it helps us understand how wages might react to specific economic or political events. As the labor market responds to a specific shock, wages and employment would tend to move toward the new equilibrium level.

Efficiency

Figure 4-1 also shows the benefits that accrue to the aggregate economy as workers and firms exchange offers in the labor market. The total revenue accruing to the firm can be calculated by adding up the value of marginal product of the first worker, the second

Theory at Work

THE INTIFADAH AND PALESTINIAN WAGES

Throughout much of the 1980s, nearly 110,000 Palestinians who resided in the occupied West Bank and Gaza Strip commuted to Israel for their jobs. Many of them were employed in the construction or agriculture industries.

As a result of the Intifadah that began in 1988—the Palestinian uprising against Israeli control of the West Bank and Gaza territories—there were major disruptions in the flow of those workers into Israel. Israeli authorities, for instance, stepped up spot checks of work permits and began to enforce the ban on Palestinians spending the night in Israel, while strikes and curfews in the occupied territories limited the mobility of commuting workers.

Within one year, the daily absenteeism rate jumped from less than 2 percent to more than 30 percent; the average number of work days in a month dropped from 22 to 17 days; and the length of time it took a commuting Palestinian to reach the work location rose from 30 minutes to three or four hours.

The Intifadah, therefore, greatly reduced the supply of Palestinian commuters in Israel. The supply and demand framework suggests that the uprising should have increased the equilibrium wage of the Palestinian commuters. And that's exactly what happened. The roughly 50 percent cut in the labor supply of Palestinian commuters increased their real wage by about 50 percent, implying that the labor demand elasticity for Palestinian commuters is about −1.0.

Source: Joshua D. Angrist, "Short-Run Demand for Palestinian Labor," *Journal of Labor Economics* 14 (July 1996): 425–453.

worker, and all workers up to E^*. Because the labor demand curve gives the value of marginal product, it must be the case that the area under the demand curve gives the value of total product.[1] Each worker receives a wage of w^*. Hence, the profits accruing to firms, which we call **producer surplus,** are given by the area of the triangle P.

Workers also gain. The supply curve gives the wage required to bribe additional workers into the labor market. In effect, the height of the supply curve at a given point measures the value of the marginal worker's time in alternative uses. The difference between what the worker receives (that is, the competitive wage w^*) and the value of the worker's time outside the labor market gives the gains accruing to workers. This quantity is called **worker surplus** and is given by the area of the triangle Q.

The aggregate, the **gains from trade,** accruing to the economy is given by the sum of producer surplus and worker surplus, or the area $P + Q$. *The competitive market maximizes the total gains from trade accruing to the economy*. To see why, consider what the gains would be if firms hired more than E^* workers, say E_H. The "excess" workers have a value of marginal product that is less than their value of time elsewhere. These workers are not being efficiently used by the labor market; they are better off elsewhere. Similarly, consider what would happen if firms hired too few workers, say E_L. The "missing" workers have a value of marginal product that exceeds their value of time elsewhere, and their resources would be more efficiently used if they worked.

An allocation of persons to firms that maximizes the total gains from trade in the labor market is called an **efficient allocation.** A competitive equilibrium generates an efficient allocation of labor resources.

[1] To simplify the discussion, assume that labor is the only factor in the production function.

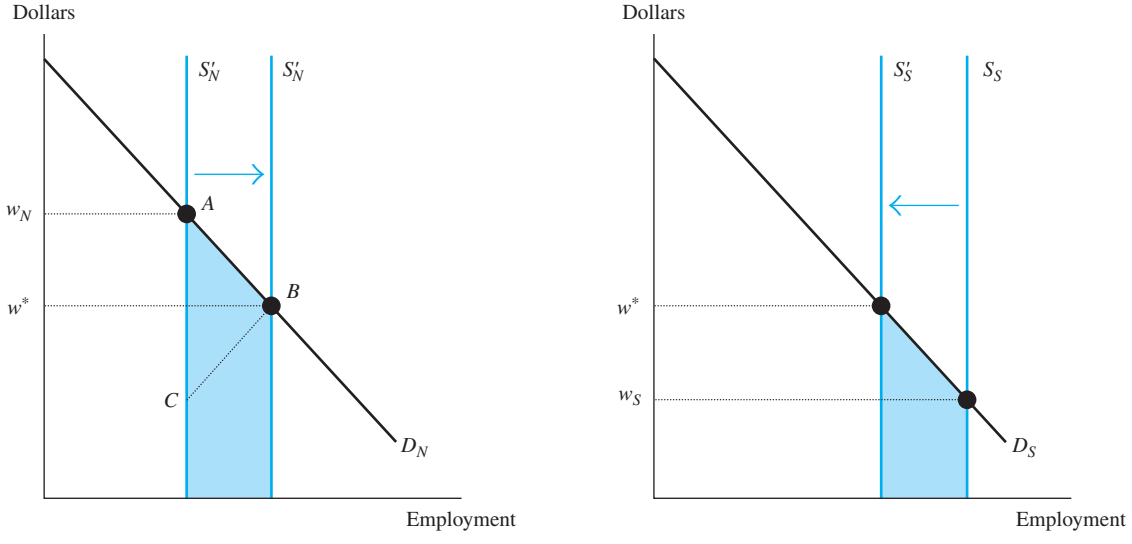
4-2 Equilibrium across Labor Markets

The discussion in the previous section examined the properties of equilibrium in a *single* competitive labor market. The economy, however, typically consists of many labor markets, even for workers who have similar skills. These labor markets might be differentiated by region (so that we can talk about the labor market in the Northeast and the labor market in California) or by industry (the labor market for production workers in the automobile industry and the labor market for production workers in the steel industry).

Suppose there are two regional labor markets, the North and the South. The two markets employ workers of similar skills; workers in the North are perfect substitutes for workers in the South. Figure 4-2 illustrates the respective labor supply and labor demand curves (S_N and D_N in the North, and S_S and D_S in the South). For simplicity, the supply curves are vertical lines, implying that supply is perfectly inelastic within each region. As drawn, the equilibrium wage in the North, w_N, exceeds the equilibrium wage in the South, w_S.

Can the wage differential between the two regions persist and represent a true competitive equilibrium? No, because the wage gap encourages southern workers to pack up and move north, where they can earn higher wages and presumably attain a higher level of utility. Employers in the North also see the wage differential and realize that they can do better

FIGURE 4-2 Competitive Equilibrium in Two Labor Markets Linked by Migration
The wage in the northern region (w_N) exceeds the wage in the southern region (w_S). Southern workers want to move north, shifting the southern supply curve to the left and the northern supply curve to the right. In the end, wages are equated across regions (at w^*). The migration reduces the value of output in the South by the size of the shaded trapezoid in the southern labor market and increases the value in the North by the size of the larger shaded trapezoid in the northern labor market. Migration increases the value of aggregate output by the triangle *ABC*.



(*a*) The Northern Labor Market

(*b*) The Southern Labor Market

by moving to the South. After all, workers are equally skilled in the two regions, and firms can make more money by hiring cheaper labor.

If workers can move across regions freely, the migration flow will shift the supply curves in both regions. The supply curve in the South would shift to the left (to S_S') as southern workers leave the region, raising the southern wage. The supply curve would shift to the right in the North (to S_N') as the southerners arrived, depressing the northern wage. If there were free entry and exit of workers in and out of labor markets, the national economy would eventually be characterized by a single equilibrium wage, w^*.

Note that wages across the two labor markets would also be equalized if firms (instead of workers) could freely enter and exit labor markets. If northern firms close their plants and move to the South, the demand curve for northern labor shifts to the left and lowers the northern wage and the demand curve for southern labor shifts to the right, raising the southern wage. The firms' incentives to move across markets evaporate once the regional wage differential disappears. As long as either workers or firms are free to enter and exit labor markets, therefore, a competitive economy will be characterized by a single equilibrium wage.

Efficiency Revisited

The "single wage" property of a competitive equilibrium has important implications for economic efficiency. Recall that the wage equals the value of marginal product of labor in a competitive market. As firms and workers move to the region that provides the best opportunities, they eliminate regional wage differentials. Therefore, workers of given skills have the same value of marginal product of labor in all markets.

The single wage property implies an efficient allocation of labor resources across markets. To see why, suppose that a benevolent dictator takes over the economy and that this dictator has the power to dispatch workers across regions. In making allocation decisions, this benevolent dictator has one overriding objective: to allocate workers to those places where they are most productive. When the dictator first takes over, he faces the initial situation illustrated in Figure 4-2, where the wage in the North is higher than in the South. This wage gap implies that the value of marginal product of labor is greater in the North than in the South.

The dictator picks a southern worker at random. What should he do with this worker? Because the dictator wants to place this worker where he is most productive, the worker is dispatched to the North. In fact, the dictator will keep dispatching workers to the North as long as the value of marginal product of labor is greater in the North than in the South. The law of diminishing returns implies that as the dictator forces more and more people to work in the North, the value of marginal product of northern workers declines and the value of marginal product of southern workers rises. The dictator will stop reallocating persons when the value of marginal product is the same in both markets. In short, the free mobility of workers from one region to another duplicates the efficient equilibrium that the dictator imposed.

We can prove that the single wage property is efficient by calculating the aggregate value of output in Figure 4-2. The value of output in a particular labor market equals the area under the demand curve. The migration of workers out of the South reduces the total value of output in the South by the shaded area of the trapezoid in the southern labor market. The migration of workers into the North increases the total value of output in the

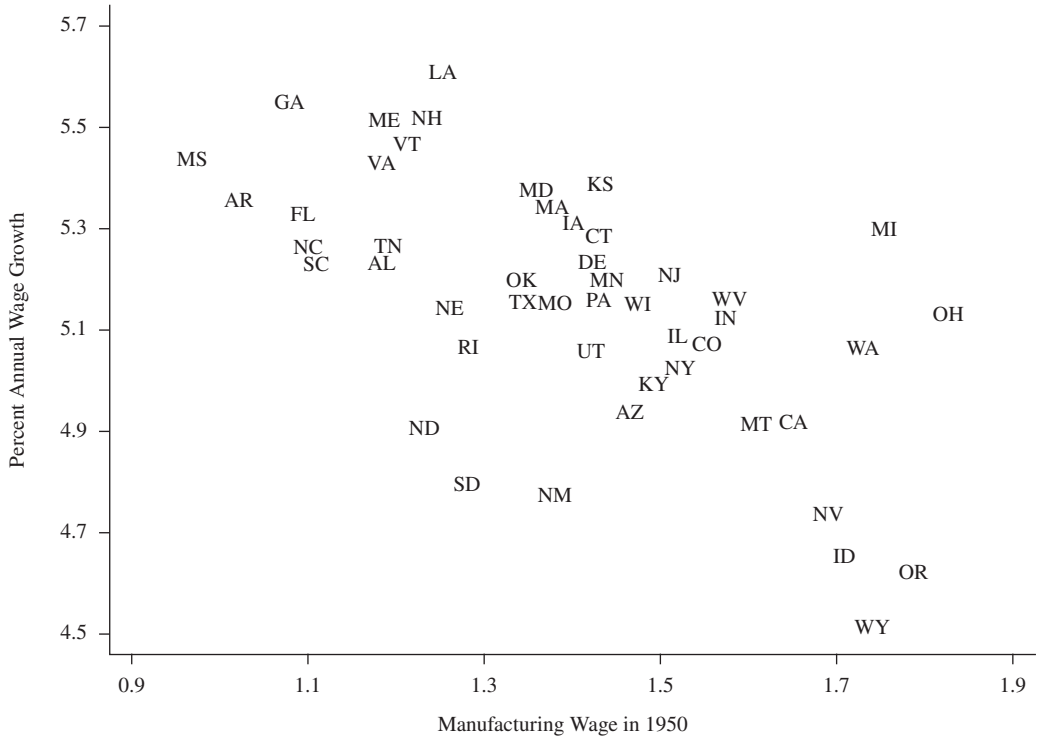
North by the shaded area of the trapezoid in the northern labor market. A comparison of the two trapezoids reveals that the area of the northern trapezoid exceeds the area of the southern trapezoid by the size of the triangle *ABC*, implying that the total value of output in the national economy increases as a result of worker migration.

The surprising implication of our analysis should be clear: *Through an "invisible hand," workers and firms that search selfishly for better opportunities accomplish a goal that no one in the economy had in mind: an efficient allocation of resources that maximizes aggregate output*.

Many studies suggest that regional wage differences in the United States (as well as in other countries) indeed narrow over time, as implied by our analysis.[2] Figure 4-3 illustrates the extent of wage convergence across states in the United States. The figure relates the

FIGURE 4-3 Wage Convergence across States, 1950–1990

Source: Olivier Jean Blanchard and Lawrence F. Katz, "Regional Evolutions," *Brookings Papers on Economic Activity* 1 (1992): 1–61.



[2] See Robert J. Barro and Xavier Sala-i-Martin, "Convergence across States and Regions," *Brookings Papers on Economic Activity* (1991): 107–158; and Olivier Jean Blanchard and Lawrence F. Katz, "Regional Evolutions," *Brookings Papers on Economic Activity* 1 (1992): 1–61. More recent evidence, however, suggests that regional convergence slowed in the past few decades; see Peter Ganong and Daniel Shoag, "Why Has Regional Income Convergence in the U.S. Stopped?" *Journal of Urban Economics* 102 (November 2017): 76–90; and Enrico Moretti, *The New Geography of Jobs,* Boston: Houghton Mifflin Harcourt, 2012.

annual growth rate in a state's manufacturing wage between 1950 and 1990 to the initial wage level in 1950. There is an obvious negative correlation between the rate of wage growth and initial wages. It is estimated that about half the wage gap across states disappears in about 30 years.

Wage convergence is also found in countries where the workforce is less mobile, such as Japan. A study of the Japanese labor market indicates that wage differentials across prefectures (a geographic unit roughly comparable to a large U.S. county) narrow at about the same rate as interstate wage differentials in the United States.[3]

Wage Convergence across Countries

The wage convergence across markets within a particular country might also be observed when we look at labor markets in different countries. There is obviously a great deal of interest in determining whether international differences in per capita income are narrowing.[4] Much of this work is motivated by a desire to understand why the income gap between rich and poor countries seems to persist.

The rate of convergence in incomes across countries plays an important role in the debate over many policy issues, particularly trade.[5] Consider, for example, the long-term effects of the North American Free Trade Agreement (NAFTA). This agreement permits the unhampered transportation of goods (but not of people) across international boundaries throughout much of the North American continent (Canada, Mexico, and the United States).

In 2000, per capita GDP in the United States was over three times as large as that in Mexico. The theory suggests that NAFTA should reduce this income gap. As U.S. firms move to Mexico to take advantage of the cheaper labor, the demand curve for Mexican labor shifts out and the wage differential between the two countries narrows. American workers who are most substitutable with Mexican workers will suffer a wage cut. But, at the same time, American employers profit from the cheaper labor and American consumers gain from the cheaper goods. In short, NAFTA likely created distinct groups of winners and losers in the American and Mexican economies.[6]

[3] Robert J. Barro and Xavier Sala-i-Martin, "Regional Growth and Migration: A Japan–United States Comparison," *Journal of the Japanese and International Economies* 6 (December 1992): 312–346. Other international studies include Christer Lundh, Lennart Schon, and Lars Svensson, "Regional Wages in Industry and Labour Market Integration in Sweden, 1861–1913," *Scandinavian Economic History Review* 53 (2005): 71–84; and Joan R. Roses and Blanca Sanchez-Alonso, "Regional Wage Convergence in Spain 1850–1930," *Explorations in Economic History* 41 (October 2004): 404–425.

[4] Robert J. Barro, "Economic Growth in a Cross-Section of Countries," *Quarterly Journal of Economics* 105 (May 1990): 501–526; N. Gregory Mankiw, David Romer, and David N. Weil, "A Contribution to the Empirics of Economic Growth," *Quarterly Journal of Economics* 107 (May 1991): 407–437; and Xavier Sala-i-Martin, "The World Distribution of Income: Falling Poverty and . . . Convergence, Period," *Quarterly Journal of Economics* 121 (May 2006): 351–397.

[5] George Johnson and Frank Stafford, "The Labor Market Implications of International Trade," in Orley C. Ashenfelter and David Card, editors, *Handbook of Labor Economics,* vol. 3B, Amsterdam: Elsevier, 1999, pp. 2215–2288.

[6] Gordon Hanson, "What Has Happened to Wages in Mexico Since NAFTA?" in Antoni Estevadeordal, Dani Rodrik, Alan Taylor, and Andres Velasco, editors, *Integrating the Americas: FTAA and Beyond,* Cambridge, MA: Harvard University Press, 2004.

The explosive growth in trade with China has been shown to have sizable adverse effects on the wages and employment of the "targeted" American workers at the same time that Chinese per capita income grew. In 1991, Chinese imports were valued at $26 billion. By 2007, Chinese imports had grown by almost 1,200 percent, to $330 billion. As a result, the share of total spending by American consumers on Chinese goods rose from less than 1 percent in 1991 to nearly 5 percent by 2007.

There is a great deal of diversity in the size of the manufacturing sector across U.S. localities.[7] And some of the manufacturing-heavy localities happen to produce the types of goods that directly compete with Chinese imports, while others are more sheltered to Chinese competition. It turns out that there is a strong link between Chinese imports and local labor market conditions. Specifically, the greater the exposure of the local labor markets to Chinese imports (in the sense of producing goods that compete directly with the imported goods), the greater the decline in manufacturing employment and the slower the rate of growth in wages. Moreover, these effects are sizable: the rise in Chinese imports in the locality with median exposure reduced both manufacturing employment and mean weekly earnings by about 1 percent.

Although increased trade inevitably affects the distribution of income within and across countries, our analysis of labor market efficiency implies that the *total* income of the countries is maximized when economic opportunities are equalized. In other words, the equalization of wages across the various countries increases the size of the economic pie available to the entire region. In theory, this additional wealth could be redistributed to the population of the various countries so as to make everyone better off. This link between free trade and economic efficiency is the point emphasized by economists when they argue in favor of more open markets.

4-3 Policy Application: Payroll Taxes and Subsidies

We can show the usefulness of the supply and demand framework by examining a government policy that specifically shifts the labor demand curve. Government programs are often funded through a payroll tax imposed on employers. In the United States, for example, firms paid a tax of 6.2 percent on the first $127,200 of a worker's annual earnings to fund the Social Security program in 2017, and an additional tax of 1.45 percent on all of a worker's earnings to fund Medicare.[8] The payroll tax on employers is even higher in other countries. In Germany, the payroll tax is 17.2 percent; in Italy, it is 21.2 percent; and in France, it is 25.3 percent.[9]

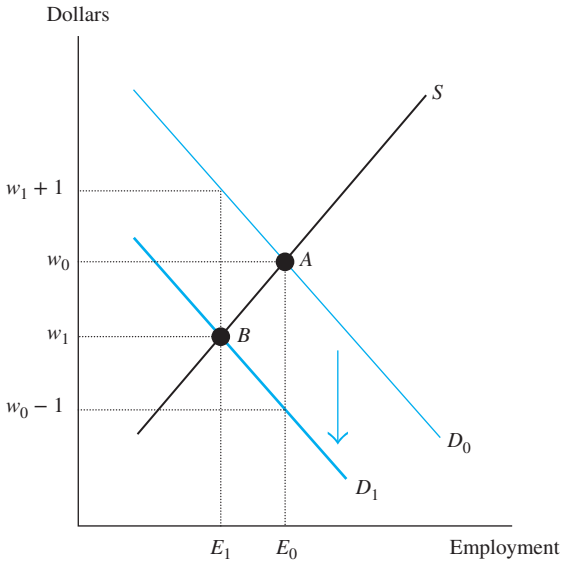
Figure 4-4 shows what happens to wages and employment when the government taxes a firm's payroll. Prior to the imposition of the tax, the industry's labor demand and labor

[7] David H. Autor, David Dorn, and Gordon H. Hanson, "The China Syndrome: Local Labor Market Effects of Import Competition in the United States," *American Economic Review* 103 (October 2013): 2121–2168.

[8] Workers are assessed a similar tax on their earnings, so the total tax rate is 15.3 percent on the first $127,200 of salary, and 2.9 percent on wages above that threshold.

[9] U.S. Bureau of the Census, *Statistical Abstract of the United States, 2012.* Washington, DC: Government Printing Office, 2011, Table 1361.

FIGURE 4-4 A Payroll Tax Imposed on Firms
A payroll tax of $1 imposed on employers shifts down the demand curve (from D_0 to D_1). The tax cuts the wage that workers receive from w_0 to w_1 and increases the cost of hiring a worker from w_0 to $w_1 + 1$.



supply curves are given by D_0 and S, respectively. Point A gives the competitive equilibrium, with E_0 workers hired at a wage of w_0 dollars.

Each point on the demand curve gives the number of workers that employers are willing to hire at a particular wage. In particular, employers are willing to hire E_0 workers if each worker costs w_0 dollars. Consider a very simple payroll tax: The firm will pay a tax of $1 for every employee-hour it hires. In other words, if the wage is $10 an hour, the total cost of hiring an hour of labor will be $11 (with $10 going to the worker and $1 going to the government). Because employers are only willing to pay a *total* of w_0 dollars to hire the marginal worker at E_0, the imposition of the payroll tax implies that employers are now only willing to pay a wage rate of $w_0 - 1$ dollars to hire that marginal worker.

A payroll tax assessed on employers, therefore, leads to a downward parallel shift in the labor demand curve to D_1, as illustrated in Figure 4-4. The new demand curve reflects the wedge that exists between the *total* amount that employers must pay to hire a worker and the amount that workers actually receive from the employer. In other words, employers take into account the *total* cost of hiring labor when they make their hiring decisions—so that the amount that they are willing to pay to workers has to shift down by $1 in order to cover the payroll tax. The payroll tax moves the labor market to a new equilibrium (point *B* in the figure). The number of workers hired declines to E_1. The equilibrium wage rate—the wage rate actually *received* by workers—falls to w_1, but the *total* cost of hiring a worker rises to $w_1 + 1$.

Although the government specifically imposed the payroll tax on employers, the labor market shifted part of the tax to workers. The cost of hiring a worker rose at the same time that the wage received by the worker declined. In a sense, firms and workers "share" the $1 cost of the payroll tax.

A Tax Imposed on Workers

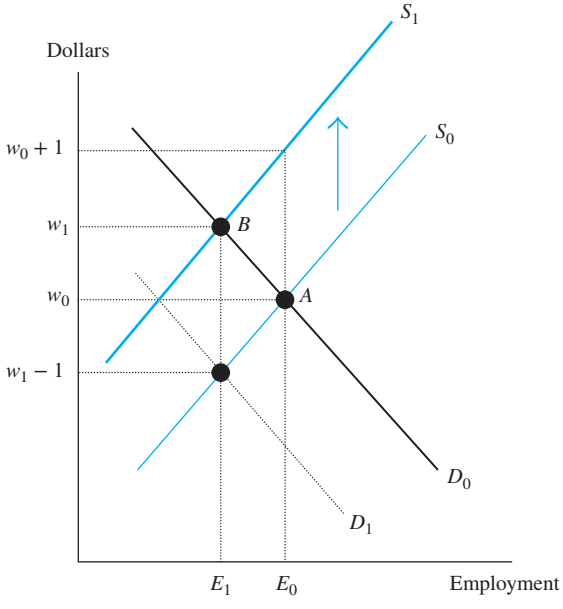
The political debate over payroll taxes often makes it sound as if workers would be better off if the payroll tax were imposed on firms, rather than on workers. It turns out, however, that this perception represents a complete misunderstanding of how a competitive labor market works. *It does not matter whether the tax is imposed on workers or firms*. The impact of the payroll tax on wages and employment is the same regardless of who bears the legal burden of paying the tax.

Suppose, in particular, that the $1 tax had been imposed on workers rather than employers. What would the new equilibrium look like? The labor supply curve gives the wage that workers require to supply a particular number of hours to the labor market. In Figure 4-5, workers will supply E_0 hours when the wage is w_0 dollars. The government now mandates that workers pay the government $1 for every hour they work. Workers, however, will still want to take home w_0 dollars to supply that marginal hour at E_0. In order to supply those many hours, therefore, the workers will now require a payment of $w_0 + 1$ dollars from the employer. In effect, a payroll tax imposed on workers shifts the supply curve up by one dollar to S_1. The shift in supply creates a wedge between the amount that workers must receive from their employers if they are to offer their services in the labor market and the amount that workers get to take home.

The labor market equilibrium then shifts from A to B. At the new equilibrium, workers receive a wage of w_1 dollars from the employer, and total employment falls from E_0 to E_1.

FIGURE 4-5 A Payroll Tax Imposed on Workers

A payroll tax imposed on workers shifts the supply curve to the left (from S_0 to S_1). The payroll tax has the same impact on the equilibrium wage and employment regardless of who it is imposed on.



Note, however, that because the worker must pay a $1 tax per hour worked, the actual after-tax wage falls from w_0 to $w_1 - 1$.

A payroll tax imposed on workers, therefore, leads to the same changes in labor market outcomes as the payroll tax imposed on firms. Both taxes reduce the take-home pay of workers; both increase the cost of an hour of labor to the firm; and both reduce employment.

In fact, we can show that the $1 payroll tax will have *exactly* the same numerical effect on wages and employment regardless of who is legally responsible for paying it. To see this, note that if the $1 payroll tax had been imposed on firms, the demand curve in Figure 4-5 would have shifted down by $1 (see the curve D_1 in the figure). The labor market equilibrium generated by the intersection of this demand curve and the original supply curve (S_0) is the same as the labor market equilibrium that resulted when the tax was imposed on workers. If the tax were imposed on firms, the worker would receive a wage of $w_1 - 1$, and the firm's total cost of hiring a worker would be w_1.

This result illustrates an important principle: The true incidence of the payroll tax (that is, who actually pays it) has little to do with the way the tax law is written or the way the tax is collected. In the end, the true incidence of the tax is determined by the competitive labor market. Even though a payroll tax imposed on the firm shifts down the demand curve, it has the same labor market impact as a revenue-equivalent payroll tax imposed on workers (which shifts up the supply curve).

When Will the Payroll Tax Be Shifted Completely to Workers?

In one extreme case, the payroll tax is shifted entirely to workers. Suppose that the tax is imposed on the firm and that the supply curve is perfectly inelastic, as illustrated in Figure 4-6. A total of E_0 workers are employed in this market regardless of the wage. As before, the imposition of the payroll tax shifts the demand curve down by $1. Prior to the tax, the equilibrium wage was w_0. After the tax, the equilibrium wage is $w_0 - 1$. The more inelastic the supply curve, therefore, the greater the fraction of the payroll taxes that workers end up paying.

Labor supply curves for men are inelastic. It would not be surprising, therefore, if most of the burden of payroll taxes is indeed shifted to workers. Although there is some disagreement regarding the exact amount of the shift, some studies suggest that workers, through a lower competitive wage, pay for as much as 90 percent of payroll taxes.[10]

Deadweight Loss

Because payroll taxes typically increase the cost of hiring a worker, these taxes reduce total employment—regardless of whether the tax is imposed on workers or firms. The after-tax equilibrium, therefore, is inefficient because the number of workers employed is not the number that maximizes the total gains from trade in the labor market.

Figure 4-7*a* illustrates again the total gains from trade accruing to the national economy in the absence of a payroll tax. The total gains from trade are given by the sum of producer surplus and worker surplus, or the area $P + Q$.

[10] Daniel S. Hamermesh, "New Estimates of the Incidence of the Payroll Tax," *Southern Economic Journal* 45 (February 1979): 1208–1219; Jonathan Gruber, "The Incidence of Payroll Taxation: Evidence from Chile," *Journal of Labor Economics* 15 (July 1997, Part 2): S102–S135; and Patricia M. Anderson and Bruce D. Meyer, "Unemployment Insurance Tax Burdens and Benefits: Funding Family Leave and Reforming the Payroll Tax," *National Tax Journal* 59 (March 2006): 77–95.

FIGURE 4-6 Inelastic Supply and a Payroll Tax Imposed on Firms
A payroll tax imposed on the firm is shifted completely to workers when the labor supply curve is perfectly inelastic. The wage is initially w_0. The \$1 payroll tax shifts the demand curve to D_1, and the wage falls to $w_0 - 1$.

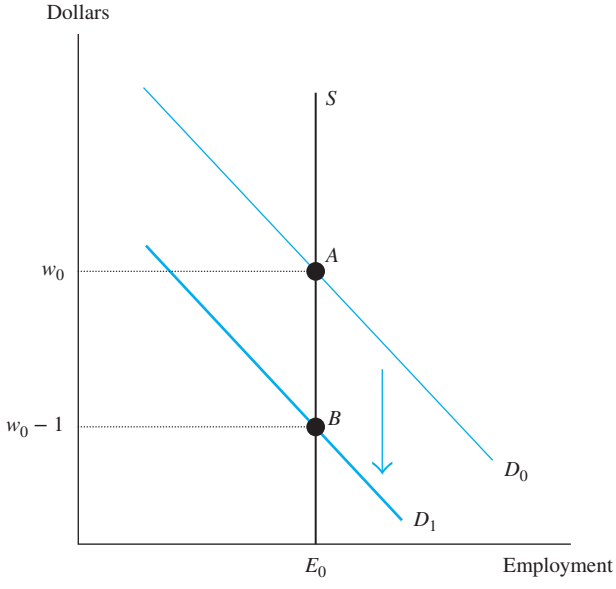


Figure 4-7*b* shows what happens after the government imposes a payroll tax. As we have seen, it does not matter if the payroll tax is imposed on firms or imposed on workers. In either case, employment declines to E_1, the cost of hiring a worker rises to w_{TOTAL}, and the worker's take-home pay falls to w_{NET}. The producer surplus is now given by the smaller triangle P^*, the worker surplus is given by the smaller triangle Q^*, and the tax revenues accruing to the government are given by the rectangle T. The total gains from trade are given by the sum of the new producer surplus and the new worker surplus, as well as the tax revenue. After all, the government will redistribute the tax revenue in some fashion and someone will benefit from those funds. Table 4-1 summarizes the relevant information.

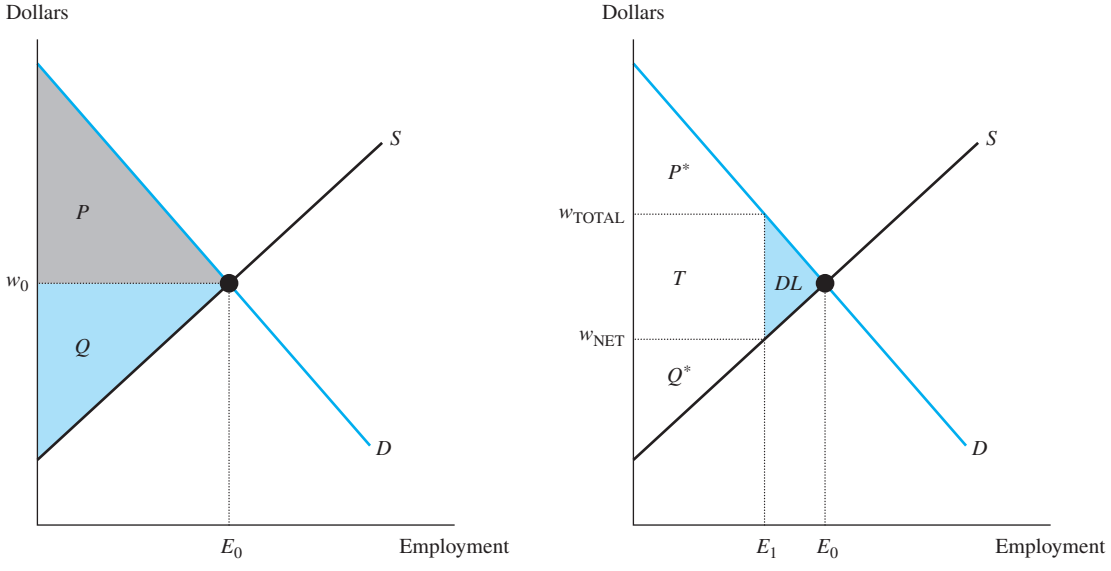
The comparison of subparts (*a*) and (*b*) in Figure 4-7 yields an important conclusion. The imposition of the payroll tax reduces the total gains from trade. There is a triangle, *DL*, that represents the **deadweight loss** (or *excess burden*) of the tax. Note that the deadweight loss measures the value of gains forgone because the tax forces employers to cut employment below the efficient level and has nothing to do with the cost of enforcing or collecting the payroll tax. The deadweight loss arises simply because the tax prevents some workers who were willing to work from being hired by employers who were willing to hire them. These forgone job matches were beneficial to society because the worker's value of marginal product exceeded the worker's value of time outside the labor market.

Employment Subsidies

The labor demand curve is shifted not only by payroll taxes but also by government subsidies designed to encourage firms to hire more workers. An employment subsidy lowers the

FIGURE 4-7 Deadweight Loss of a Payroll Tax

(*a*) In a competitive equilibrium, E_0 workers are hired at a wage of w_0. The triangle *P* gives the producer surplus and *Q* gives the worker surplus. The total gains from trade equal $P + Q$. (*b*) The payroll tax reduces employment to E_1; raises the cost of hiring to w_{TOTAL}; and reduces the worker's take-home pay to w_{NET}. The triangle P^* gives the producer surplus; the triangle Q^* gives the worker surplus; and the rectangle *T* gives the tax revenues. The net loss to society, or deadweight loss, is given by the triangle *DL*.



(*a*) No-Tax Equilibrium

(*b*) Payroll Tax Equilibrium

TABLE 4-1 Welfare Implications of a Payroll Tax

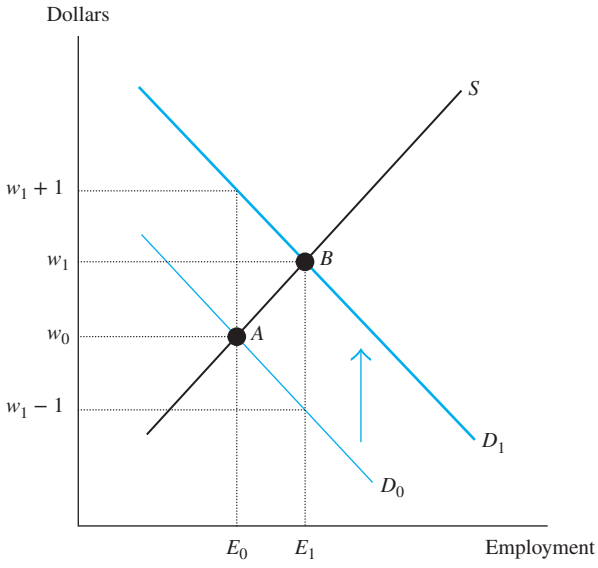
	No-Tax Equilibrium	Payroll Tax Equilibrium
Producer surplus	*P*	P^*
Worker surplus	*Q*	Q^*
Tax revenues	—	*T*
Total gain from trade	$P + Q$	$P^* + Q^* + T$
Deadweight loss	—	*DL*

firm's cost of hiring. In the typical subsidy program, the government grants the firm a tax credit, say of $1, for every person-hour it employs.

Because the subsidy reduces the cost of hiring a person-hour by $1, it shifts the demand curve up by that amount, as illustrated in Figure 4-8. The new demand curve (D_1) gives the price that firms are willing to pay to hire a particular number of workers after they take account of the employment subsidy.

Labor market equilibrium moves from point *A* to *B*. At the new equilibrium, firms hire more workers (from E_0 to E_1). The subsidy also increases the wage that workers actually receive (from w_0 to w_1), and *reduces* the wage that firms actually have to pay out of their own pocket (from w_0 to $w_1 - 1$).

FIGURE 4-8 The Impact of an Employment Subsidy
An employment subsidy of $1 per worker hired shifts the demand curve from D_0 to D_1, increasing employment. The wage that workers receive rises from w_0 to w_1. The cost of hiring falls from w_0 to $w_1 - 1$.



The labor market impact of these subsidies can be sizable and will depend on the elasticities of the labor supply and labor demand curves. For instance, if the labor supply elasticity is 0.3 and the labor demand elasticity is −0.5, a subsidy that reduces the cost of hiring by 10 percent would increase the wage by 4 percent and increase employment by 2 percent.[11]

The largest employment subsidy program in the United States was the New Jobs Tax Credit (NJTC). The program began soon after the recession of 1973–1975 and was in effect from mid-1977 through 1978. The NJTC gave firms a tax credit of 50 percent on the first $4,200 paid to a worker. The firm could claim no more than $100,000 as a tax credit for any given year. Because only the first $4,200 of earnings was eligible for a credit, this program was designed to encourage the employment of low-wage workers. A survey of the evidence concluded that the NJTC generated about 400,000 permanent new jobs.[12] The total cost of the tax credit was roughly $4.5 billion, so each new job cost taxpayers an average of $11,250.

[11] Lawrence F. Katz, "Wage Subsidies for the Disadvantaged," in Richard B. Freeman and Peter Gottschalk, editors, *Generating Jobs,* New York: Russell Sage Press, 1998, pp. 21–53.

[12] Jeffrey Perloff and Michael Wachter, "The New Jobs Tax Credit—An Evaluation of the 1977–78 Wage Subsidy Program," *American Economic Review* 69 (May 1979): 173–179; and John Bishop, "Employment in Construction and Distribution Industries: The Impact of the New Jobs Tax Credit," in Sherwin Rosen, editor, *Studies in Labor Markets,* Chicago, IL: University of Chicago Press, 1981.

4-4 Policy Application: Mandated Benefits

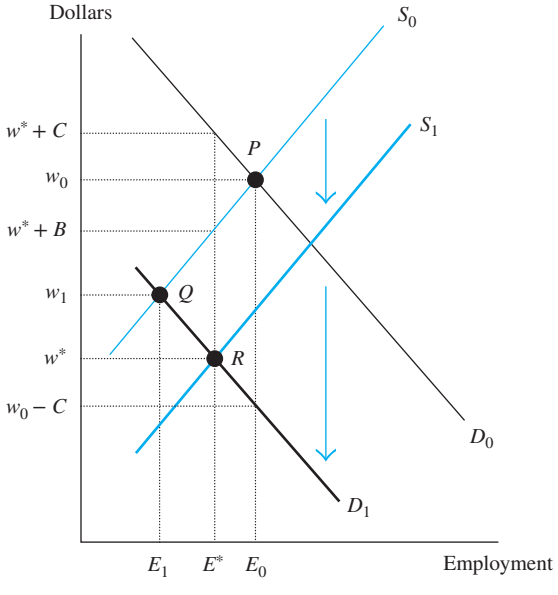
The government can ensure that workers receive particular benefits by mandating that firms provide those benefits. In the United States, for example, the federal government mandates that employers keep the workplace safe or provide accommodations to disabled workers. How do such **mandated benefits** affect equilibrium wages and employment?

It is useful to think in terms of a specific mandated benefit; for example, the provision of spinach pie to workers during the lunch break. Although this might sound a bit far-fetched, it is quite instructive for understanding how the labor market response to mandates differs from the response to payroll taxes.

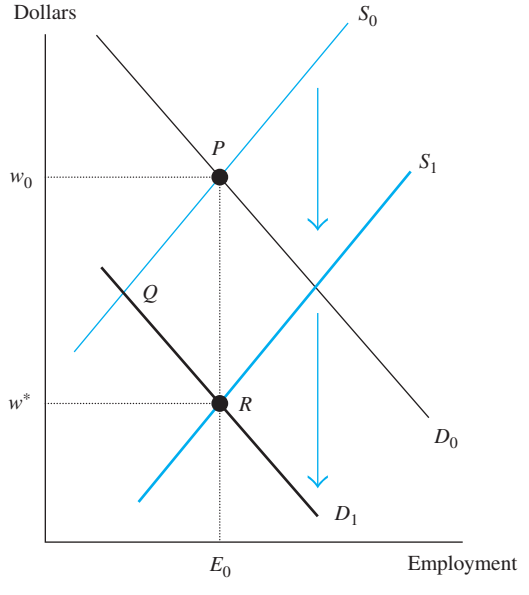
Figure 4-9*a* illustrates how the mandate affects a competitive equilibrium.[13] The initial equilibrium is at point P, with wage w_0 and employment E_0. Suppose that the mandated provision of spinach pie costs C dollars per worker. The mandated provision of this benefit results in a parallel downward shift of the demand curve to D_1, where the vertical difference between the two demand curves is C dollars. The firm is willing to hire E_0 workers only if the total cost of employment for the marginal worker is w_0. The mandated provision

FIGURE 4-9 The Impact of a Mandated Benefit

(*a*) It costs firms C dollars to provide a mandated benefit, shifting the demand curve from D_0 to D_1. Workers value the benefit only by B dollars, so the supply curve shifts down by less. Employment at the new equilibrium (point R) is higher than would have been the case if the firm had been assessed a payroll tax of C dollars (point Q), but lower than in a no-tax equilibrium (point P). (*b*) When the cost of providing the mandate equals the worker's valuation, the resulting equilibrium replicates the competitive no-tax equilibrium in terms of employment, total cost of hiring workers, and total compensation received by workers.



(*a*) Cost of Mandate Exceeds Worker's Valuation



(*b*) Cost of Mandate Equals Worker's Valuation

of spinach pie implies that the firm is now only willing to pay a wage of $w_0 - C$ to that marginal worker.

Suppose workers despise spinach pie, regardless of what the government says about its nutritional value. The government may mandate the firm to provide the benefit; the firms may indeed serve up a slice of spinach pie at lunchtime; but no one can force the workers to eat it. The workers simply take their slice and quickly dispose of it in the trash can. In other words, workers attach no value to this particular benefit. The new labor market equilibrium would then be at point Q, where firms spend a total of $w_1 + C$ dollars to hire a worker (w_1 for the wage and C for the pie), and employment falls to E_1. Note that the equilibrium resulting from a government mandate where workers attach no value to the mandated benefit is what we would have observed if the government had instead enacted a payroll tax of C dollars.

However, it is possible that the typical worker finds the spinach pie tasty, and values the mandated benefit. Suppose that each worker in the industry values the provision of the spinach pie at B dollars, where $B < C$. In other words, workers are willing to pay somewhat less for the spinach pie than what it costs firms to provide it. The fact that the spinach pie makes workers better off implies that the mandated benefit affects not only the demand curve, but also the supply curve. The initial supply curve S_0 in Figure 4-9*a* indicates that E_0 workers are willing to work as long as each receives a *total* compensation of w_0 dollars. Because workers value the spinach pie at B dollars, the E_0 workers are now willing to work as long as firms pay them a wage of $w_0 - B$. In effect, the mandated benefit leads to a parallel downward shift of the supply curve by B dollars, creating the new supply curve S_1.

Because it is costly for firms to provide the spinach pie and because workers value the pie, the new equilibrium is given by the intersection of the new supply and demand curves (point R), so that E^* workers are employed. Although employment falls from E_0 to E^*, it falls by less than it would have fallen if the government had instead imposed a payroll tax of C dollars on firms. In that case, employment would have dropped from E_0 to E_1.

The new equilibrium wage is w^*. But this wage does not represent the value of the employment package from the perspective of either workers or firms. It costs the firm $w^* + C$ dollars to hire a worker; and the worker values the compensation package at $w^* + B$ dollars. In contrast to the initial competitive equilibrium, workers receive less compensation and firms face higher costs. However, in contrast to the payroll tax equilibrium, both firms and workers are better off—workers have higher compensation and firms face lower costs.

Figure 4-9*b* illustrates one special case that is of interest. Suppose that the mandated provision of a spinach pie costs C dollars to the firm *and* that workers value this pie at C dollars. In other words, workers value the mandated benefit just as much as it costs the firm to provide it (so that $B = C$). The supply curve and the demand curve both shift down by *exactly* the same amount (that is, C dollars). At the new equilibrium (point R), employment is still E_0. Similarly, workers value their compensation package at $w^* + C$, and the firm's cost is $w^* + C$. This quantity equals the competitive wage w_0.

The analysis of mandated benefits, therefore, reveals an important property of a competitive equilibrium. As long as the mandated benefit provides some value to workers, the mandated benefit is preferable to a payroll tax because it leads to a smaller cut in employment.

[13] Lawrence H. Summers, "Some Simple Economics of Mandated Benefits," *American Economic Review* 79 (May 1989): 177–183.

Put differently, the government mandate reduces the deadweight loss associated with a payroll tax. In fact, if the cost of providing the mandated benefit exactly equals the value that workers attach to that benefit, there is no deadweight loss. Firms end up hiring exactly the same number of workers they would have hired in a competitive no-tax equilibrium.

Obamacare and the Labor Market

In 2009, nearly two-thirds of Americans under the age of 65 were covered by employer-provided health insurance, but 16 percent did not have any health insurance coverage at all.[14] The long-running debate over whether employers should be required to provide health insurance to all workers culminated in the 2010 enactment of the Patient Protection and Affordable Care Act (ACA), which has come to be known as "Obamacare."

The ACA introduced a complex set of new regulations, mandates, subsidies, penalties, and taxes into the health insurance marketplace, with many of the provisions going into effect on January 1, 2014. Our discussion of payroll taxes and mandated benefits suggests that mandated increases in health insurance participation could have significant labor market effects, including changes in the market wage and in the number of workers employed.

A recent study shows the labor market effects associated with health-related increases in hiring costs using a clever identification strategy.[15] Beginning around 2000, partly because of a substantial increase in malpractice payments, the premiums for physician malpractice insurance soared, which, in turn, increased the cost of employer-provided health insurance. The increase in premium varied greatly across states (depending on the ease with which doctors can be held accountable), suggesting that one can use the state variation in malpractice payments as an instrument to identify how increases in the cost of employer-provided health insurance affect wages and employment. A 10 percent increase in health insurance premiums reduced the probability of employment by 1.2 percentage points, reduced the number of hours worked by 2.4 percent, and lowered the wage of workers with employer-provided health insurance by around 2 percent.

The ACA is a collection of many different programs, each of which can affect labor market outcomes independently (as well as interact with all the other provisions). Among these programs are:

1. An "employer mandate" requiring firms that employ 50 or more full-time workers to offer health insurance to their workforce.
2. An "individual mandate" requiring all individuals to be covered by a health insurance plan. Persons who are not covered by health insurance face a penalty. Congress, however, repealed the individual mandate in 2017.
3. A system of subsidies for persons with incomes between 100 and 400 percent of the federal poverty level to purchase their insurance through an exchange set up by the ACA.
4. Expansion of Medicaid eligibility to include families who have incomes below 133 percent of the federal poverty line (but this expansion only applies in some states).

[14] U.S. Bureau of the Census, *Statistical Abstract of the United States 2012,* Washington, DC: Government Printing Office, 2011, Table 155.

[15] Katherine Baicker and Amitabh Chandra, "The Labor Market Effects of Rising Health Insurance Premiums," *Journal of Labor Economics* 3 (July 2006): 609–634.

Some of the provisions have obvious labor market impacts. Firms, for instance, will inevitably find that the marginal cost of hiring the 50th full-time worker can be substantial, thereby inhibiting employment expansion at that threshold. Similarly, individuals will find that working "too many" hours may put them above the qualifying poverty threshold, which will trigger a suspension of the subsidies and a substantial increase in the cost of insurance, inducing a corresponding labor supply effect. At the same time, however, the fact that individuals can now easily purchase health insurance at various exchanges could make the workforce more mobile and more efficient (because workers will no longer need to be tied down to a particular job to keep their health insurance).

Although it is far too early to measure the net impact of the ACA's many provisions, there are already conflicting findings: The ACA will either decrease employment by 3 percent in the next decade or have no employment impact whatsoever.[16] One prediction, however, is bound to be correct. The complexity of the legislation, and the many unknown interactions among its many provisions, will likely provide full employment for the many economists who will try to document its consequences.

4-5 The Labor Market Impact of Immigration

Because of major policy changes, the United States witnessed a major resurgence in immigration after 1965. In the 1950s, only about 250,000 immigrants entered the country annually. Since 2000, the annual inflow consists of over 1 million legal and illegal immigrants. These sizable supply shifts have reignited the debate over immigration policy.

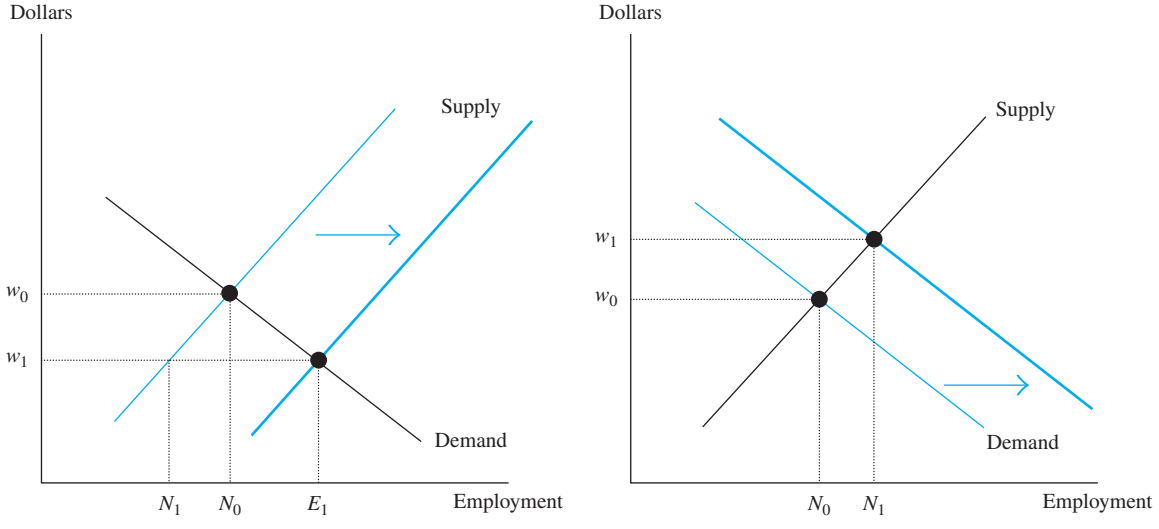
There has also been an immigration surge in many other developed countries. According to the United Nations, 3.4 percent of the world's population (or about 258 million people) reside in a country where they were not born.[17] By 2017, the fraction of foreigners in the country's population was 14.8 percent in Germany, 12.2 percent in France, 15.3 percent in the United States, 21.5 percent in Canada, and 29.6 percent in Switzerland. In many of these countries, the central issue in the immigration debate revolves around the impact of immigration on the employment opportunities of native-born workers.

The simplest model of the labor market impact of immigration starts by assuming that immigrants and natives are perfect substitutes in production. The two groups have the same types of skills and are competing for the same types of jobs. Figure 4-10*a* illustrates the impact of immigration in the short run, with capital held fixed. As immigrants enter the labor market, the supply curve shifts out, increasing total employment from N_0 to E_1 and reducing wages (from w_0 to w_1). Fewer native-born workers are willing to work at this lower wage, so the employment of native workers actually falls, from N_0 to N_1. In a sense, immigrants "take jobs away" from natives by reducing the wage and convincing some native workers that it is no longer worthwhile to work.

[16] Casey Mulligan, *Side Effects and Complications: The Economic Consequences of Health Care Reform,* Chicago: University of Chicago Press, 2015; and Pauline Leung Alexandre Mas, "Employment Effects of the ACA Medicaid Expansion," National Bureau of Economic Research Working Paper No. 22540, August 2016.

[17] United Nations, Department of Economic and Social Affairs. *Trends in International Migrant Stock: The 2017 Revision,* http://esa.un.org/migration/.

FIGURE 4-10 The Short-Run Impact of Immigration
(*a*) *Perfect substitutes.* The two groups compete in the same labor market. Immigration shifts the supply curve to the right. The wage falls from w_0 to w_1 and total employment increases from N_0 to E_1. The number of natives who work at the lower wage drops from N_0 to N_1. (*b*) *Complements.* The two groups do not compete in the same labor market. Immigration makes natives more productive, shifting up the demand curve even though capital is fixed. Both the native wage and native employment increase.



(*a*) Perfect Substitutes (*b*) Complements

The short-run impact of immigration when native workers and immigrants are perfect substitutes, therefore, is unambiguous. As long as the demand curve is downward sloping and capital is fixed, immigration moves the labor market down the demand curve, reducing the wage and employment of native-born workers.

But the assumption that natives and immigrants are perfect substitutes is questionable. It may be that the two groups are not competing for the same types of jobs. For instance, immigrants may be particularly adept at labor-intensive agricultural production. This frees up the more skilled native workforce to perform tasks that make better use of their talents. The presence of immigrants increases native productivity because natives can now specialize in tasks that are better suited to their skills. Immigrants and natives complement each other in the labor market.

If the two groups are complements, an increase in the number of immigrants raises the value of marginal product of natives, shifting up their labor demand curve. As Figure 4-10*b* shows, this increase in native productivity raises the native wage from w_0 to w_1. And some natives who previously did not find it worthwhile to work see the higher wage rate as an incentive to enter the labor market, increasing native employment from N_0 to N_1.

The Short Run and the Long Run

Suppose that immigrants and natives are perfect substitutes. In the short run, the immigrant supply shock means that employers can hire workers at a lower wage, raising the returns to capital and increasing profits. The increased profitability attracts capital into the market, as

old firms expand and new firms open up to take advantage of the cheap labor. The increase in the capital stock, therefore, shifts the labor demand curve to the right, attenuating the initial negative wage impact of immigration.

The crucial question is: By how much will the demand curve shift to the right in the long run? If the demand curve were to shift just a little, the competing native workers would still receive lower wages. If, on the other hand, the demand curve shifted to the right dramatically, the negative wage effects would disappear or even turn positive.

The size of the rightward shift in the demand curve depends on the production technology. To illustrate, suppose that the production function can be described by the well-known Cobb–Douglas production function:

$$q = A K^{\alpha} E^{1-\alpha} \tag{4-1}$$

where A is a constant and α is a parameter that lies between 0 and 1. The production function in equation (4-1) has an important property. If we double labor (E) and double capital (K), output will double. This property is called **constant returns to scale.**[18]

Profit maximization in a competitive labor market requires that the price of capital r (which equals the rate of return to capital) is given by the value of marginal product of capital and that the wage w is given by the value of marginal product of labor. For simplicity, suppose that the price of the output is arbitrarily set to \$1. Using elementary calculus, we can then show that the marginal productivity conditions implied by the Cobb–Douglas are

$$r = \$1 \times \alpha A K^{\alpha-1} E^{1-\alpha} \tag{4-2}$$

$$w = \$1 \times (1-\alpha) A K^{\alpha} E^{-\alpha} \tag{4-3}$$

A little algebraic manipulation shows that we can rewrite these two equations as

$$r = \alpha A \left(\frac{K}{E}\right)^{\alpha-1} \tag{4-4}$$

$$w = (1-\alpha) A \left(\frac{K}{E}\right)^{\alpha} \tag{4-5}$$

Immigration increases the number of workers. Examination of equations (4-4) and (4-5) implies that this increase in E raises the rate of return to capital r and lowers the wage w.

Over time, the higher return to capital will stimulate an increase in the capital stock K. In the long run, after the capital adjusts fully to the supply shock, the rate of return to capital falls back to its "normal" level. This argument implies that the rate of return to capital is fixed in the long run, so that the value of r must be exactly the same before the supply shock and after capital has fully adjusted. Equation (4-4) shows that the only way that the rate of return to capital can be constant in the long run is if the capital–labor ratio (K/E) is also constant in the long run. In other words, if immigration increases the number of workers by, say, 20 percent, the capital stock also must have increased by 20 percent.

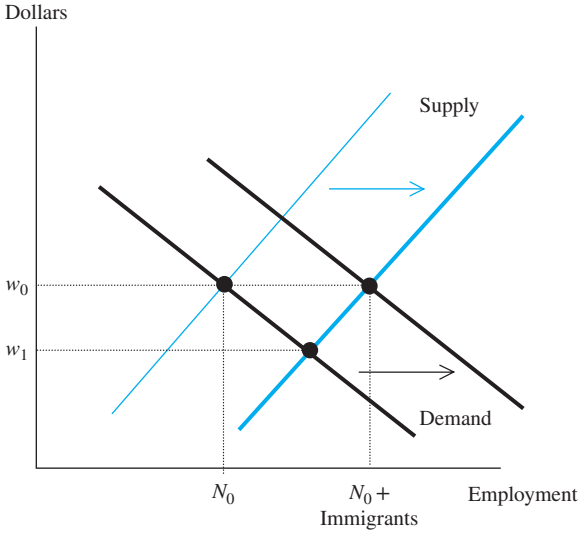
[18] If doubling all inputs leads to less than double the output, the technology has *decreasing returns to scale*. If doubling all inputs leads to more than double the output, the technology has *increasing returns to scale*.

This theoretical insight has interesting implications for the long-run impact of immigration. Equation (4-5) implies that if the capital–labor ratio is constant in the long run, then *the wage must also be constant*. In other words, immigration lowers the wage initially. Over time, the capital stock increases as employers take advantage of cheaper workers. In the end, however, the capital stock completely adjusts to bring the economy back to where it began, with the same rate of return to capital and the same wage rate.[19]

Figure 4-11 illustrates the long-run effect. The labor market is initially in equilibrium at wage w_0, with N_0 natives employed at that wage. In the short run, the supply curve shifts to the right and the wage falls to w_1. In the long run, the demand curve also shifts to the right. *And* it must shift by a sufficient amount to bring the labor market back to its pre-immigration equilibrium. In the end, the wage again equals w_0. At this wage, the number of natives employed in the long run exactly equals the number employed prior to the supply shock.

We do not know how long it takes for the long run to arrive. It is unlikely that the capital stock adjusts instantaneously. It is also unlikely that the capital stock never adjusts. The key insight from the theory is that immigration will have an adverse wage impact on (substitutable) native workers over some time frame, and this impact will weaken as the economy adjusts to the immigrant influx.

FIGURE 4-11 Long-Run Impact of Immigration, Immigrants and Natives Are Perfect Substitutes
Immigration initially shifts out the supply curve, and the wage falls from w_0 to w_1. Over time, capital expands as firms take advantage of the cheaper labor, shifting out the labor demand curve. If the aggregate production function has constant returns to scale, it must be the case that, after all capital adjustments have taken place, the wage returns to its initial level of w_1. In addition, the long-run level of native employment is exactly what it was prior to the immigrant influx.



[19] This theoretical implication does not hinge on the assumption that the production function is Cobb–Douglas. The conclusion that immigration has no long-run labor market impact in the receiving labor market holds whenever the aggregate production function has constant returns to scale.

Spatial Correlations

The discussion suggests a simple way to determine empirically if immigrants and natives are complements or substitutes in production. If they are substitutes, natives should earn less if they reside in labor markets where immigrants are in abundant supply. If they are complements, natives earn more in those labor markets where immigrants cluster.

Much of the empirical research exploits this implication of the theory. The empirical studies typically compare native earnings in cities where immigrants are a substantial fraction of the workforce (for example, Los Angeles or New York) with native earnings in cities where immigrants are a relatively small fraction (such as Pittsburgh or Nashville). The regression model is given by

$$w_{it} = \beta p_{it} + \text{Other variables} \qquad \textbf{(4-6)}$$

where w_{it} is the native wage in city i at time t, and p_{it} is a measure of immigration in that city at that time, such as the percent of the workforce that is foreign-born.

The cross-city correlation between wages and immigration, estimated by the coefficient β, is called a *spatial correlation*. Of course, native wages would vary among labor markets even if there were no immigration. The "other variables" in the regression include factors that also generate wage dispersion across cities, such as differences in native skills or in industrial composition. As an example of the fixed effects methodology introduced in the chapter on labor supply, the empirical studies often include fixed effects for each city. These fixed effects control for city-specific factors that have a permanent impact on local wages. The inclusion of fixed effects implies that the impact of immigration is being estimated by "differencing" the data within each city and observing how a city's wage responds to changes in the number of immigrants settling in that city.

Many of the studies that estimate the regression in equation (4-6) report a negative, but weak, correlation between local wages and immigration.[20] In other words, although the native wage is somewhat lower in those cities where immigrants reside, the wage difference does not seem to be large.

But correlations need not measure the causal effect of immigration. For example, immigrants probably want to settle in high-wage cities with robust labor markets. This settlement pattern would generate a *positive* spurious correlation between immigration and native wages, making it very difficult to detect the potential negative effect implied by the theory (if the two groups are substitutes).

One solution is to find an instrument that somehow leads to an exogenous increase in the number of immigrants settling in a given city. In other words, we need a variable that leads to different numbers of immigrants settling across cities, but this variable must have nothing to do with regional wage differences. The typical instrument used in many of the studies is a lagged measure of immigration in the city. The presumption being that new

[20] Jean B. Grossman, "The Substitutability of Natives and Immigrants in Production," *Review of Economics and Statistics* 54 (November 1982): 596–603; and Joseph G. Altonji and David Card, "The Effects of Immigration on the Labor Market Outcomes of Less-Skilled Natives," in John M. Abowd and Richard B. Freeman, editors, *Immigration, Trade, and the Labor Market,* Chicago: University of Chicago Press, 1991, pp. 201–234. The evidence is summarized in Francine D. Blau and Christopher Mackie, editors, *The Economic and Fiscal Consequences of Immigration,* Washington, DC: National Academy of Sciences Press, 2015, Chapter 5.

immigrants are most likely to settle in those cities where earlier waves settled, perhaps because of lower information costs resulting from close-knit ethnic networks.[21]

The instrument is valid if the settlement patterns of past immigrant waves are in no way correlated with the wage variation across cities today. But the settlement decision of the earlier waves was not random. Those early waves probably settled in cities that offered high-wage at the time they arrived, and high-wage cities may tend to remain high-wage cities for a long time. In other words, the same high wages that attracted the early immigrant waves are attracting the new waves, invalidating the instrument. Recent research suggests that the long-term persistence of high or low wages in particular cities contaminates the regression in equation (4-6), making it difficult to interpret the spatial correlation as the causal effect of immigration.[22]

The Mariel Boatlift

Because the spatial correlation is contaminated by the fact that immigrants settle in high-wage areas, there is a search for alternative strategies that might identify the wage impact of immigration. One such strategy is to look for natural experiments where a large number of immigrants are randomly dropped off in a particular location at a particular time.

On April 20, 1980, Fidel Castro declared that Cuban nationals wishing to move to the United States could leave freely from the port of Mariel. By September 1980, about 125,000 Cubans had chosen to undertake the journey. Almost overnight, Miami's labor force had unexpectedly grown by 7 percent.

An early study concluded that the average wage in Miami was barely affected by the Mariel supply shock.[23] Figure 4-12*a* illustrates the trend in the (inflation-adjusted) hourly wage of white workers in Miami. In 1979, just prior to Mariel, the typical worker earned around $6.40 per hour, and he still earned $6.20 per hour by 1985. Of course, it is hard to interpret the Miami trend without knowing what was going on elsewhere. Put differently, we need to observe the wage trend in a control group, a set of cities unaffected by the Mariel supply shock. The figure compares the Miami experience with that of four cities that had roughly comparable employment conditions at the time (Atlanta, Houston, Los Angeles, and Tampa). In 1979, the typical worker in Miami earned 50 cents less per hour than the typical worker in the control group. By 1985, the gap was 60 cents, a trivial difference of 10 cents an hour, or $4 for a full workweek.

The explosion in refugee flows worldwide has sparked renewed interest in the Mariel experience.[24] Not surprisingly, the recent research, which examines the available data at a

[21] David Card, "Immigrant Inflows, Native Outflows, and the Local Market Impacts of Higher Immigration," *Journal of Labor Economics* 19 (January 2001): 22–64.

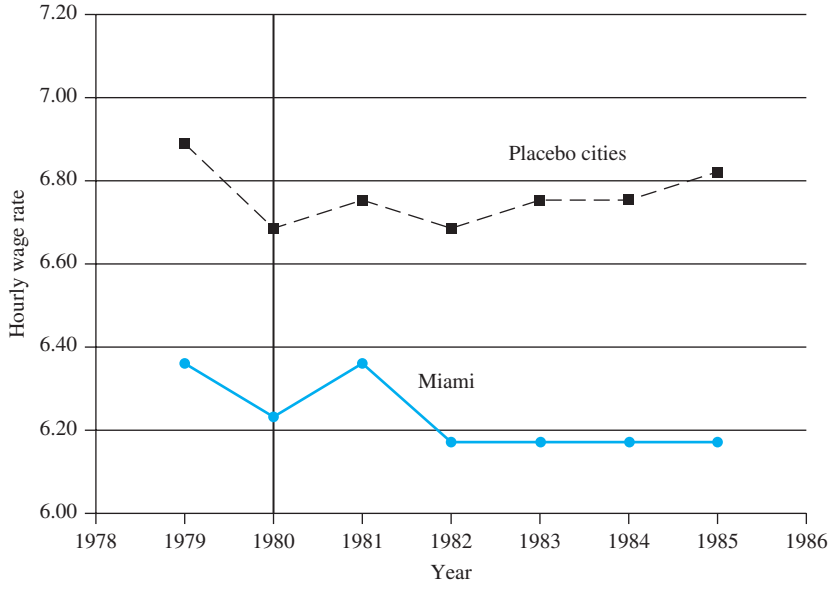
[22] David A. Jaeger, Joakim Ruist, and Jan Stuhler, "Shift-Share Instruments and the Impact of Immigration," Working Paper, Universidad Carlos III de Madrid, November 2017.

[23] David Card, "The Impact of the Mariel Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review* 43 (January 1990): 245–257.

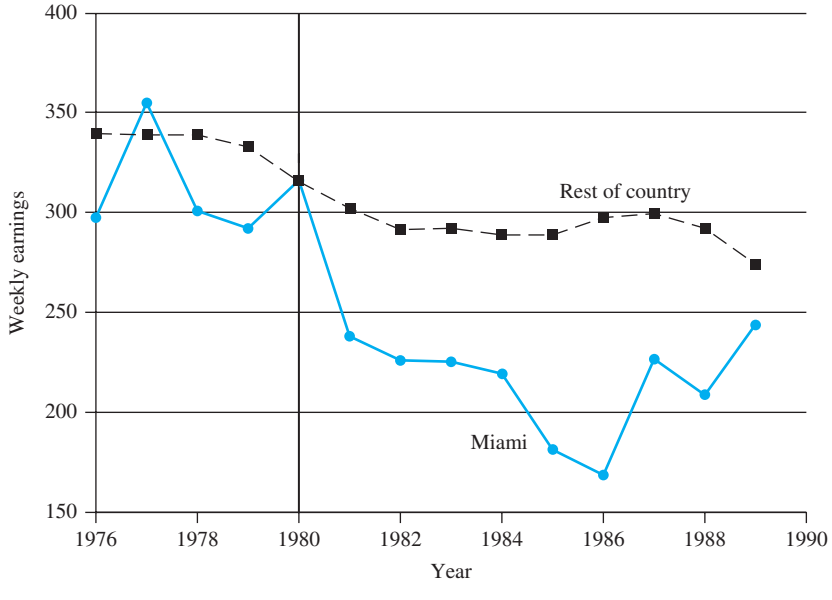
[24] George J. Borjas, "The Wage Impact of the Marielitos: A Reappraisal," Industrial and Labor Relations Review 70 (October 2017): 1077–1110; Giovanni Peri and Vasil Yasenov, "The Labor Market Effects of a Refugee Wave: Applying the Synthetic Control Method to the Mariel Boatlift," NBER Working Paper No. 21801, Cambridge, MA: National Bureau of Economic Research, December 2015; and Michael A. Clemens and Jennifer Hunt. "The Labor Market Effects of Refugee Waves: Reconciling Conflicting Results," NBER Working Paper No. 23433, May 2017.

FIGURE 4-12 The Wage Impact of the Mariel Boatlift

Source: (*a*) Adapted from David Card, "The Impact of the Mariel Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review* 43 (January 1990), p. 250; (*b*) Adapted from George J. Borjas, "The Wage Impact of the *Marielitos:* A Reappraisal," *Industrial and Labor Relations Review* 70 (October 2017): 1077–1110.



(*a*) Trend in average wage



(*b*) Trend in average wage of high school dropouts

more meticulous level, reaches conflicting findings. Almost two-thirds of the Mariel refugees did not have high school diplomas, so that the supply shock increased the number of high school dropouts in the Miami labor market by almost 20 percent.

Much of the recent research focuses specifically on this group to determine if the supply shock had wage consequences. Figure 4-12*b* shows the trend in the average wage of prime-age, non-Hispanic men without a high school diploma in Miami and in the rest of the country. It seems as if the wage of this group took a dramatic nosedive after 1980, and it took a decade for their wage to recover. Other studies claim that the trend in the wage of low-skill workers in Miami depends on the definition of the "low-skill" workforce, that the trend may be affected by changes in the racial composition of the sample, and that there may be a lot of sampling error because of the small number of workers surveyed in the Miami metropolitan area at the time. The increasing importance of refugee flows throughout the world guarantees that the study of the Mariel experience will continue as researchers fine tune the analysis and search for alternative data sources that can provide insights into the consequences of such flows.[25]

Immigration and the National Labor Market

The entry of immigrants into a particular city may lower the wage of competing workers, but that is unlikely to be the end of the story. Natives now have incentives to change their behavior in ways that take advantage of the altered economic landscape. One potential response is to leave the cities that immigrants targeted and move into cities that received few immigrants and now offer relatively higher wages. These shifts in native supply would diffuse the impact of immigration over the national economy, and imply that comparisons of geographic wage differences might provide little information about the true wage impact of immigration.

As a result, some of the research has moved away from geographic comparisons and instead examines trends in the wage of specific skill groups in the national labor market. The "skill-cell approach" tries to determine if the wage changes experienced by specific skill groups is related to the number of immigrants that entered each of those groups.[26]

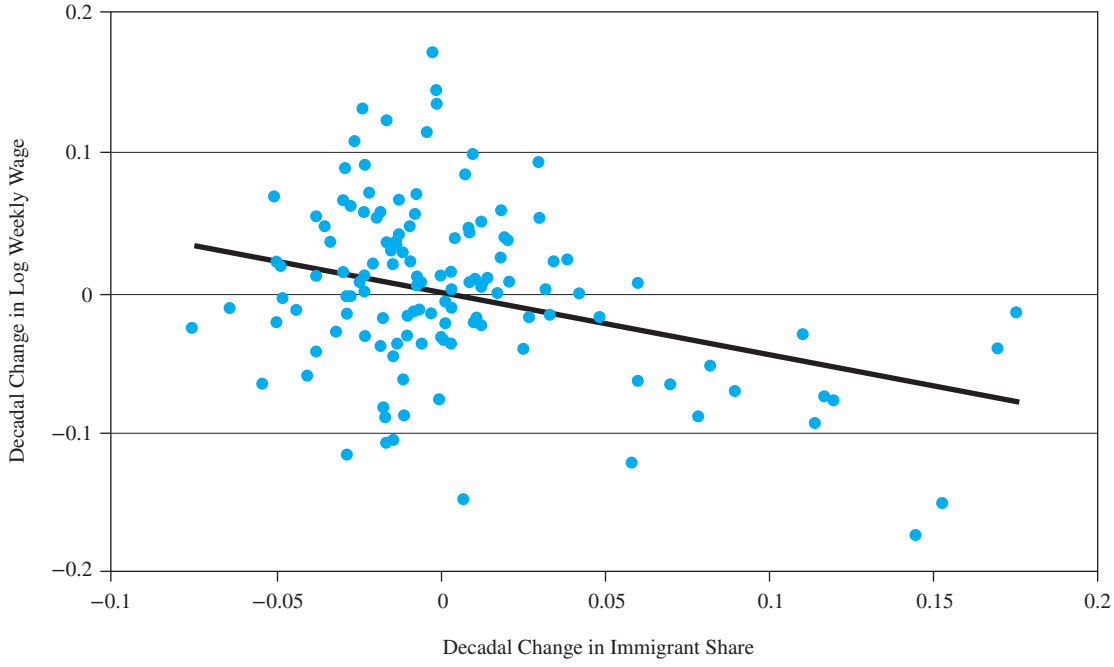
Figure 4-13 summarizes some of this evidence. Define a skill group as the set of workers with a particular combination of education and labor market experience (for example, high school dropouts with 6–10 years of experience, or college graduates with 20–24 years of experience). Each point in the scatter diagram relates the wage growth experienced by a particular skill group of natives over a particular decade to the change in the percent of the group that is foreign born. There is an obvious negative correlation between the two variables; wages grew fastest for those skill groups least affected by immigration. The regression line suggests that a 10 percent increase in the size of the skill group reduces the wage of that group by 3–4 percent.

[25] See, for example, George J. Borjas and Joan Monras, "The Labor Market Consequences of Refugee Supply Shocks" *Economic Policy* 32: 361–413; and Ximena Del Carpio and Mathis Wagner, "The Impact of Syrian Refugees on the Turkish Labor Market," *Journal of Labor Economics,* 2018.

[26] George J. Borjas, "The Labor Demand Curve *Is* Downward Sloping: Reexamining the Impact of Immigration in the Labor Market," *Quarterly Journal of Economics* 118 (November 2003): 1335–1374.

FIGURE 4-13 Scatter Diagram Relating Wages and Immigration for Native Skill Groups, 1960–2000

Source: George J. Borjas, "The Labor Demand Curve *Is* Downward Sloping: Reexamining the Impact of Immigration in the Labor Market," *Quarterly Journal of Economics* 118 (November 2003): 1335–1374. Each point represents the decadal change in the log weekly wage and the immigrant share (that is, the percent of immigrants in the workforce) for a native group of working men defined by years of education and work experience.



The skill-cell approach has been expanded to estimate a full-blown model that specifies the production functions linking output, capital, and the quantity of labor in each of the skill groups. This model-based approach uses the immigrant supply shock in each group as the instrument that shifts supply and identifies the labor demand curve. One benefit from this model-based approach, in contrast to the simple correlation implied by the regression line in Figure 4-13, is that we can estimate how the wage of a particular skill group (for example, college graduates) is affected by the immigration of workers in other skill groups (such as high school dropouts).

Table 4-2, drawn from a recent National Academy of Sciences report, uses the simplest version of this model-based approach to simulate the total wage effect of the immigrants who entered the United States between 1990 and 2010. Even after accounting for all the potential complementarities across skill groups, the 1990–2010 supply shock reduced the wage of the typical worker by 3.2 percent in the short run. As we saw earlier, the theory implies that the long-run wage effect of immigration must be 0.0 percent: The average worker in the labor market is unaffected after all capital adjustments take place. Note, however, that immigration has distributional effects even in the long run, with the average

TABLE 4-2

Percent Wage Effect of the 1990–2010 Immigrant Supply Shock, Accounting for Cross Effects

Source: Francine D. Blau and Christopher Mackie, editors, *The Economic and Fiscal Consequences of Immigration,* Washington, DC: National Academy of Sciences Press, 2015, Table 5-1.

	Short Run	Long Run
Education group:		
High school dropouts	−6.3	−3.1
High school graduates	−2.8	0.4
Some college	−2.3	0.9
College graduates	−3.3	−0.1
Postcollege	−4.1	−0.9
All workers	−3.2	0.0

wage of high school dropouts falling by about 3 percent and the average wage of workers with some college education rising by about 1 percent.[27]

The skill-cell approach has been used to examine the link between immigration and wages in other countries. One particularly interesting study examines the impact of *emigration* on wages in Mexico.[28] Emigration (almost entirely to the United States) has reduced the size of the Mexican workforce by about 10 percent. This outflow should increase wages in Mexico. There is indeed a strong positive correlation between the number of emigrants in a particular skill group and the wage growth experienced by that group in Mexico. A 10 percent reduction in the size of the group raised the wage of the workers who remained in Mexico by about 3 percent.

The Immigration and Minimum Wage Debates

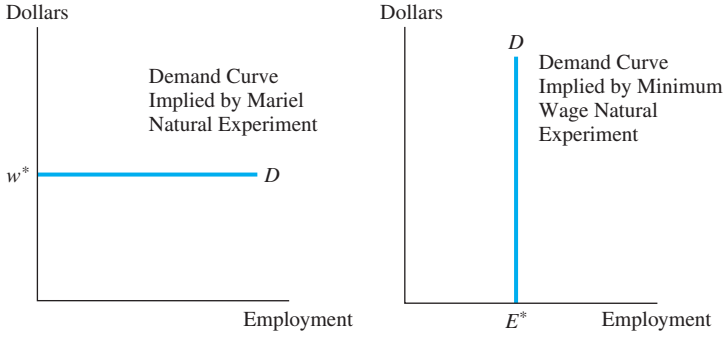
The immigration and minimum wage debates share one other crucial feature. Both debates focus on exactly the same parameter—the elasticity of labor demand. In the minimum wage context, we care about how an exogenous shift in the wage affects employment; in the immigration context, we care about how an exogenous shift in supply shifts the wage.

Suppose we take the results from the original Mariel study at face value, and infer that immigration has little impact on the native wage—even in the short run. Figure 4-14*a* shows the short-run labor demand curve implied by that inference. It is a perfectly elastic curve, indicating that wages are constant regardless of the number of workers.

[27] The National Academy also reports simulations that allow for complementarities between immigrants and natives who have observationally similar skills, and for the possibility that high school dropouts and high school graduates are perfect substitutes. Both of these alternative assumptions attenuate the adverse wage impact of immigration. The evidence on the validity of these alternative assumptions is mixed; see David Card, "Immigration and Inequality," *American Economic Review* 99 (May 2009): 1–21; Gianmarco I. P. Ottaviano and Giovanni Peri, "Rethinking the Effect of Immigration on Wages," *Journal of the European Economic Association* 10 (February 2012): 152–197, and George J. Borjas, Jeffrey Grogger, and Gordon H. Hanson, "On Estimating Elasticities of Substitution," *Journal of the European Economic Association* 10 (February 2012): 198–210.

[28] Prachi Mishra, "Emigration and Wages in Source Countries: Evidence from Mexico," *Journal of Development Economics* 82 (January 2007): 180–199.

FIGURE 4-14 The Short-Run Labor Demand Curve Implied by Different Natural Experiments
(*a*) The short-run labor demand curve is perfectly elastic if the data from the Mariel natural experiment indicates that increased immigration did not affect the wage. (*b*) The short-run labor demand curve is perfectly inelastic if the data from the New Jersey–Pennsylvania minimum wage natural experiment indicates that the minimum wage does not affect employment.



In the chapter on labor demand, we discussed the study that attempted to measure the impact of the minimum wage on employment in the fast-food industry.[29] That empirical exercise compared employment in New Jersey and Pennsylvania prior to and after the imposition of a state minimum wage in New Jersey. Because the minimum wage only increased in New Jersey, one would expect that fast-food employment in New Jersey declined relative to fast-food employment in Pennsylvania. But no such decline occurred.

Suppose again we take the evidence from the New Jersey–Pennsylvania natural experiment at face value. We can then infer that minimum wages have little impact on employment. Figure 4-14*b* illustrates the short-run labor demand curve implied by this natural experiment. It is a perfectly inelastic curve, indicating that employment is constant regardless of the level of the wage.

Needless to say, at least one of these two demand curves must be wrong. The short-run labor demand curve cannot be both perfectly elastic and perfectly inelastic at the same time. One could perhaps argue that the data are the data—and that at a particular time and in a particular place that is what the labor demand curve looked like. But this approach makes the inferences from experimental evidence completely useless—as the evidence from one particular experiment cannot be used to predict what would happen as a result of policy shifts at other times and in other places.

In short, the evidence from natural experiments, particularly in areas where the empirical evidence is highly polarizing because of its political implications, should be interpreted with caution and skepticism. Before proceeding to buy into the argument that the minimum wage does not affect short-run employment *and* that immigration does not affect short-run wages, it is worth asking whether such a claim is internally consistent with any underlying theoretical framework.

[29] David Card and Alan B. Krueger, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review* 84 (September 1994): 772–793.

4-6 The Immigration Surplus

The flow of workers across labor markets helps the economy move toward a more efficient outcome. Although immigration may have an adverse impact on the wage of competing native workers, immigrants may also generate efficiency gains for the aggregate economy.[30]

Consider the short-run analysis illustrated in Figure 4-15. The labor supply curve is initially given by S and the demand curve by D. For simplicity, suppose that supply is inelastic, so there are N native workers regardless of the wage. Competitive equilibrium implies that the N natives are employed at a wage of w_0.

Recall that the labor demand curve gives the value of marginal product schedule, so that each point on the demand curve tells us the contribution of the last worker hired. As a result, the area under the demand curve gives the total product of all workers hired. The area in the trapezoid $ABN0$ measures the value of national income prior to immigration.

What happens when immigrants enter the country? Suppose that immigrants and natives are perfect substitutes. The supply curve shifts to S' and the market wage falls to w_1. National income is now given by the area in the trapezoid $ACM0$. The figure shows that the total wage bill paid to immigrants is given by the area in the rectangle $FCMN$, so that the increase in national income kept by natives is given by the area in the triangle BCF. This triangle is called the **immigration surplus** and measures the increase in wealth produced by immigrants and accruing to natives.

Why is there an immigration surplus? Because the market wage equals the productivity of the *last* worker hired. Immigrants then increase national income by more than what it costs to employ them. Put differently, all the immigrants hired, except for the last one, add more to the economy than they get paid.

The analysis in Figure 4-15 implies that if the demand curve is perfectly elastic (so that immigrants had no impact on the native wage), immigrants would be paid their entire value of marginal product and natives would gain nothing. A positive immigration surplus requires that native wages fall when immigrants enter the country. Therefore, immigration redistributes income from labor to capital. Native workers lose the area in the rectangle w_0BFw_1, and this quantity plus the immigration surplus goes to employers. Although native workers earn less, their losses are more than offset by the increase in income accruing to native-owned firms.

Calculating the Immigration Surplus

The formula for the area of a triangle is one-half times the base times the height. Figure 4-15 implies that the immigration surplus is

$$\text{Immigration surplus} = \frac{1}{2} \times (w_0 - w_1) \times (M - N) \qquad \textbf{(4-7)}$$

It is convenient to define the surplus as a fraction of national income. After rearranging terms in the equation, we get[31]

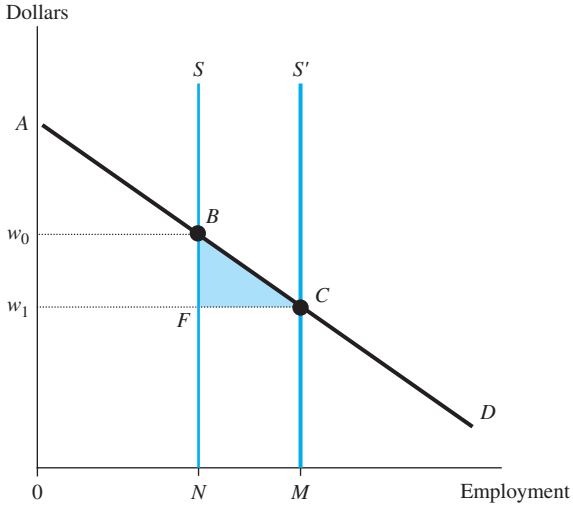
[30] George J. Borjas, "The Economic Benefits from Immigration," *Journal of Economic Perspectives* 9 (Spring 1995): 3–22.

[31] In particular, we can rewrite the immigration surplus as

$$\frac{\text{Immigration Surplus}}{\text{GDP}} = \frac{1}{2} \times \frac{w_0 - w_1}{w_1} \times \frac{M - N}{M} \times \frac{w_1 M}{\text{GDP}}$$

FIGURE 4-15 The Immigration Surplus

Prior to immigration, there are N native workers and national income is given by the trapezoid $ABN0$. Immigration increases labor supply to M workers and national income is given by the trapezoid $ACM0$. Immigrants are paid a total of $FCMN$ dollars as salary. The immigration surplus gives the increase in national income that accrues to natives and is given by the area in the triangle BCF.



$$\frac{\text{Immigration surplus}}{\text{GDP}} = \frac{1}{2} \times (\%\ \text{change in native wage rate}) \times (\%\ \text{change in employment}) \times (\text{labor's share of national income}) \quad \textbf{(4-8)}$$

where GDP gives the country's gross domestic product, and "labor's share of national income" is the fraction of GDP that accrues to workers.

Immigration has increased labor supply in the United States by about 15 percent. As discussed earlier, some evidence suggests that wages may fall by about 3 percent for every 10-percent increase in supply, so that a 15-percent increase in supply would lower wages by 4.5 percent. Finally, it is well known that labor's share of national income is about 0.7. This implies that immigration increased the income of natives by 0.24 percent (or $0.5 \times 0.045 \times 0.15 \times 0.7$). The GDP of the United States in 2018 neared $19 trillion, so that the economic gains from immigration are relatively small in the context of such a large economy, about $46 billion per year.

This estimate of the immigration surplus is a short-run estimate. In the long run, as long as there are constant returns to scale, neither the rate of return to capital nor the wage is affected by immigration. As a result, the long-run immigration surplus must be equal to zero. Immigrants increase GDP in the long run, but the entire increase in national income is paid to immigrants for their services.

4-7 Policy Application: High-Skill Immigration

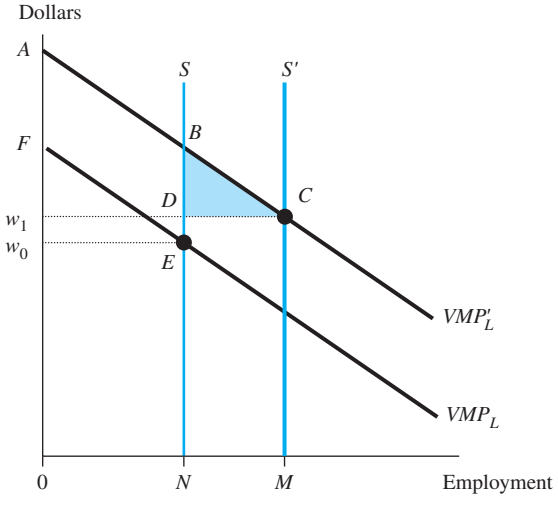
The calculation of the immigration surplus in a competitive labor market suggests that even a large supply shock may not generate relatively large gains for the native population. Nevertheless, there is a widespread perception that some types of immigration, and particularly the immigration of high-skill workers, can be hugely beneficial. This perception relies on a crucial departure from the textbook model, the assertion that high-skill immigrants generate **human capital externalities.** The sudden presence of high-skill immigrants exposes natives to new forms of knowledge, increases their human capital, and makes them more productive.

It is easy to illustrate how human capital externalities can substantially increase the immigration surplus. If high-skill immigrants had positive spillover effects on native productivity, an influx of immigrants induces an outward shift in the labor demand curve because the value of marginal product for a native worker rises.

Figure 4-16 illustrates the short-run model. Immigration not only shifts the supply curve, but also shifts the demand curve from VMP_L to VMP_L'. The change in national income accruing to natives is then given by the sum of the triangle *BCD and* the area of the trapezoid *ABEF*, which measures the impact of immigration on the total product of native workers. It is obvious that if the human capital externalities are sufficiently important, the beneficial spillover effects from high-skill immigration could be an important driver of economic growth.

Several recent studies attempt to document the empirical importance of these externalities. This research typically examines case studies where a country is "hit" by an exogenous

FIGURE 4-16 The Immigration Surplus in the Presence of Positive Externalities
There are N native workers. Immigration increases labor supply to M workers and the positive human capital externalities shift the demand curve to VMP_L'. The wage rises from w_0 to w_1. Immigrants receive *DCMN* in wage payments. Native income increases by the sum of the trapezoid *ABEF* and the triangle *BCD*.



supply shock of high-skill immigrants and then traces out its consequences for the productivity of the affected natives.[32]

Nazi Germany

Immediately after seizing power in 1933, the National Socialist Party enacted legislation known as the *Law for the Restoration of the Professional Civil Service.* This Orwellian-named statute mandated the dismissal of all Jewish professors from German universities. A remarkable 18 percent of German mathematics professors were dismissed between 1932 and 1934.[33]

The dismissals included some of the most famous mathematicians of the era, including John von Neumann, Richard Courant, and Richard von Mises. Many of the dismissed mathematicians eventually managed to migrate to other countries, mainly the United States. Von Neumann, for instance, settled at Princeton University where, after teaming with an economist, Oskar Morgenstern, he wrote a landmark treatise, *The Theory of Games and Economic Behavior.*

The Jewish mathematicians had not been randomly employed across German universities, so some departments barely noticed the departure of the luminaries, while other departments lost more than 50 percent of the faculty. The most affected departments included some of the best mathematics departments in the country, including the University of Göttingen and the University of Berlin. A remarkable exchange between the Nazi Minister of Education and David Hilbert, one of the most famous mathematicians of the twentieth century, summarizes the impact:

> *Minister:* How is mathematics in Göttingen now that it has been freed of Jewish influence?
> *Hilbert:* Mathematics in Göttingen? There is really none any more.

If highly skilled mathematicians produce beneficial externalities for their students, we would expect to see the impact of the dismissals on the doctoral students "left behind." Using archival records from German universities, it is possible to compare the professional careers of doctoral students who were enrolled in German universities before and after the dismissals. The departure of so many leading mathematicians should presumably affect the output of the affected doctoral students (as measured for example, by the number of research papers the student eventually publishes or by the number of citations those publications received). The presence of human capital externalities would imply that post-1933 students graduating from the departments that suffered a serious decline in quality (for example, the departure of von Neumann) should have lower lifetime productivity than the students who graduated from those same departments just prior to 1933.

[32] A related literature examines the impact of the high-skill H-1B visa program in the United States by correlating the supply shock with local labor market outcomes. See William R. Kerr and William F. Lincoln, "The Supply Side of Innovation: H-1B Visa Reforms and U.S. Ethnic Invention," *Journal of Labor Economics* 28 (July 2010): 473–508; and Giovanni Peri, Kevin Shih, and Chad Sparber, "STEM Workers, H-1B Visas and Productivity in U.S. Cities," *Journal of Labor Economics* 33 (July 2015; Part 2): S225–S255.

[33] Fabian Waldinger, "Quality Matters: The Expulsion of Professors and the Consequences for PhD Student Outcomes in Nazi Germany," *Journal of Political Economy* 118 (August 2010): 787–831.

FIGURE 4-17 The Dismissal of Jewish Mathematicians and Student Productivity

Source: Fabian Waldinger, "Quality Matters: The Expulsion of Professors and the Consequences for PhD Student Outcomes in Nazi Germany," *Journal of Political Economy* 118 (August 2010), p. 813.

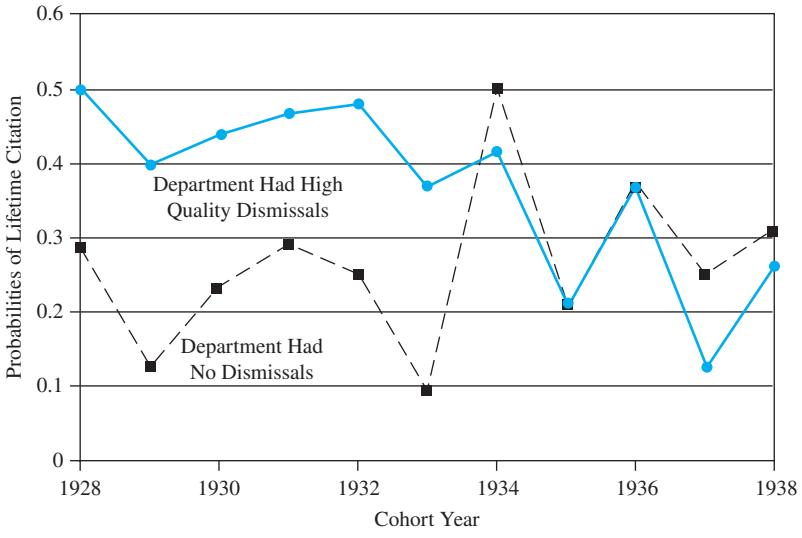


Figure 4-17 summarizes the evidence. It shows the probability of receiving one lifetime citation for each cohort of students in two types of departments: Departments where no dismissals occurred and departments where high-quality professors were dismissed. It is evident that students enrolled in the departments that suffered the heaviest losses experienced a relative decline in their productivity.

The correlation between student output and faculty quality, therefore, suggests that human capital externalities matter. The absence of Jewish mathematical luminaries depressed the future output of those students who were deprived of their mentorship.

The Soviet Union

Soon after the collapse of the Soviet Union in 1992, over 1,000 Soviet mathematicians (or roughly 10 percent of the stock) left the country, with a third eventually settling in the United States.[34]

For decades prior to 1992, Soviet and Western mathematicians had essentially lived "separate lives," with little contact between the two communities. A key event cementing the separation was the so-called Luzin affair. In 1936, Nikolai Luzin, a mathematician at Moscow State University and a member of the USSR Academy of Sciences, became the target of a Stalinist witch-hunt. The allegations included the generic charge of promoting anti-Soviet propaganda *and* the accusation that Luzin saved his best theorems for publication in Western journals. Soviet mathematicians quickly grasped the lesson: They should

[34] George J. Borjas and Kirk B. Doran, "The Collapse of the Soviet Union and the Productivity of American Mathematicians," *Quarterly Journal of Economics* 127 (August 2012): 1143–1203.

only publish in Soviet journals. In addition, the Soviet government imposed strict restrictions on communications with Western peers, on whether they could travel, and on access to Western materials.

Just as speakers of one language, when separated geographically for many generations, eventually develop separate and different dialects, so Soviet and Western mathematicians began to specialize in very different fields. The two most popular Soviet fields were partial differential equations and ordinary differential equations, and those two fields accounted for about 18 percent of all pre-1990 publications. In contrast, the two most popular American fields were statistics and operations research, and those two fields accounted for 16 percent of all American publications.

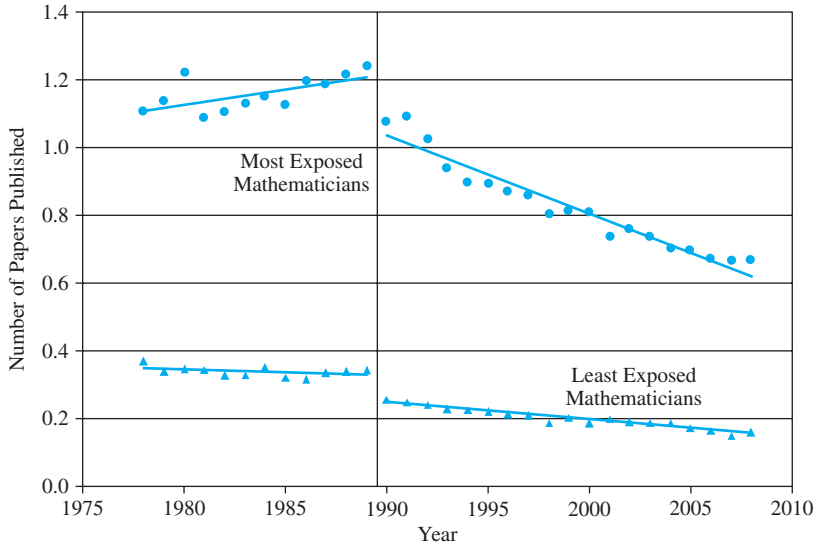
Using publication records from the American Mathematical Society, it is possible to track the publication record of every American mathematician before and after the arrival of the Soviet émigrés. We could then measure the impact of the supply shock on the mathematicians whose research agenda most overlapped with that of the Soviets.

There are two possible effects. The first is implied by the law of diminishing returns. An increase in the number of mathematicians deriving theorems in, say, partial differential equations reduces the market value of such mathematicians and will make the pre-existing American mathematicians less productive. The second is implied by human capital externalities. Exposing American mathematicians to the new theorems and techniques brought in by their Soviet counterparts could lead to a renaissance in mathematical ideas and increase the productivity of American mathematicians working in those fields.

Figure 4-18 illustrates the impact of the Soviet influx on the productivity of American mathematicians. The "most exposed" group had specialized in Soviet-style research topics

FIGURE 4-18 The Collapse of the Soviet Union and American Mathematics

Source: George J. Borjas and Kirk B. Doran, "The Collapse of the Soviet Union and the Productivity of American Mathematicians," *Quarterly Journal of Economics* 127 (August 2012), p. 1172.



Theory at Work

HURRICANES AND THE LABOR MARKET

Hurricanes develop over warm water, where the ocean's temperature exceeds 80 degrees Fahrenheit. Due to the high temperatures needed to fuel the storm, most hurricanes that strike the United States first touch land in the states that surround the Gulf of Mexico or the Southeastern states, particularly Florida. Hurricanes are categorized according to the Saffir-Simpson Scale based on their wind speed and are given a number from 1 to 5. Category 1 hurricanes have wind speeds ranging from 74 to 95 miles per hour at the time of landfall. Category 4 hurricanes have wind speeds between 131 and 155 miles per hour. Andrew, a category 5 hurricane, had wind speeds above 180 miles per hour when it first hit land in 1992.

Although we can predict with confidence that the hurricane season will generate some hurricanes and that Florida will likely be hit by some of them, the exact timing and path of the hurricanes cannot be forecast. As a result, each of these hurricanes generates exogenous shocks in the Florida counties that are directly hit. The randomness of the path provides a natural experiment that can be used to analyze how the deadly storms alter labor market conditions. Because so many hurricanes have hit Florida in the past two decades, we can use the available data to estimate difference-in-differences models that examine the economic impact on the affected Florida counties relative to the economic events in the unaffected counties.

When a strong hurricane strikes a particular county, some people will flee—causing at least a temporary decline in the number of workers available. The hurricane-induced reduction in supply suggests that wages would rise and employment would fall in the counties directly affected by the hurricane. Many of these "refugees" would likely move to neighboring counties at least in the short run. This implies that the supply of labor would increase in the neighboring counties, and that the wage in those counties may actually fall.

The table below summarizes the labor market consequences of the hurricanes that hit Florida between 1988 and 2005. The evidence seems consistent with the simple story that labor supply shifts induced by the hurricanes led to corresponding employment and wage shifts both in the county hit by the hurricane and in surrounding counties.

EMPLOYMENT AND WAGES IN FLORIDA COUNTIES HIT BY CATEGORIES 4 AND 5 HURRICANES (RELATIVE TO CHANGE OBSERVED IN THE AVERAGE COUNTY)

Source: Ariel R. Belasen and Solomon W. Polachek, "How Disasters Affect Local Labor Markets: The Effects of Hurricanes in Florida," *Journal of Human Resources* 44 (Winter 2009), Table 4.

	% change in employment	% change in earnings
Effect on county directly hit	−4.5	+4.4
Effect on neighboring county	+0.8	−3.3

prior to 1990. These mathematicians had a slight upward pre-1990 trend in the average number of papers published annually. In contrast, the American mathematicians in the "least exposed" group, those who specialized in fields that had little in common with Soviet research interests, had a slight downward trend. After 1990, there was a precipitous decline in the publication rate of the group whose research agenda overlapped most with the Soviets. In this particular context, therefore, it seems that competitive effects outweigh the potential benefits from human capital spillovers.

4-8 The Cobweb Model

The typical analysis of labor market equilibrium assumes that markets adjust instantaneously to shifts in either supply or demand curves, so that wages and employment change swiftly from the old equilibrium levels to the new. Many labor markets, however, may not adjust so quickly to supply and demand shocks.

Consider, in particular, the market for new engineering graduates. It has long been recognized that this market fluctuates regularly between periods of excess demand for labor and periods of excess supply, creating a cyclical trend in the entry wage of engineering graduates. Two key assumptions underlie the model often used to explain how such wage trends arise: (1) It takes time to produce a new engineer; and (2) persons decide whether or not to become engineers by looking at conditions in the engineering labor market *at the time they enter school.*[35]

Figure 4-19 presents the supply and demand curves for new engineers. The labor market is initially in equilibrium where the supply curve S intersects the demand curve D. There are E_0 new engineering graduates and the entry wage is w_0. Suppose there is a sudden increase in the demand for newly trained engineers (such as the one created by the race to get a man on the moon in the 1960s). The demand curve shifts to D', and firms would like to hire E^* new engineers at a wage of w^*.

But this is easier said than done. It will be very difficult for firms to find those additional engineers. New engineers do not come out of thin air simply because firms want to hire them. It takes time to train them. Because engineering schools are only producing E_0 engineers annually, the *short-run* supply curve is, in fact, perfectly inelastic at E_0 workers. The combination of this inelastic supply curve (that is, a vertical line at E_0) and the demand shift increases the entry wage of engineers to w_1.

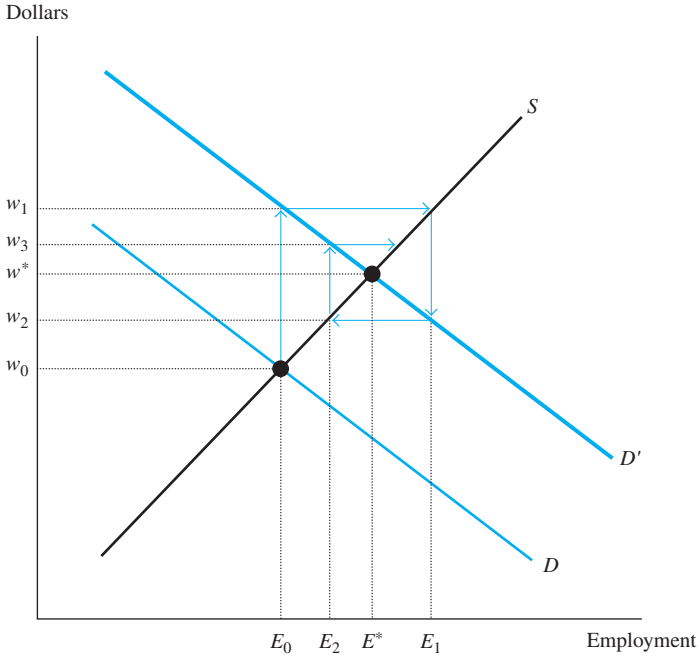
While all this is happening, a new generation of students is deciding whether to enter the engineering profession. These students see the high wage of w_1, and this high wage induces many of them to become engineers. In fact, the supply curve in Figure 4-15 indicates that at the wage of w_1, a total of E_1 persons will want to enroll in engineering schools.

After a few years, therefore, E_1 new engineers enter the marketplace. When this cohort of engineers enters the market, the short-run supply of new engineers is perfectly inelastic at E_1 workers. The market situation is then summarized by this inelastic supply curve and the demand curve D' (assuming that demand conditions have not changed any further). Equilibrium occurs at a wage of w_2, which is substantially below the wage that the new engineers thought they were going to get. Because high school and college graduates presumed that they would get the high wage of w_1 that prevailed at the time they made their career decisions, there was an oversupply of engineers a few years down the line.

But this is not the end of the story. Still another generation of students is making a career choice. At the low wage of w_2, the engineering profession is not attractive, and few persons will want to enter it. The supply curve implies that only E_2 persons become

[35] Richard B. Freeman, "A Cobweb Model of the Supply and Starting Salary of New Engineers," *Industrial and Labor Relations Review* 29 (January 1976): 236–246; and Richard B. Freeman, "Supply and Salary Adjustments to the Changing Science Manpower Market: Physics, 1948–1973," *American Economic Review* 65 (March 1975): 27–39.

FIGURE 4-19 The Cobweb Model in the Market for New Engineers
The initial equilibrium wage is w_0. The demand for engineers shifts to D', and the wage will eventually increase to w^*. Because new engineers are not produced instantaneously and because students might misforecast future opportunities, a cobweb is created as the labor market adjusts to the increase in demand.



engineers at a wage of w_2. When these students enter the labor market, the entry wage rises to w_3 because there was an undersupply of engineers. This high wage induces the next generation of students to oversupply, and so on.

The **cobweb model** creates a cobweb around the equilibrium point as the engineering labor market adjusts to the initial demand shock. The entry wage exhibits a systematic pattern of booms and busts as the market slowly drifts toward its long-run equilibrium wage w^* and employment E^*.[36]

Assumptions of the Cobweb Model

The cobweb model makes two key assumptions. The first is reasonable: It takes time to produce a new engineer, so the supply of engineers can be thought of as being perfectly inelastic in the short run. The second is questionable: Students are very myopic when they are deciding whether to become engineers. Students choose a career based entirely on the

[36] As drawn in Figure 4-19, wages and employment in the engineering market drift toward their equilibrium levels over time. It should be evident that, depending on the values of the elasticities of supply and demand, the cobweb model can also generate booms and busts where wages and employment diverge *away* from equilibrium.

wage they currently observe in the engineering market and do not "look into the future" when making the decision.

But potential engineers obviously have strong incentives to know about trends in the wage of newly minted engineers. If they knew these trends, they might deduce what would happen when their cohort enters the market. In fact, even if many of the students did not bother collecting the relevant information, *someone would!* The data could be sold to prospective engineers, who would be willing to pay for very valuable information about their future earnings.

In short, the cobwebs are generated because students are misinformed. They do not take into account the wage trends in the engineering labor market when choosing a career. If students had better information about the past booms and busts, they would be more hesitant to enter the career when current wages are high and more willing to enter when current wages are low. And the cobweb model would unravel.

There is evidence of cobwebs in many professional markets, suggesting that students do systematically misforecast future earnings opportunities.[37] In fact, it's not just students who misforecast the future; even professionals find it difficult to predict future earnings opportunities.[38] This inherent uncertainty forces decision-makers to place too heavy a weight on the wages they currently observe, helping generate cobwebs.

4-9 Monopsony

Up to this point, we have analyzed the properties of equilibrium in competitive labor markets. Each firm in the industry faces the same price p when trying to sell its output, regardless of how much output it sells. And each firm in the industry pays the going wage w to all workers, regardless of how many workers it hires.

A **monopsony** is a firm that faces an upward-sloping supply curve of labor.[39] In contrast to a competitive firm that can hire as much labor as it wants at the going wage, a monopsonist must pay higher wages in order to attract more workers. The one-company town, such as a coal mine in a remote location, is the stereotypical example of a monopsony. The only way the firm can induce more people to work is to raise the wage so as to meet the reservation wage of the nonworkers.

It is tempting to dismiss the relevance of the monopsony model because one-company towns are rare in mobile, industrialized economies. But a particular firm may have an upward-sloping supply curve—the key feature of a monopsony—even when it faces competition in the labor market. The circumstances that give rise to upward-sloping supply curves for seemingly competitive firms will be discussed below.

[37] Julian R. Betts, "What Do Students Know about Wages? Evidence from a Study of Undergraduates," *Journal of Human Resources* 31 (Winter 1996): 27–56; and Jeff Dominitz and Charles F. Manski, "Eliciting Student Expectations of the Returns to Schooling," *Journal of Human Resources* 31 (Winter 1996): 1–26.

[38] Jonathan Leonard, "Wage Expectations in the Labor Market: Survey Evidence on Rationality," *Review of Economics and Statistics* 64 (February 1982): 157–161.

[39] A detailed examination of the monopsony model is given by Alan Manning, "A Generalised Model of Monopsony," *Economic Journal* 116 (January 2006): 84–100.

Perfectly Discriminating Monopsonist

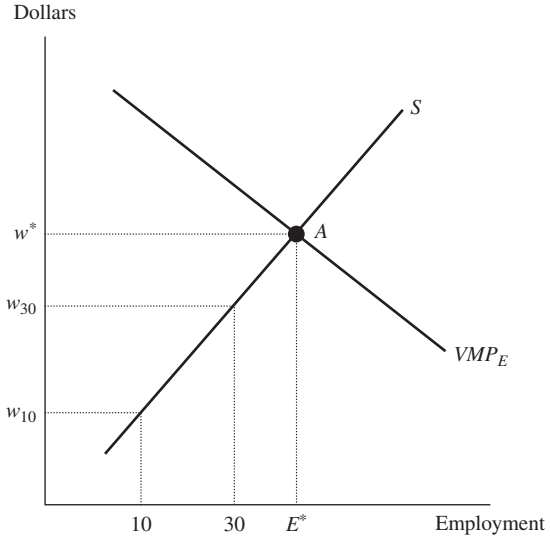
We first consider a **perfectly discriminating monopsonist,** a monopsonist that can get away with paying different wages to different workers. Figure 4-20 illustrates the labor market conditions faced by this firm. The monopsonist faces the upward-sloping labor supply curve given by *S*. This labor supply curve implies that the monopsonist need only pay a wage of w_{10} to attract the 10th worker, and must pay a wage of w_{30} to attract the 30th worker. As a result, the supply curve of labor is identical to the marginal cost of hiring labor.

Because a monopsonist cannot influence prices in the output market, it can sell as much output as it wants at a constant price p. The revenue from hiring an extra worker equals the price times the marginal product of labor, or the value of marginal product. Hence, the labor demand curve for the monopsonist, as for a competitive firm, is given by the value of marginal product curve.

Regardless of whether firms operate in a competitive market or not, profit maximization requires that the dollar value of the last worker hired equals the cost of hiring that last worker. A perfectly discriminating monopsonist will then hire up to the point where the last worker's contribution to firm revenue (or VMP_E) equals the marginal cost of labor. This equilibrium occurs at point *A*, where supply equals demand. The perfectly discriminating monopsonist hires E^* workers, exactly the same employment that would have been observed if the labor markets were competitive. The wage w^*, however, is *not* the competitive wage. It is instead the wage that the monopsonist must pay to attract the last worker hired. All other workers receive lower wages, with each worker receiving her reservation wage.

FIGURE 4-20 The Hiring Decision of a Perfectly Discriminating Monopsonist

A perfectly discriminating monopsonist faces an upward-sloping supply curve and can hire different workers at different wages. The labor supply curve gives the marginal cost of hiring. Profit maximization occurs at point *A*. The monopsonist hires the same number of workers as a competitive market, but each worker gets paid his reservation wage.



Nondiscriminating Monopsonist

The second type of monopsony is a **nondiscriminating monopsonist,** a monopsonist that must pay all workers the same wage, regardless of the worker's reservation wage. Because the nondiscriminating monopsonist must raise the wage to all workers when he wishes to hire one more worker, the labor supply curve no longer gives the marginal cost of hiring.

The numerical example in Table 4-3 illustrates this point. At a wage of \$4, no one is willing to work. At a wage of \$5, the firm attracts one worker, total labor costs equal \$5, and the marginal cost of hiring that worker is \$5. If the firm wishes to hire two workers, it must raise the wage to \$6. Total labor costs then equal \$12, and the marginal cost of hiring the second worker increases to \$7. As the firm expands, therefore, it incurs an ever-higher marginal cost.

Figure 4-21 illustrates the relation between the labor supply curve and the marginal cost of hiring curve for a nondiscriminating monopsonist. Because wages rise as the monopsonist tries to hire more workers, the marginal cost of labor curve (MC_E) is upward sloping, rises even faster than the wage, and lies above the supply curve. The marginal cost of hiring involves not only the high wage paid to the additional worker, but also the expense of paying that higher wage to all previously hired workers.[40]

The profit-maximizing monopsonist hires at the point where the marginal cost of labor equals the value of marginal product, or point A in the figure. If the monopsonist hires fewer than E_M workers, the value of marginal product exceeds the marginal cost of labor, and the firm should expand. But if the monopsonist hires more than E_M workers, the marginal cost of labor exceeds the contribution of those workers and the monopsonist should lay off some employees. The profit-maximizing condition for a nondiscriminating monopsonist is then given by

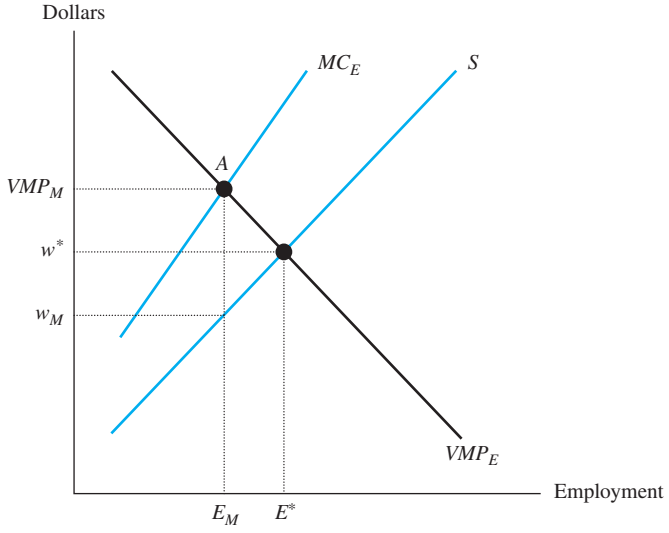
$$MC_E = VMP_E \qquad \textbf{(4-9)}$$

TABLE 4-3
Calculating the Marginal Cost of Hiring for a Nondiscriminating Monopsonist

Wage (w)	Number of Persons Willing to Work at that Wage (E)	$w \times E$	Marginal Cost of Labor
\$4	0	\$0	—
\$5	1	5	\$5
\$6	2	12	7
\$7	3	21	9
\$8	4	32	11

[40] Using calculus, it can be shown that the relationship between the wage and the marginal cost of hiring is given by $MC_E = w\left(1 + \frac{1}{\sigma}\right)$, where σ is the labor supply elasticity (that is, the percentage change in quantity supplied for a given percentage change in the wage). A competitive firm faces a perfectly elastic labor supply curve, so that the labor supply elasticity is infinite and the marginal cost of labor equals the wage. If the labor supply curve is upward sloping, the elasticity of labor supply will be positive and the marginal cost of labor exceeds the wage. Note that the labor supply elasticity that is of interest in the monopsony context, which measures the rate at which the firm must increase wages to attract more workers, is conceptually different from the labor supply elasticity giving the relation between hours of work and wages for an individual worker. As a result, the evidence on labor supply elasticities summarized in the chapter on labor supply is of little use for determining the degree of monopsony power enjoyed by particular firms.

FIGURE 4-21 The Hiring Decision of a Nondiscriminating Monopsonist
A nondiscriminating monopsonist pays the same wage to all workers. The marginal cost of hiring exceeds the wage, and the marginal cost curve lies above the supply curve. Profit maximization occurs at point A; the monopsonist hires E_M workers and pays them a wage of w_M.



Note that the labor supply curve indicates that the monopsonist need only pay a wage of w_M to attract E_M workers to the firm.

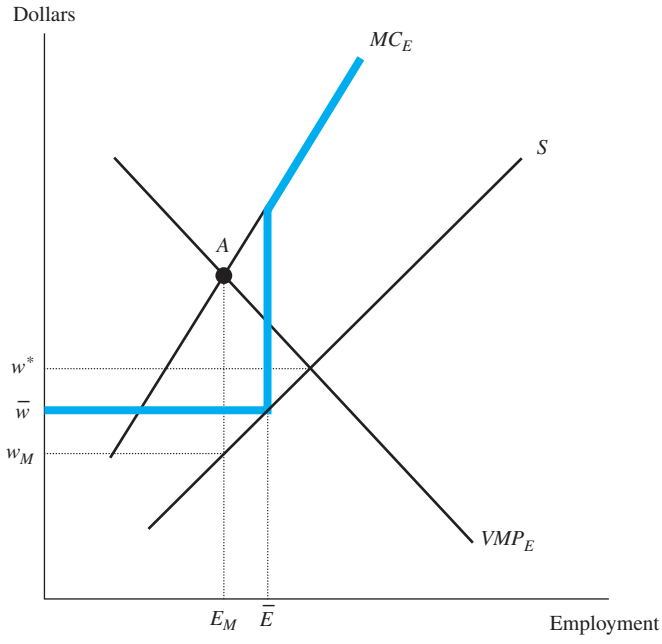
The labor market equilibrium in Figure 4-21 has two important properties. First, a nondiscriminating monopsonist employs fewer workers than would be employed if the market were competitive. The competitive level of employment is given by the intersection of supply and demand, or E^* workers. As a result, there is underemployment in a monopsony. The allocation of resources in a nondiscriminating monopsony is not efficient.

Second, the monopsony wage w_M is less than the competitive wage, w^*, and is also less than the value of the worker's marginal product, VMP_M. In a monopsony, therefore, workers are paid less than their value of marginal product and are, in this sense, "exploited."

Monopsony and the Minimum Wage

The imposition of a minimum wage in a nondiscriminating monopsony can *increase* both wages and employment. Figure 4-22 shows the monopsonist initially in equilibrium at point A, hiring E_M workers at a wage of w_M. Suppose the government imposes a wage floor of $\overline{w}$. The firm can now hire up to $\overline{E}$ workers at the minimum wage because these workers were all willing to work at wage equal to or below the minimum. In other words, the marginal cost of labor is equal to the minimum wage as long as the firm hires up to $\overline{E}$ workers. If the firm wants to hire more than $\overline{E}$ workers, however, the marginal cost of hiring reverts back to its old level MC_E (because the monopsonist must pay more than the minimum wage to attract the marginal worker as well as raise the wage of all previously hired workers). The marginal cost of labor curve, therefore, is given by the bold line in the figure: a perfectly elastic segment up to $\overline{E}$ and the upward-rising segment beyond that threshold.

FIGURE 4-22 The Minimum Wage in a Nondiscriminating Monopsony
The minimum wage may increase both wages and employment when imposed on a monopsonist. A minimum wage set at $\overline{w}$ increases employment to $\overline{\text{E}}$.



A profit-maximizing monopsonist will still want to equate the marginal cost of hiring with the value of marginal product. As drawn in Figure 4-22, the monopsonist hires $\overline{\text{E}}$ workers and pays them the minimum wage. Note that the minimum wage legislation increased both employment (from E_M to $\overline{\text{E}}$) and the wage (from w_M to $\overline{w}$). Moreover, there is no unemployment. Everyone who is looking for a job at the minimum wage $\overline{w}$ can find one.

In fact, the figure suggests that the government can do even better. It could set the minimum wage at the competitive level w^* (where supply equals demand). The monopsony would then employ the same number of workers that would be employed if the market were competitive, workers would be paid the competitive wage, and there would be no unemployment. A well-designed minimum wage, therefore, can completely eliminate the power of monopsonists and prevent the exploitation of workers.

The chapter on labor demand summarized evidence suggesting that minimum wage increases may not reduce the number of persons employed in the fast-food industry.[41] In fact, some of the evidence suggested that fast-food establishments *increased* their employment after a minimum wage increase. It has been proposed that these positive employment effects may have occurred because the fast-food industry can be viewed as a monopsony in terms of employing low-skilled teenage labor. Because these workers have few other alternatives, fast-food restaurants could perhaps provide the "one-company" environment necessary for a monopsony.

[41] David Card and Alan B. Krueger, *Myth and Measurement: The New Economics of the Minimum Wage,* Princeton, NJ: Princeton University Press, 1997.

Do Many Firms Have Upward-Sloping Labor Supply Curves?

The one-company town is the classic example of a market with an upward-sloping labor supply curve. If this firm wishes to expand, it has to raise the wage to attract more people into the workforce. This situation gives "monopsony power" to the single firm in the industry: the ability to pay its workers less than the value of marginal product, allowing the firm to make excess profits.

It turns out that individual firms might have some monopsony power even when there are many firms competing for the same type of labor. We have argued that one channel through which a competitive equilibrium is eventually attained is worker mobility—workers moving across firms to take advantage of better job opportunities. When some firms pay relatively high wages, the mobility of workers across firms reduces the wage gap and eventually equilibrates wages throughout the economy. The "law of one price" depends crucially on the assumption that workers can costlessly move from one job to another.

But it is costly to switch jobs. These costs are incurred as workers search for other jobs and as the workers move themselves and their families to unfamiliar economic and social environments. The moving costs imply that it does not make sense for a worker to accept every better-paying job offer that comes along. After all, the moving costs could well exceed the pay increase. Put differently, moving costs introduce inertia into the labor market. A firm wishing to hire more workers may have to pay a wage premium to induce workers already employed in other firms to quit those jobs, incur the moving costs, and join the firm. In effect, moving costs generate an upward-sloping supply curve for a firm. A firm wishing to hire more and more workers will have to keep raising its wage to compensate workers for the cost of switching jobs.

A firm may also have an upward-sloping supply curve if the employer finds it harder to monitor workers as it expands. The larger the firm and the more workers it employs, the larger the possibilities for workers to "shirk" their responsibilities on the job and for that shirking to go undetected. Employers know that instead of wading through a spreadsheet searching for errors, most of us would prefer to constantly check our mobile phones for the latest gossip in social media.

One possible solution to this monitoring problem is to offer workers a high wage. This high wage would make workers realize that they have much to lose if they are caught shirking and are fired. As the firm expands employment and finds it more difficult to monitor its workers, therefore, the firm may pay a higher wage to keep its workers in line. In fact, larger firms do pay higher wages.[42]

The crucial insight to draw from this discussion is that upward-sloping supply curves for particular firms may arise even when there are many firms competing for the same workers. In short, many firms in seemingly competitive markets could have some degree of monopsony power.

The insight that monopsony power need not be restricted to the extreme case of a one-company town has inspired research that attempts to estimate the labor supply elasticity to

[42] Charles Brown, James Hamilton, and James Medoff, *Employers Large and Small*, Cambridge, MA: Harvard University Press, 1990.

a given firm. A recent study, for instance, examines how the supply of registered nurses (RNs) to a particular hospital responds to changes in the RN wage.[43]

Before 1991, the U.S. Department of Veteran Affairs (VA) used a national pay scale that roughly determined RN wages in all of its facilities, regardless of whether those facilities were located in high or low cost-of-living areas. As an example, the starting RN hourly wage in Milwaukee in 1990 was $11.20 in non-VA hospitals and $11.65 in VA hospitals, so that the VA wage offer was quite competitive. In contrast, the starting RN hourly wage in San Francisco was $16.30, but the VA starting wage lagged far behind at $14.00. This policy obviously affected the VA's ability to recruit nurses in high-wage regions.

The Nurse Pay Act of 1990 addressed the problem by changing how the VA set wages in local facilities. In particular, the legislation tied the VA wage offer to the wage that prevailed in the local labor market. If the wage in VA hospitals was below the prevailing wage, the RN wage in the VA hospital would be raised immediately. However, if the wage in VA hospitals was above the prevailing wage, the VA wage would be held constant in nominal terms until the two wages reached parity.

In other words, the legislation mandated a rapid wage increase for nurses in VA hospitals in San Francisco, presumably attracting many new nurses to those facilities, but little wage change in VA hospitals in Milwaukee, where the supply of RNs would presumably remain constant.

The difference-in-difference exercise reported in Table 4-4 illustrates how we can use the Nurse Pay Act as an instrument to estimate the labor supply elasticity to VA hospitals. Between 1990 and 1992, the wage of RNs increased by 12.5 percent in VA hospitals and by 9.9 percent in non-VA hospitals, or a difference of 2.6 percentage points. At the same time, the number of RNs working at VA hospitals increased by 8.3 percent, as compared to an increase of only 5.6 percent in non-VA hospitals, or a difference of 2.7 percentage points. The labor supply elasticity is given by the ratio of the percent change in the number of workers employed to the percent change in the wage, or 2.7 ÷ 2.6, or about 1.0. In other words, a 1 percent increase in the wage offered by VA hospitals pay attracts 1 percent more nurses to those hospitals.[44]

TABLE 4-4

RN Wages and Employment, 1990–1992

Source: Douglas O. Staiger, Joanne Spetz, and Ciaran S. Phibbs, "Is There Monopsony in the Labor Market? Evidence from a Natural Experiment," *Journal of Labor Economics* 28 (April 2010), p. 223.

	VA Hospitals	Non-VA Hospitals
Percent change in wage	12.5	9.9
Percent change in RN employment	8.3	5.6

[43] Douglas O. Staiger, Joanne Spetz, and Ciaran S. Phibbs, "Is There Monopsony in the Labor Market? Evidence from a Natural Experiment," *Journal of Labor Economics* 28 (April 2010): 211–236. Good summaries of this more generalized approach to monopsony are given by Alan Manning, *Monopsony in Motion.* Princeton, NJ: Princeton University Press, 2003; and Orley C. Ashenfelter, Henry Farber, and Michael R. Ransom, "Labor Market Monopsony," *Journal of Labor Economics* 28 (April 2010): 203–210.

[44] Related studies that estimate the labor supply elasticity to specific firms include Torberg Falch, "The Elasticity of Labor Supply at the Establishment Level," *Journal of Labor Economics* 28 (April 2010): 237–266; and Michael Ransom and David P. Sims, "Estimating the Firm's Labor Supply Curve in a "New Monopsony" Framework: School Teachers in Missouri," *Journal of Labor Economics* 28 (April 2010): 331–355.

Summary

- A competitive equilibrium leads to an efficient allocation of resources. No other allocation of workers to firms generates higher gains from trade.
- A fraction of the payroll taxes imposed on firms is passed on to workers. The more inelastic the labor supply curve, the higher the fraction of payroll taxes that is shifted to workers. A payroll tax has the same impact on wages and employment regardless of whether it is imposed on workers or on firms.
- The payroll tax creates a deadweight loss.
- In the short run, immigration reduces the wage of workers who have skills similar to those of immigrants and increases the wages of workers who have skills that complement those of immigrants. In the long run, these wage effects are attenuated as the capital stock adjusts to the presence of immigrants.
- In the short run, immigration redistributes wealth from workers to employers, but the net income of natives increases.
- Markets for professional workers are sometimes characterized by systematic booms and busts, or cobwebs.
- A nondiscriminating monopsonist hires fewer workers than would be hired in a competitive labor market and pays them a lower wage.
- The imposition of a minimum wage on a monopsony can increase both the wage and the number of workers employed.
- A particular firm may have some monopsony power, even in labor markets that may seem competitive, when workers find it costly to move across firms.

Key Concepts

constant returns to scale, *141*
cobweb model, *158*
deadweight loss, *133*
efficient allocation, *124*
gains from trade, *124*
human capital externalities, *152*
immigration surplus, *150*
invisible hand theorem, *122*
mandated benefits, *136*
monopsony, *159*
nondiscriminating monopsony, *161*
perfectly discriminating monopsony, *160*
producer surplus, *124*
worker surplus, *124*

Review Questions

1. What is the producer surplus? What is the worker surplus? Show that a competitive market equilibrium maximizes the gains from trade.
2. Discuss the implications of equilibrium for a competitive economy containing many regional markets when labor and firms are free to enter and exit the various markets. Why is the resulting allocation of labor efficient?
3. Show what happens to producer surplus, worker surplus, and the gains from trade as workers migrate from a low-wage to a high-wage region.

4. Describe the impact of a payroll tax on wages and employment in a competitive industry. Why is part of the tax shifted to workers? What is the deadweight loss of the payroll tax?
5. Why does the payroll tax have the same impact on wages and employment regardless of whether it is imposed on workers or on firms?
6. How do mandated benefits affect labor market outcomes? Why do these outcomes differ from those resulting from a payroll tax? What is the deadweight loss arising from mandated benefits?
7. What is the short-run impact of immigration on the wage of native workers? What is the long-run impact?
8. What is the immigration surplus and how is it affected by human capital externalities?
9. Describe the trends in wages and employment implied by the cobweb model for the engineering market. What would happen to the cobwebs if a consulting firm sold information on the history of wages and employment in the engineering market?
10. Describe the hiring decision of a perfectly discriminating monopsonist and of a nondiscriminating monopsonist. In what sense do monopsonists "exploit" workers?
11. Show how the imposition of a minimum wage on a monopsony can increase both wages and employment.

Problems

4-1. Figure 4-9 discusses the changes to a labor market equilibrium when the government mandates an employee benefit for which the cost exceeds the worker's valuation (panel a) and for which the cost equals the worker's valuation (panel b).

(a) Provide a similar graph to those in Figure 4-9 when the cost of the benefit is less than the worker's valuation and discuss how the equilibrium level of employment and wages has change. Is there deadweight loss associated with the mandated benefit?

(b) Why is the situation in part (a) in which a mandated benefit would cost less than the worker's valuation less important for public policy purposes than when the cost of the mandated benefit exceeds the worker's valuation?

4-2. In the United States, labor supply tends to be inelastic relative to labor demand, and according to law, payroll taxes are essentially assessed evenly between workers and firms. Given the above situation, are workers or firms more likely to bear the additional burden of an increased payroll tax in the United States? Could this burden be shifted to the firms by assessing the increase in payroll taxes on just firms rather than having firms and workers continue to be assessed payroll taxes equally?

4-3. Suppose the supply curve of physicists is given by $w = 10 + 5E$, while the demand curve is given by $w = 50 - 3E$. Calculate the equilibrium wage and employment level. Suppose now that the demand for physicists increases to $w = 70 - 3E$. Assume the market is subject to cobwebs. Calculate the wage and employment level in each round as the wage and employment levels adjust to the demand shock. What are the new equilibrium wage and employment level?

4-4. Suppose labor demand for low-skilled workers in the United States is $w = 24 - 0.1E$ where E is the number of workers (in millions) and w is the hourly wage. There are

120 million domestic U.S. low-skilled workers who supply labor inelastically. If the U.S. opened its borders to immigration, 20 million low-skill immigrants would enter the U.S. and supply labor inelastically. What is the market-clearing wage if immigration is not allowed? What is the market-clearing wage with open borders? How much is the immigration surplus when the U.S. opens its borders? How much surplus is transferred from domestic workers to domestic firms?

4-5. There are two reasons why the immigration surplus exists when immigration is accompanied by human capital externalities. Both reasons are evident in Figure 4-16. The first is represented by triangle *BCD*. The second is represented by trapezoid *ABEF*. Explain the underlying source of each area. Explain why human capital externalities are important.

4-6. Let total market demand for labor be represented by $E_D = 1{,}000 - 50w$ where E_D is total employment and w is the hourly wage.

(a) What is the market clearing wage when total labor supply is represented by $E_S = 100w - 800$? How many workers are employed? How much producer surplus is received at the equilibrium wage?

(b) Suppose the government imposes a minimum wage of $16. What is the new level of employment? How much producer surplus is received under the minimum wage?

4-7. Let total market demand for labor be represented by $E_D = 1{,}200 - 30w$ where E_D is total employment and w is the hourly wage. Suppose 750 workers supply their labor to the market perfectly inelastically. How many workers will be employed? What will be the market clearing wage? How much producer surplus is received?

4-8. A firm faces perfectly elastic demand for its output at a price of $6 per unit of output. The firm, however, faces an upward-sloped labor supply curve of

$$E = 20w - 120$$

where E is the number of workers hired each hour and w is the hourly wage rate. Thus, the firm faces an upward-sloped marginal cost of labor curve of

$$MC_E = 6 + 0.1E$$

Each hour of labor produces five units of output. How many workers should the firm hire each hour to maximize profits? What wage will the firm pay? What are the firm's hourly profits?

4-9. Ann owns a lawn mowing company. She has 400 lawns she needs to cut each week. Her weekly revenue from these 400 lawns is $20,000. If given an 18-inch deck push mower, a laborer can cut each lawn in two hours. If given a 60-inch deck riding mower, a laborer can cut each lawn in 30 minutes. Labor is supplied inelastically at $10 per hour. Each laborer works 8 hours a day and 5 days each week.

(a) If Ann decides to have her workers use push mowers, how many push mowers will Ann rent and how many workers will she hire?

(b) If she decides to have her workers use riding mowers, how many riding mowers will Ann rent and how many workers will she hire?

(c) Suppose the weekly rental cost (including gas and maintenance) for each push mower is $250 and for each riding mower is $2,400. What equipment will Ann rent? How many workers will she employ? How much profit will she earn?

(d) Suppose the government imposes a 20 percent payroll tax (paid by employers) on all labor and offers a 20 percent subsidy on the rental cost of capital. What equipment will Ann rent? How many workers will she employ? How much profit will she earn?

4-10. Figure 4-6 shows that a payroll tax will be completely shifted to workers when the labor supply curve is perfectly inelastic. In this case, for example, a new $2 payroll tax will lower the wage by $2, will not affect employment, and will not result in any deadweight loss. Suppose instead that labor supply is perfectly elastic at a wage of $10. In this case, what would be the effect on wages, employment, and deadweight loss from a $2 payroll tax?

4-11. In the Cobweb model of labor market equilibrium (Figure 4-19), the adjustments in employment can be small with adjustment being fast, or the adjustments in employment can be large with adjustment being slow. The result that comes about depends on the elasticity of labor supply. Which result (small and fast versus large and slow) is associated with very inelastic labor supply? Which result is associated with elastic labor supply? What is the economic intuition behind this result?

4-12. A monopsonist's demand for labor can be written as $VMP_E = 40 - 0.005E_D$. Labor is supplied to the firm according to $w = 5 + 0.01E_S$. Thus, the firm's marginal cost of hiring workers when it hires off of this supply schedule is $MC_E = 5 + 0.02E_S$.

(a) How much labor does the monopsony firm hire and at what wage when there is no minimum wage?

(b) How much labor does the monopsony firm hire and at what wage when it must pay a minimum wage of $25?

4-13. Suppose the economy's labor market is competitive and that labor demand can be written as $w = 50 - 0.3E$ while labor supply can be written as $w = 8 + 0.2E$ where E is the total amount of employment in millions. What is the market clearly wage? How many people are employed? What is the total value of producer surplus? What is the total amount of worker surplus?

4-14. Suppose the Cobb-Douglas production function given in equation 4-1 applies to a developing country. Instead of thinking of immigration from a developing to a developed country, suppose a developed country invests large amounts of capital (foreign direct investment, or FDI) in a developing country.

(a) How does an increase in FDI affect labor productivity in the developing country? How will wages respond in the short-run?

(b) What are the long-run implications of FDI, especially in terms of potential future immigration from the developing country?

4-15. Empirical work suggests that labor demand is very elastic while labor supply is very inelastic. Assume too that payroll taxes are about 15% and legislated to be paid half by the employee and half by the employer.

(a) What would happen to worker wages if payroll taxes were eliminated?

(b) What would happen to employment costs paid by firms if payroll taxes were eliminated?

(c) What would happen to producer and worker surplus if payroll taxes were eliminated? Which measure is relatively more sensitive to payroll taxes? Why?

(d) Why might workers not want payroll taxes eliminated?

Selected Readings

Joshua D. Angrist, "Short-Run Demand for Palestinian Labor," *Journal of Labor Economics* 14 (July 1996): 425–453.

David H. Autor, David Dorn, and Gordon H. Hanson, "The China Syndrome: Local Labor Market Effects of Import Competition in the United States," *American Economic Review* 103 (October 2013): 2121–2168.

Ariel R. Belasen and Solomon W. Polachek, "How Disasters Affect Local Labor Markets: The Effects of Hurricanes in Florida," *Journal of Human Resources* 44 (Winter 2009): 251–276.

Olivier Jean Blanchard and Lawrence F. Katz, "Regional Evolutions," *Brookings Papers on Economic Activity* 1 (1992): 1–61.

George J. Borjas, "The Labor Demand Curve Is Downward Sloping: Reexamining the Impact of Immigration in the Labor Market," *Quarterly Journal of Economics* 118 (November 2003): 1335–1374

David Card, "The Impact of the Mariel Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review* 43 (January 1990): 245–257.

Jonathan Gruber, "The Incidence of Payroll Taxation: Evidence from Chile," *Journal of Labor Economics* 15 (July 1997, Part 2): S102–S135.

Prachi Mishra, "Emigration and Wages in Source Countries: Evidence from Mexico," *Journal of Development Economics* 82 (January 2007): 180–199.

Douglas O. Staiger, Joanne Spetz, and Ciaran S. Phibbs, "Is There Monopsony in the Labor Market? Evidence from a Natural Experiment," *Journal of Labor Economics* 28 (April 2010): 211–236.

Chapter 5

Compensating Wage Differentials

It's just a job. Grass grows, birds fly, waves pound the sand. I beat people up.
—*Muhammad Ali*

The free entry and exit of workers and firms in a competitive labor market leads to a single wage equilibrium *as long as all jobs are alike and all workers are alike.*

But workers are different and jobs are different, so the real-world labor market is not characterized by a single wage. Workers differ in their skills. And jobs differ in the amenities they offer. Some jobs, for instance, are located in sunny California, and others in the tundras of Alaska; some expose workers to dangerous chemicals, others to the wonders of delicious chocolates and gourmet meals.

Because workers care about whether they work in California or the Arctic and whether they work amid toxic waste or in a luxurious French restaurant, we should think of a job offer not just in terms of how much money the job pays, but in terms of the entire job package that includes both wages and working conditions. This chapter examines the impact of differences in job amenities on the determination of wages and employment.

The idea that job characteristics matter was first raised by Adam Smith in 1776. In the first statement of what labor market equilibrium is about, Smith argued that **compensating wage differentials** arise to compensate workers for the nonwage characteristics of jobs. As Smith put it in a renowned passage of *The Wealth of Nations:*[1]

> The whole of the advantages and disadvantages of different employment of labour and stock must, in the same neighbourhood, be either perfectly equal or continually tending to equality. If in the same neighbourhood there was any employment either evidently more or less advantageous than the rest, so many people would crowd into it in the one case, and so many would desert it in the other, that its advantages would soon return to the level of other employments. This at least would be the case in a society where things were left to follow their rational course, where there was perfect liberty and where every-man was perfectly free both to choose what occupation he thought proper, and to change it as often as he thought proper.

[1] Adam Smith, *The Wealth of Nations,* Chicago: University of Chicago Press, 1976 (1776), p. 111.

It is not the wage that is equated across jobs in a competitive market, but the "whole of the advantages and disadvantages" of the job. Firms that have unpleasant working conditions must offer some offsetting advantages (such as a higher wage) to attract workers; firms that offer pleasant working conditions can get away with paying lower wages (in effect, making workers pay for the pleasant environment).

The nature of labor market equilibrium in the presence of compensating wage differentials differs radically from the equilibrium implied by the traditional supply–demand framework. In the traditional model, the wage guides the allocation of workers across firms so as to achieve an efficient allocation of resources. Workers and firms move to whichever market offers them the best opportunities, equating wages, and the value of marginal product in the process. In the end, workers and firms are anonymous and it does not matter who works where.

The introduction of compensating differentials breaks the anonymity. Workers differ in their preferences for job characteristics and firms differ in the working conditions they offer. The theory of compensating differentials tells a story of how workers and firms "match and mate" in the labor market. Workers who are looking for a particular set of job amenities search out those firms that provide it. As a result, the allocation of labor to firms is not random and it matters who works where.

The theory of compensating wage differentials also provides a starting point for analyzing one of the central questions in economics: Why do different workers get paid differently? This chapter focuses on the role played by the characteristics of jobs. Some of the remaining chapters will focus on the role played by the characteristics of workers.

5-1 The Market for Risky Jobs

We begin by deriving a compensating wage differential in a very simple, and policy-relevant, context.[2] There are only two types of jobs, and they are differentiated by the probability of getting injured on the job, denoted by ρ. Some jobs offer a completely safe environment, and the probability of injury is equal to zero. Other jobs have an inherently risky environment, and the probability of injury equals one.

Suppose the worker has complete information about the risk associated with every job. In other words, the worker knows if the job is "safe" ($\rho = 0$) or "risky" ($\rho = 1$). This is an important assumption because the risks may not be detectable for years. Prior to the 1960s, for example, asbestos products were used regularly in the construction industry. Few people knew that continuous exposure to asbestos had adverse health effects. It took a long time for the scientific evidence to accumulate. We will discuss below how the analysis is affected when the worker is not fully informed about the value of ρ.

Workers care about the wage (w) they earn on the job. And they also care about whether they will get hurt. We write the worker's utility function as

$$\text{Utility} = f(w, \rho) \qquad \textbf{(5-1)}$$

[2] Sherwin Rosen, "The Theory of Equalizing Differences," in Orley C. Ashenfelter and Richard Layard, editors, *Handbook of Labor Economics,* vol. 1, Amsterdam: Elsevier, 1986, pp. 641–692.

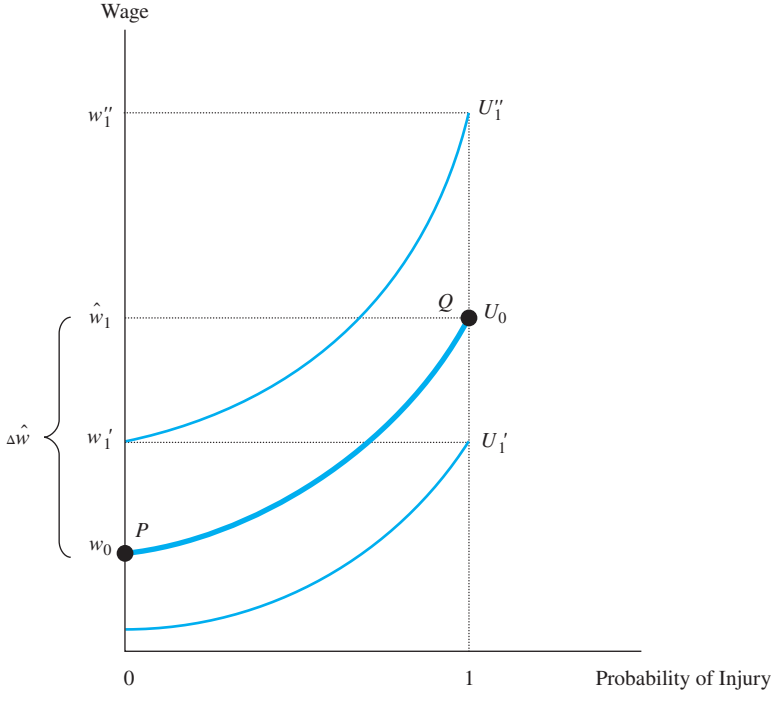
The marginal utility of income gives the change in utility resulting from a $1 increase in the worker's income, holding constant the probability of injury. We assume that workers prefer higher wages, so that the marginal utility of income is positive. The marginal utility of risk gives the change in utility resulting from a one-unit change in the probability of injury, holding constant the worker's income. Assume initially that the marginal utility of risk is negative. Most of us would probably get less utility from working at a job where we are more likely to get hurt.[3]

Suppose the safe job offers a wage rate of w_0 dollars. Figure 5-1 illustrates the worker's indifference curve (U_0) that goes through the point summarizing the employment package offered by the safe job. At point *P,* the worker gets a wage of w_0 and has a zero probability of injury.

The indifference curves describing the trade-off between the wage and the probability of injury must be upward sloping because risk is a "bad." Suppose that the worker is at point *P.* The only way to persuade the worker to switch to the riskier job *and* hold her

FIGURE 5-1 Indifference Curves between the Wage and the Probability of Injury

The worker earns a wage of w_0 and gets U_0 utils if she chooses the safe job. She would prefer the safe job if the risky job paid a wage of w_1', but would prefer the risky job if that job paid a wage of w_1''. The worker is indifferent if the risky job pays $\hat{w}_1$. The worker's reservation price is $\Delta\hat{w} = \hat{w}_1 - w_0$.



[3] Some people may enjoy working in a risky environment and the marginal utility of risk would be positive for those workers. We ignore the existence of "risk lovers" until later in the discussion.

utility constant is by increasing her wage. She would obviously be worse off if she moved to a riskier job and her wage fell. The curvature of the indifference curve reflects the usual assumption that indifference curves are convex.

The Supply of Labor to Risky Jobs

The indifference curve U_0 tells us how much the worker dislikes being injured. She would obviously prefer working in the safe job if the risky job only paid w_1'. Her utility in the safe job (U_0) would then exceed her utility in the risky job U_1'. Similarly, she would prefer working in the risky job if the risky job paid w_1''. Her utility would then increase to U_1''. The worker, however, would be indifferent between the two jobs if the risky job paid the wage $\hat{w}_1$.

We define the worker's **reservation price** as the amount of money it would take to bribe her into accepting the risky job—or the difference $\Delta\hat{w} = \hat{w}_1 - w_0$. If the worker's income were to increase by $\Delta\hat{w}$ dollars as she switched from the safe job to the risky job, she would not care about being exposed to the additional risk. The reservation price, therefore, is the answer to the age-old question: How much would it take for you to do something that you would rather not do?

Different workers have different attitudes toward risk. Depending on how we draw a worker's indifference curves, the reservation price $\Delta\hat{w}$ could be a small number or a large number. If the indifference curves between income and risk were relatively flat, $\Delta\hat{w}$ would be small. If the indifference curves were steep, $\Delta\hat{w}$ would be large. The greater the worker's dislike for risk, the greater the bribe required for switching from the safe job to the risky job, and the greater the reservation price $\Delta\hat{w}$.

Figure 5-2 illustrates the supply curve to risky jobs. The supply curve tells us how many workers are willing to offer their services to the risky job as a function of the wage differential between the risky job and the safe job. Because we assumed that all workers dislike risk, no worker would be willing to work at the risky job when the wage differential is zero (or negative). As the wage differential rises, there will come a point where the worker who dislikes risk the least is "bought off" and decides to work in the risky job. This threshold is illustrated by the reservation price $\Delta\hat{w}_{MIN}$ in Figure 5-2. As the wage differential between the risky job and the safe job keeps increasing, more and more workers are bribed into the risky occupation, and the number of workers who choose to work in risky jobs keeps rising. The supply curve to the risky job, therefore, is upward sloping.

The Demand for Labor by Risky Firms

Just as workers decide whether to work in a safe or risky job, a firm must decide whether to provide a safe or risky work environment to its workers. The firm's choice depends on which is more profitable.

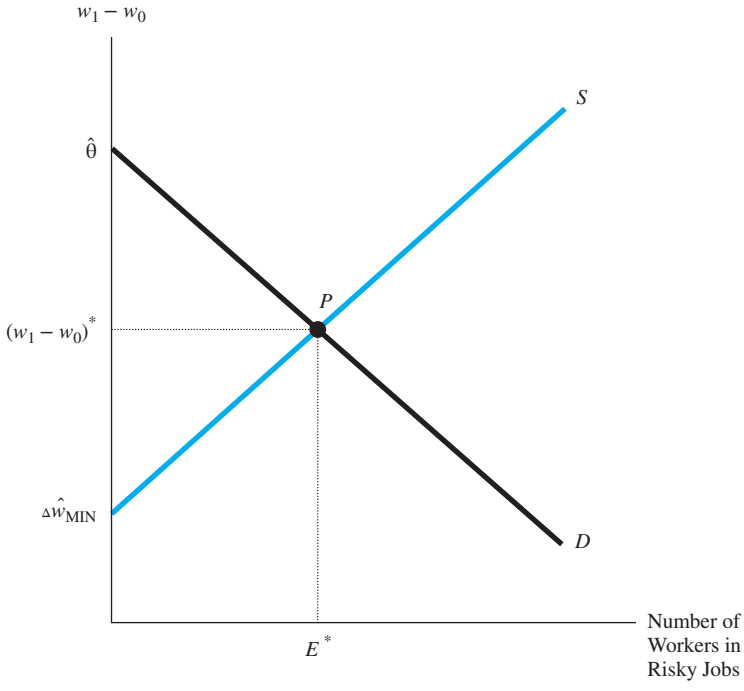
To easily illustrate the firm's decision, suppose the firm is going to hire E^* workers regardless of which environment it chooses. If the firm chooses to offer a safe work environment, the production function is

$$q_0 = \alpha_0 E^* \qquad (5\text{-}2)$$

The parameter α_0 gives the marginal product of labor in a safe environment, the increase in output when the firm hires one more worker. The value of marginal product of labor in a safe firm equals $p \times \alpha_0$, where p is the price of the output.

FIGURE 5-2 The Market Compensating Differential

The supply curve slopes up because more workers are willing to work in the risky job as the wage premium paid by the risky job rises. The demand curve slopes down because fewer firms will offer risky working conditions if risky firms have to pay very high wages to attract workers. The market compensating differential equates supply and demand and gives the bribe required to attract the last worker hired by risky firms.



If the firm offers a risky environment, the production function is

$$q_1 = \alpha_1 E^* \qquad \textbf{(5-3)}$$

where α_1 is the marginal product of labor in a risky environment. The value of marginal product of labor in a risky firm equals $p \times \alpha_1$.

We now must address a crucial question: How does the marginal product of labor differ between safe and risky environments?

Safety is not cheap. The firm must allocate labor and capital to *produce* a safe environment—diverting resources from the production of output. It takes a lot of work to remove asbestos fibers from preexisting structures or to make a building earthquake-proof, and that effort could have been put to producing more output. This diversion of resources suggests that the marginal product of labor is higher in a risky environment, so that $\alpha_1 > \alpha_0$. Note that if the marginal product of labor were higher in safe firms, we would never observe any risky firms. After all, not only would workers be more productive in safe firms, but the firm could get away with paying them lower wages because workers value safety.

The firm's profits equal the difference between the firm's revenues (the price of the output p times the output produced) and the firm's costs (the wage the firm has to pay times the number of workers hired). Profits depend on whether the firm offers a safe or a risky environment. The profits under each of the possibilities are given by

$$\pi_0 = p\,\alpha_0\,E^* - w_0\,E^* \tag{5-4}$$

$$\pi_1 = p\,\alpha_1\,E^* - w_1\,E^* \tag{5-5}$$

where π_0 gives the firm's profits if it chooses to be a safe firm, and π_1 gives the profits if it chooses to be a risky firm. Both revenues and costs are affected by the firm's decision. A risky firm has greater revenues (because more output is produced), but also incurs higher costs (because it must pay a higher wage to attract workers).

A profit-maximizing firm offers a risky environment if $\pi_1 > \pi_0$. Define $\theta = p\alpha_1 - p\alpha_0$ as the per-worker dollar gain (that is, the difference in the value of marginal product) when the firm switches from a safe to a risky environment. Algebraic manipulations of equations (5-4) and (5-5) show that the firm's decision rule is

$$\begin{aligned} &\text{Offer a safe work environment if } w_1 - w_0 > \theta \\ &\text{Offer a risky work environment if } w_1 - w_0 < \theta \end{aligned} \tag{5-6}$$

If the additional labor cost exceeds the per-worker productivity gain (or $w_1 - w_0 > \theta$), the firm is better off by offering a safe environment. If the per-worker productivity gain exceeds the additional labor cost (or $w_1 - w_0 < \theta$), the firm maximizes profits by offering a risky environment.

Different firms have different technologies for producing safety—implying that the parameter θ differs across firms. For example, universities do not have to allocate many resources to produce a safe environment for the staff, so that the per-worker productivity gain θ is small. But coal mines find it much more difficult to produce safety. The productivity gain associated with offering a risky environment in a coal mine is probably substantial and θ is large.

The market labor demand curve by risky firms is derived by "adding up" labor demand across firms. If the wage gap between risky and safe firms is very large, no firm would choose to become a risky firm and the demand for risky workers is zero. As the wage differential falls, there will come a point where the firm that has the most to gain by becoming a risky firm decides that it is worth incurring the additional labor cost. This firm has a threshold value of θ equal to $\hat{\theta}$ in Figure 5-2. As the wage differential between the risky job and the safe job keeps falling, more and more firms will find it profitable to offer a risky environment and the number of workers demanded by risky firms rises. The labor demand curve for risky jobs, therefore, is downward sloping, as illustrated in Figure 5-2.

Equilibrium

The compensating wage differential and the number of workers employed by risky firms are determined by the intersection of the supply and demand curves, as given by point P

in Figure 5-2. The compensating wage differential in the market is $(w_1 - w_0)^*$, with E^* workers employed in risky jobs. If the wage differential exceeds this equilibrium level, more persons are willing to work in risky firms than are being demanded, and the wage differential would fall. But if the wage differential fell below the equilibrium level, too few workers are willing to work in risky jobs relative to the demand, and the wage differential would rise.

Note that the market compensating wage differential $(w_1 - w_0)^*$ is positive. In equilibrium, risky jobs pay more than safe jobs. This result follows from our assumption that all workers dislike risk; if firms offering a risky environment wish to hire anyone, they will have to pay a higher wage.

It is tempting to interpret the wage differential $(w_1 - w_0)^*$ as a measure of the *average* distaste for risk among workers (that is, as a measure of the average reservation price). This interpretation, however, is not correct. The equilibrium compensating wage differential is the wage premium required to attract the *marginal* worker (that is, the last worker hired) into the risky job. In other words, the equilibrium wage differential measures the reservation price of the last worker hired and has nothing to do with the average distaste for risk in the population.

As a result, all workers except for the marginal worker are *overcompensated* by the market. Every worker but the last worker hired was willing to work at the risky job at a lower wage. A competitive labor market with fully informed workers provides more than adequate compensation for the risks that workers encounter on the job.

Equilibrium When Some Workers Like Risk

We have assumed that all workers dislike risk. But some workers may prefer to work in jobs where they face a high probability of injury. In other words, some workers (just like the helmet-less motorcyclists who fly down the highway at 100 mph) might get utility from being in situations where they can "test their courage." The reservation price for workers who like risk is negative because they are willing to pay for the right to be employed in risky jobs. The supply curve drawn in Figure 5-3 allows for the possibility that some workers have negative reservation prices and are willing to work in the risky job even though the risky job pays less than the safe job.

Suppose that the demand for workers by risky firms is very small. There are, for example, an extremely limited number of job openings for test pilots. The market demand curve could then intersect the market supply curve at a point like P in the figure, which would imply a *negative* compensating wage differential for the E^* workers employed in risky jobs. Even though almost everyone in the population dislikes risk, the demand for labor by risky firms is so small that those firms can get away with only hiring workers who are willing to pay to be in those jobs.

The equilibrium illustrated in Figure 5-3 reinforces our understanding of exactly what a compensating wage differential measures. Even though most of us would think it sensible that the theory should predict that workers employed in risky jobs should earn more than workers employed in safe jobs, it takes two to tango. If some workers are willing to pay for the right to be exposed to a risky environment, and if the demand by such jobs is sufficiently small, the market differential will go in the opposite direction.

Theory at Work

JUMPERS IN JAPAN

On March 11, 2011, a deadly earthquake measuring 9.0 on the Richter scale struck the east coast of Japan. Within minutes, tsunami waves more than 100 feet high struck the coast, and some of those waves traveled as much as six miles inland. The tsunami waves were extremely destructive, wiping out entire towns in a matter of minutes. Despite Japan's preparedness for such tragic events, there were at least 20,000 fatalities from the combined destruction of the earthquake and tsunami.

A number of cooling systems in some of the nuclear reactors that help provide electricity to Japan began to fail, and evacuations were required around the Fukushima nuclear plant, where thousands of tons of water became radioactive. The stabilization of the situation is obviously a very dangerous task, requiring workers to be exposed to radioactive material.

Because of the seriousness of the situation at the Fukushima reactor, the Tokyo Electric Power Company (TEPCO) needed workers, called "jumpers," who would dart into highly radioactive areas to conduct a particular task and leave as quickly as possible so as to minimize their exposure. TEPCO needed jumpers to help connect the equipment that would help pump out the contaminated water. In the words of a TEPCO official: "The pump could be powered from an independent generator, and all that someone would have to do is bring one end of the pump to the water and dump it in, and then run out."

Despite the simplicity of the task, the excursion was obviously very dangerous. In the words of a news report, "The radiation might be so intense that jumpers can only make one such foray in their entire lives, or risk serious radiation poisoning." TEPCO and its subcontractors began to recruit jumpers and offered eye-popping wages. The going pay for a jumper was around 200,000 yen, or roughly $2,400, for less than an hour's work. A worker's reaction summarizes the tragedy: "Ordinarily I'd consider that a dream job, but my wife was in tears and stopped me, so I declined."

Source: Terril Yue Jones, "'Jumpers' Offered Big Money to Brave Japan's Nuclear Plant," *Reuters,* April 1, 2011.

5-2 The Hedonic Wage Function

The model presented in the previous section derived the compensating wage differential in a labor market where there are only two types of firms, a risky firm and a safe firm. We now generalize the model to a market where there are many types of firms.[4] The probability of injury on the job, again denoted by ρ, can take on any value between 0 and 1.

Indifference Curves of Different Workers

For convenience, we assume that workers dislike risk. Different workers, however, dislike risk differently. Figure 5-4 illustrates the indifference curves for three different workers, A, B, and C (with associated utilities U_A, U_B, and U_C). The slope of each indifference curve tells us how much the wage would have to increase if the particular worker were to voluntarily switch to a slightly riskier job. The slope of an indifference curve, therefore, gives the reservation price that the worker attaches to moving to a slightly riskier job.

Worker A has the steepest indifference curve, and hence the highest reservation price. This worker, therefore, dislikes risk the most. At the other extreme, worker C has the flattest indifference curve and the lowest reservation price. Although worker C does not like risk, she does not mind it that much.

[4] Sherwin Rosen, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy* 82 (January–February 1974): pp. 34–55.

FIGURE 5-3 Market Equilibrium When Some Workers Prefer Risky Jobs
A worker who prefers a risky job is willing to pay for the right to work in that environment. If the demand for such workers is small, the market compensating differential may be negative. At point *P,* where supply equals demand, workers employed in risky jobs earn less than workers employed in safe jobs.

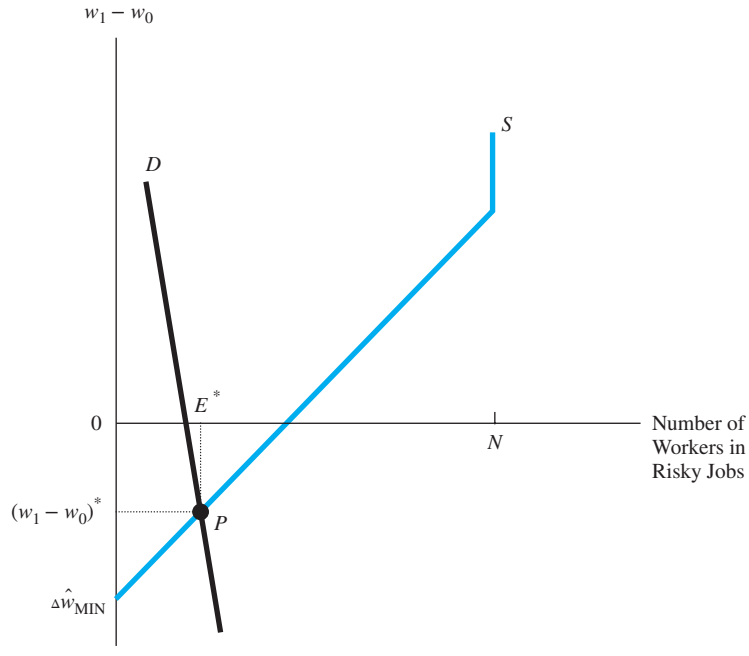
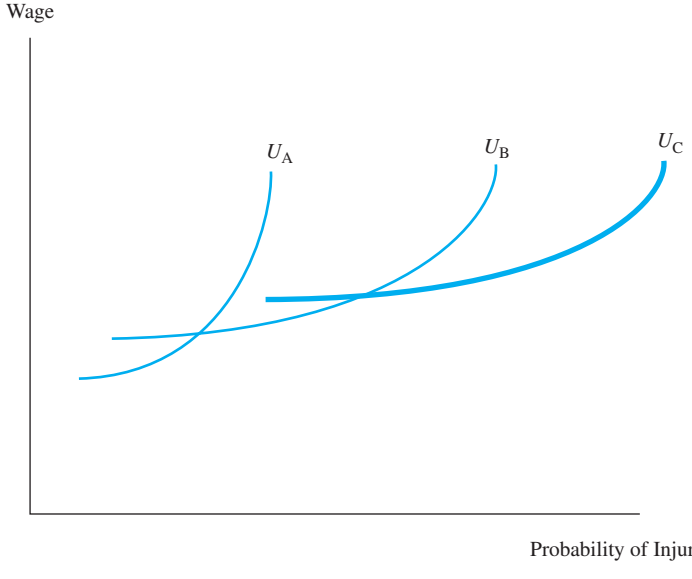


FIGURE 5-4 Indifference Curves for Different Types of Workers
Different workers have different preferences for risk. Worker A strongly dislikes a high probability of injury. Worker C does not mind it as much.



Note that the indifference curves drawn in Figure 5-4 intersect. This would seem to contradict one of our basic tenets about indifference curves. The figure, however, illustrates the indifference curves of *different workers.* Even though the indifference curves of one worker cannot intersect, the indifference curves of workers who differ in their attitudes about getting injured on the job can certainly intersect.

The Isoprofit Curve

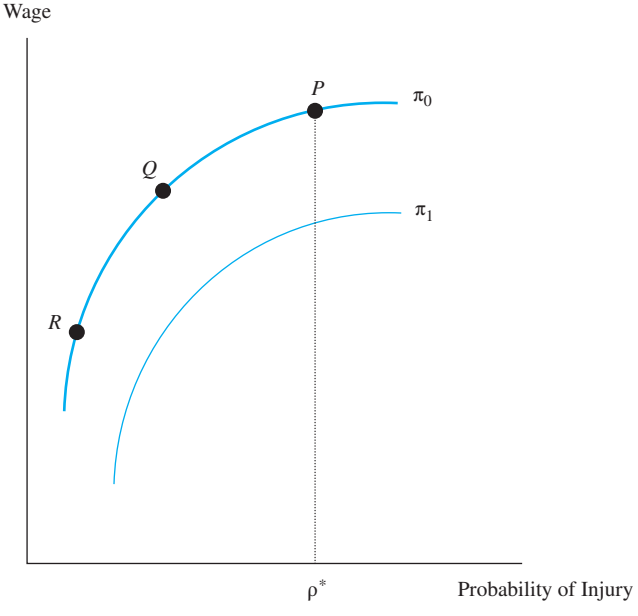
Firms compete for workers by offering different job packages, where a job package consists of a particular value of the wage w and of the probability of injury ρ. To show how firms choose a particular job package, we introduce a new concept, an **isoprofit curve.** As implied by its name, all points along an isoprofit curve yield the same profits. A profit-maximizing employer is indifferent among the various combinations of w and ρ that lie along a particular isoprofit curve.

Figure 5-5 illustrates the family of isoprofit curves for a particular employer. These curves have a number of important properties.

1. *Isoprofit curves are upward sloping because it costs money to produce safety.* Suppose the firm makes the job offer at point P on the isoprofit curve that yields π_0 profit. What must happen to the wage if the firm wants to become a safer firm *and* hold profits constant? As we noted earlier, a firm must use some resources to make the workplace safer. Profits can then be held constant only if the firm reduces the wage that it pays its workers as it tries to make the workplace safer (and moves toward point Q). Hence,

FIGURE 5-5 The Isoprofit Curve

An isoprofit curve gives all the combinations between the wage and the probability of injury that yield the same profits. It is costly to produce safety. If profits are held constant, a firm with probability ρ^* can make the workplace safer only if it reduces wages. The isoprofit curve is upward sloping and higher isoprofit curves yield lower profits.



isoprofit curves slope up. If isoprofit curves were downward sloping, it would mean that the firm could "buy" safety, raise the wage, and have the same profits. This would contradict our assumption that it is costly to produce safety.

2. *Higher isoprofit curves yield lower profits.* Points on the isoprofit labeled π_0 are less profitable than points on the π_1 isoprofit. For any probability of injury (such as ρ^* in the figure), a wage cut moves the firm to a lower isoprofit curve. This wage cut, however, increases profits.
3. *Isoprofit curves are concave.* The concavity implies that the law of diminishing returns also applies to the production of safety. Consider a firm at point *P* in the π_0 curve. The firm offers a risky work environment. There are many cheap things the firm can do to improve the safety of the workplace. To prevent injury from earthquakes, for example, the firm can nail bookcases to the wall and tighten the screws on lighting fixtures. These activities would reduce the probability of injury at little cost, so that the firm can lower ρ and hold profits constant by only slightly reducing the wage. The isoprofit curve between points *P* and *Q,* therefore, would be relatively flat. Suppose, however, that the firm wishes to make the workplace even safer. All the cheap things have already been done. To further reduce ρ to point *R,* the firm has to incur substantial expenditures. To provide additional protection from earthquakes, the firm might have to reinforce the building's foundation or move to another location. Further reductions in the probability of injury, therefore, are very costly and the firm has to greatly reduce the wage in order to hold profits constant. The segment of the isoprofit curve between points *Q* and *R,* therefore, would be steeper.

Suppose the firm operates in a competitive market with free entry and exit. When firms earn excess profits, many firms will enter the industry and depress profits. If profits were to become negative, firms would leave the industry, pushing up prices and increasing profits for the remaining firms. In the end, the only feasible trade-off between the wage and the probability of injury lies along the zero-profit isoprofit curve.

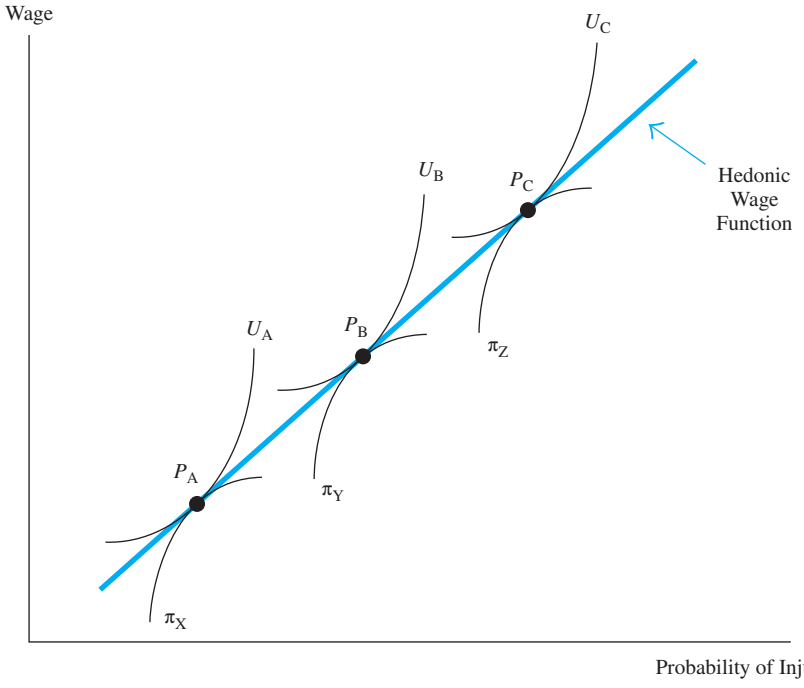
Equilibrium

The zero-profit isoprofit curve gives the menu of (w, ρ) job packages available to a particular firm. Some firms will find it easy to offer a safe environment to their workers, but other firms will find it difficult. Figure 5-6 illustrates the zero-profit isoprofit curves for three firms: π_X for firm X, π_Y for firm Y, and π_Z for firm Z. Firm X (which might be producing computer software) finds it easy to provide a very safe work environment, while firm Z (perhaps building experimental fighter planes) finds it virtually impossible to provide a safe work environment.

Workers maximize utility by choosing the job package that places them on the highest possible indifference curve. Worker A, who dislikes the risk of injury the most, maximizes utility at point P_A, and ends up at firm X, which happens to be the firm that finds it easiest to provide a safe work environment. But worker C, who minds the risk of injury the least, maximizes utility at point P_C and ends up at firm Z, the firm that finds it difficult to provide a safe work environment.

There is a nonrandom sorting of workers and firms. Safe firms are matched with workers who value safety a lot, and risky firms are matched with workers who value safety the least. In effect, workers self-select themselves across the spectrum of firms. The matching

FIGURE 5-6 The Hedonic Wage Function
Different firms have different isoprofit curves and different workers have different indifference curves. The labor market marries workers who prefer low probabilities of injury (such as worker A) with firms that find it easy to provide a safe environment (like firm X); and workers who do not mind risk as much (worker C) with firms that find it difficult to provide a safe environment (firm Z). The hedonic wage function gives the observed relationship between wages and job characteristics.



of workers to firms differs radically from the equilibrium implied by the standard supply–demand framework. In the usual equilibrium, firms and workers are indistinguishable, implying a random sorting of workers and firms. In contrast, the compensating differential model "marries" workers and firms that have common interests.

The points P_A, P_B, and P_C in Figure 5-6 give the (w, ρ) combinations that will actually be observed in the labor market. If we connect these points, we generate what is called the **hedonic wage function,** which summarizes the relationship between the wage that workers get paid and job characteristics. Because workers dislike working in risky firms, and because it is expensive to provide safety, the hedonic wage function is upward sloping.

The slope of the hedonic wage function gives the wage premium offered by a slightly riskier job. At point P_A, the hedonic wage function is tangent to worker A's indifference curve, so that the slope of the hedonic wage function gives worker A's reservation price. At point P_C, the hedonic wage function is tangent to worker C's indifference curve, and the slope of the hedonic wage function gives worker C's reservation price. In short, the slope of the hedonic wage function measures the reservation price of workers. As we shall see, this theoretical property of the hedonic wage function has important policy implications.

TABLE 5-1
Injury Rates in the United States, By Industry, 2008

Industry Group	Deaths (per 100,000 Workers)	Number of Disabling Injuries (in 1,000s)
Total	2.9	3,200
Agriculture	29.0	60
Mining	21.1	10
Construction	8.9	260
Manufacturing	2.3	390
Wholesale trade	3.8	80
Retail trade	0.9	380
Transportation and warehousing	13.0	160
Utilities	4.0	20
Information	1.0	30
Financial activities	0.6	70
Professional & business services	2.2	150
Educational & health services	0.5	510
Leisure & hospitality	0.9	270
Other services	1.8	110
Government	1.8	700

Notes: A disabling injury is one which results in death, some degree of physical impairment, or renders the person unable to perform regular activities for a full day beyond the day of the injury.

Source: U.S. Department of Commerce, *Statistical Abstract of the United States, 2011,* Washington, DC: Government Printing Office, 2011, Table 656.

5-3 Policy Application: How Much Is a Life Worth?

As Table 5-1 shows, there is a lot of variation in injury rates across workers employed in different industries. The annual rate of fatal injuries per 100,000 workers was 29.0 in agriculture, 13.0 in transportation, and 0.6 in financial services. Many studies estimate the hedonic function that relate wages and the probability of injury on the job.[5] The studies typically estimate regressions of the form

$$w_i = \alpha\, \rho_i + \text{Other variables} \qquad \textbf{(5-7)}$$

where w_i gives the wage of worker i and ρ_i gives the probability of injury on the worker's job.

The coefficient α gives the wage change associated with a one-unit increase in the probability of injury, holding constant other factors that might generate wage differences across workers, including the worker's educational attainment and age, and the location of the job. The empirical studies typically find a positive correlation between wages and hazardous or unsafe work conditions, regardless of how the hazard or the unsafe nature of the work environment is measured.[6]

[5] The most influential study is that of Richard Thaler and Sherwin Rosen, "The Value of Saving a Life: Evidence from the Labor Market," in Nestor Terleckyj, editor, *Household Production and Consumption,* New York: Columbia University Press, 1976, pp. 265–298.

[6] See Jeff Biddle and Gary Zarkin, "Worker Preferences and Market Compensation for Job Risks," *Review of Economics and Statistics* 70 (November 1988): 660–667; John Garen, "Compensating Wage Differentials and the Endogeneity of Job Riskiness," *Review of Economics and Statistics* 70 (February 1988): 9–16; Thomas J. Kniesner, W. Kip Viscusi, Christopher Woock, and James P. Ziliak, "The Value of a Statistical Life: Evidence from Panel Data," *Review of Economics and Statistics* 94 (February 2012): 74–87; and Morley Gunderson and Douglas Hyatt, "Workplace Risks and Wages: Canadian Evidence from Alternative Models," *Canadian Journal of Economics* 34 (May 2001): 377–395.

Perhaps the most interesting evidence pertains to the correlation between the wage and the probability of fatal injuries on the job. Workers who are exposed to a high probability of fatal injuries earn more. A survey of the evidence concludes that a 0.001 increase in the probability of fatal injury (so that, on average, an additional worker out of every thousand will die of job-related injuries in any given year) may increase annual earnings by about $8,700 (in 2017 dollars).[7]

Calculating the Value of Life

These correlations allow us to calculate the "value of life." Let's compare two jobs. Workers employed in firm X have a probability of fatal injury equal to ρ_x and earn w_x dollars per year. Workers employed in firm Y have a probability of fatal injury exceeding that in firm X by 0.001. The empirical evidence indicates that the riskier job, on average, pays about $8,700 more. We summarize the data as follows:

Firm	Probability of Fatal Injury	Annual Earnings
X	ρ_x	w_x
Y	$\rho_x + 0.001$	$w_x + \$8{,}700$

Suppose that firms X and Y each employs 1,000 workers. Because firm Y's probability of fatal injury is 0.001 higher, firm Y is likely to have one additional fatality in any given year. Workers in firm Y willingly accept the additional risk because *each* gets a compensating differential of $8,700.

Recall the theoretical implication that the hedonic wage function is tangent to the worker's indifference curves. As a result, the wage increase resulting from a 0.001 increase in the probability of fatal injury is *exactly* what it takes to convince the marginal worker in firm Y to accept the slightly riskier job and hold her utility constant. In other words, it is the marginal worker's reservation price.

This interpretation of the slope of the hedonic wage function suggests that each of the workers in firm Y is willing to give up $8,700 per year to reduce the probability of fatal injury in their job by 0.001. Put differently, the 1,000 workers employed in firm Y are willing to give up $8.7 million (or $8,700 × 1,000 workers) to save the life of that one worker who will likely die that year. The workers in firm Y, therefore, value a life at $8.7 million.

This is obviously not the answer we would get if the workers knew beforehand which one of the 1,000 was scheduled to suffer a fatal injury that year and we were to ask that unlucky person how much she would be willing to pay to avoid her fate. Our calculation instead gives the amount that workers are jointly willing to pay to reduce the likelihood that one of them will suffer a fatal injury in any given year. Put differently, it is the **value of a statistical life.**

Not surprisingly, there is a lot of variation in the estimated correlation between wages and the probability of fatal injury on the job. As a result, there is uncertainty about what the "true" value of a statistical life is.

[7] W. Kip Viscusi, "The Value of Risks to Life and Health," *Journal of Economic Literature* 31 (December 1993): 1912–46; see also Orley S. Ashenfelter, "Measuring the Value of a Statistical Life: Problems and Prospects," *Economic Journal* 116 (March 2006): C10–C23; and Per-Olov Johansson, "Is There a Meaningful Definition of the Value of a Statistical Life?" *Journal of Health Economics* 20 (January 2001): 131–139.

Theory at Work

THE VALUE OF LIFE ON THE INTERSTATE

In 1987, the federal government gave states the option of raising the speed limit on their rural interstate highways, from 55 to 65 mph. By the end of 1987, 38 states had raised the maximum speed limit on their rural interstates despite the warning that such an increase would lead to more highway fatalities.

Proponents of the legislation argued that increasing the speed limit would benefit travelers by reducing travel time. A report by the Indiana Department of Transportation makes the trade-off clear: "Speed limits represent trade-offs between risk and travel time ... reflecting an appropriate balance between the societal goals of safety and mobility." The states that chose to increase the speed limit were implicitly indicating that the value of time saved by driving faster was worth more than the value of the lives of the additional fatalities.

The data indicate that the states that adopted the higher speed limit experienced an increase in fatality rates on the affected highways. The fatality rate (that is, the number of fatalities per 100 million vehicle-miles of travel) rose by around 35 percent. At the same time, the amount of time required to travel a mile fell by about 4 percent. Each fatality in effect, "saved" 125,000 hours of travel time. If we calculate the dollar savings implied by the shorter travel times at the mean wage, one additional fatality saved $2.3 million (in 2017) in travel costs. By willingly adopting the higher speed limit, the states were indicating that those savings were equal to or larger than the value of life on the interstate.

Source: Orley Ashenfelter and Michael Greenstone, "Using Mandated Speed Limits to Measure the Value of a Statistical Life," *Journal of Political Economy* 112 (February 2004): S226–S267.

Part of the problem arises because the wage impact of a 0.001 increase in the probability of fatal injury depends on which set of workers we are analyzing. Are the data referring to workers who switch from a job with a 0.001 probability to a job with a 0.002 probability, or to workers who switch from a job with a 0.050 probability to a job with a 0.051 probability? The workers who end up in the "low-risk" jobs (the jobs with a 0.001 probability) obviously differ from the workers who end up in the "high-risk" jobs (the jobs with the 0.050 probability). The wage impact of a 0.001 increase in the probability of fatal injury will typically depend on what type of a 0.001 increase we have in mind.

The concept and estimates of the value of a statistical life have influenced how the government regulates safety hazards. When making construction decisions, highway departments routinely compare the cost of a safer highway design with the dollar savings associated with fewer fatalities. In 2004, both the California Department of Transportation (Caltrans) and the U.S. Department of Transportation used a value of a statistical life of around $3 million to guide their decisions.[8] The Environmental Protection Agency (EPA) wanted to limit the exposure of workers in glass manufacturing to arsenic poisoning. The cost of this regulation per statistical life saved would have been $142 million. It was not cost-effective, and the proposed regulation was rejected.[9]

[8] Ashenfelter, "Measuring the Value of a Statistical Life: Problems and Prospects."

[9] W. Kip Viscusi, *Fatal Trade-offs: Public and Private Responsibilities for Risk,* New York: Oxford University Press, 1992.

5-4 Policy Application: Safety and Health Regulations

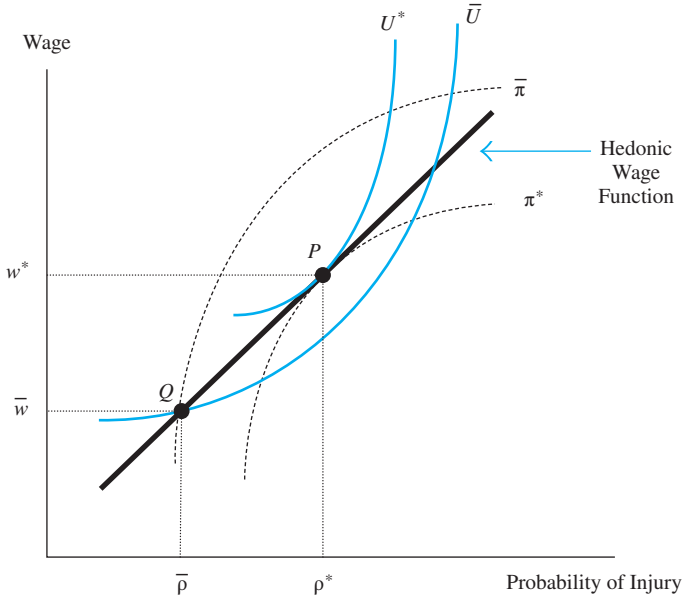
Since the enactment of the Occupational Safety and Health Act of 1970, the federal government in the United States has played a major role in setting safety standards at the workplace. The legislation created the Occupational Safety and Health Administration (OSHA), whose job is to protect the health and safety of the workforce. OSHA sets standards setting the maximum amount of cotton dust in the air in textile plants, requires that employers set up the workplace in a way that prevents workers from falling off platforms, and imposes many other restrictions on the job environment.

These regulations raise a number of important questions. Are workers better off as a result of the regulations? How do the safety standards alter the nature of the labor market equilibrium that generates compensating wage differentials? And do these government mandates actually reduce the probability of injury on the job?

The OSHA regulations effectively set a ceiling of $\overline{\rho}$ on the permissible injury rate. Figure 5-7 illustrates the impact of this ceiling on the labor market. Prior to the regulation, the worker chose the employment package at point *P,* which offered wage w^* and exposed her to probability of injury ρ^*. The worker got U^* utils and the employer earned π^* dollars of profits.

The regulation declares this employment package to be illegal and forces the worker to accept a job at point *Q* on the hedonic wage function. The new job offers the maximum

FIGURE 5-7 Impact of OSHA Regulation on Wages, Utility, and Profits
A worker maximizes utility by choosing the job at point *P,* which offers a package of w^* and ρ^*. The government sets a ceiling of $\overline{\rho}$ on the probability of injury, shifting the worker and the firm to point *Q.* The worker gets a lower wage (from w^* to $\overline{w}$), has less utility (from U^* to $\overline{U}$), and the firm earns lower profits (from π^* to $\overline{\pi}$).



injury rate of $\overline{\rho}$, but pays a lower wage of $\overline{w}$. The new employment package *must* lower the worker's utility to $\overline{U}$. After all, the worker was employed in the job that maximized her utility prior to the regulation.

The regulation also affects the firm's profitability. The firm can no longer offer the employment package of w^* and ρ^*. To comply with the regulation, the firm also moves to point Q on the hedonic wage function, placing the firm on the higher isoprofit curve $\overline{\pi}$, and reducing the firm's profits. A sufficiently large reduction in profits could prompt the firm to shut down.

Impact of Regulations When Workers Are Unaware of the Risks

Figure 5-7 shows that mandated safety standards reduce both the utility of affected workers and the profitability of affected firms. So why bother regulating safety standards at all?

One justification is that workers do not know the true risks associated with particular jobs. Construction workers in the 1950s and 1960s, for instance, did not know that continued exposure to asbestos fibers would eventually create serious health problems. It is worth pointing out, however, that neither firms nor government bureaucrats had complete information either, suggesting that perhaps the problem could not have been handled properly at the time.

But suppose that employers know full well the risks associated with the job, and that workers systematically underestimate the risk they are being exposed to. For instance, workers might be very optimistic about their own chances of escaping injury when they are employed as test pilots, even though a dispassionate and unblinking look at the data would suggest otherwise.

Consider the hedonic wage function in Figure 5-8. The worker gets a wage of w^* dollars but believes that she is being exposed to ρ_0, rather than the true injury probability of ρ^*. Because of this misperception, the worker thinks she is getting U_0 utils. In fact, she is only getting U^*.

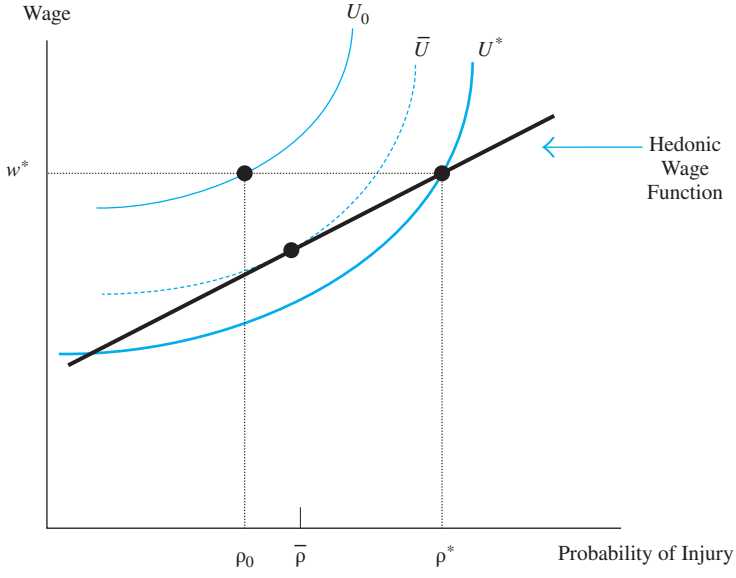
When workers misperceive their chances of getting injured, the government can step in and raise the worker's utility. The government can impose a ceiling on the injury rate anywhere between ρ_0 and ρ^*, and this ceiling would increase the worker's *actual* utility. If the government sets the ceiling at $\overline{\rho}$, for example, the worker's utility would be $\overline{U}$, which although lower than the worker's perceived utility of U_0 actually makes the worker better off. Safety regulations, therefore, can improve the workers' well-being if workers consistently underestimate the true risk.[10]

It might seem redundant to ask if requiring employers to provide a safer work environment actually leads to a safer work environment. But it has been difficult to establish that OSHA regulations significantly improve safety in the workplace. The available evidence suggests that OSHA has only slightly reduced the injury rates in firms and that the impact of the mandates has declined over time.[11]

[10] W. Kip Viscusi and W. A. Magat, "An Investigation of the Rationality of Consumer Valuations of Multiple Health Risks," *Rand Journal of Economics* 18 (Winter 1987): 465–479; and W. Kip Viscusi, "Sources of Inconsistency in Societal Responses to Health Risks," *American Economic Review* 80 (May 1990): 257–261.

[11] John W. Ruser and Robert S. Smith, "Reestimating OSHA's Effects," *Journal of Human Resources* 26 (Spring 1991): 212–236; and Wayne B. Gray and John Mendeloff, "The Declining Effects of OSHA Inspections on Manufacturing Injuries: 1979 to 1998," *Industrial and Labor Relations Review* 58 (July 2005): 571–587.

FIGURE 5-8 Workers Misperceive Risks on the Job
Workers earn a wage of w^* and incorrectly believe that their probability of injury is ρ_0. The probability is actually ρ^*. The government can mandate that firms do not offer a probability of injury higher than $\bar{\rho}$, increasing the worker's actual utility from U^* to $\bar{U}$.



5-5 Compensating Differentials and Job Amenities

Although we derived the hedonic wage function in terms of a single job characteristic, the probability of injury on the job, the model is equally applicable to other job characteristics, such as whether the job involves repetitive and monotonous work, whether the job is located in a pleasant physical setting, and whether the job involves strenuous physical work. As long as *all* persons in the labor market agree on whether a particular job characteristic is a "good" or a "bad," good job characteristics would be associated with low wage rates and bad job characteristics would be associated with high wage rates.

The empirical studies typically estimate the hedonic wage function in equation (5-7) by correlating the worker's wage with various job characteristics. Despite the central role played by the theory of compensating differentials in our understanding of labor market equilibrium, the evidence does not provide a ringing endorsement of the theory. A survey of the evidence concluded that "tests of the theory of compensating wage differentials are inconclusive with respect to every job characteristic except the risk of death."[12] For instance, jobs that demand physical strength are presumably more unpleasant than other

[12] Charles Brown, "Equalizing Differences in the Labor Market," *Quarterly Journal of Economics* 94 (February 1980): 113–134.

jobs, and hence would be expected to pay higher wage rates. But, in fact, jobs requiring workers to have substantial physical strength often pay less.[13]

Why Do Compensating Differentials Often Go the "Wrong" Way?

Our theoretical discussion suggests why many empirical tests of the theory may contradict our expectations. Simply put, the "correct" direction of the wage differential typically reflects our own preferences and biases. We are obviously reasonable people, so jobs *we* find disagreeable should pay more. But the compensating wage differential observed in the labor market measures what it took to get the *marginal* worker to accept that particular offer. If that marginal worker likes risky jobs or working outdoors doing strenuous physical activity, the market wage differential will be in what seems to be the wrong direction.

In addition, estimates of the compensating wage differential associated with particular job characteristics are valid only if all the other factors that influence a worker's wages are held constant. Because more able workers have a larger earnings potential, those workers might want to spend some of their additional income on job amenities. More able workers will then have higher wages *and* a surplus of "good" job amenities. This correlation would work against the compensating wage differential hypothesis. Because a worker's ability is seldom observed, the failure of the estimated correlations to show the right sign may be instead suggesting that we have not estimated the correct regression model.

One way of ridding the analysis from this type of ability bias is to track a particular worker over time as she changes jobs and purchases different packages of job amenities. Because a worker's innate ability does not change from job to job, the correlation between the change in the wage and the change in the job amenity might better measure the compensating wage differential. It turns out that the correlation between the change in a worker's wage and the change in the package of job amenities is more consistent with the compensating differentials hypothesis.[14]

Compensating Differentials and Layoffs

A key justification for the unemployment insurance (UI) system is that workers need to be protected from the vagaries of the competitive labor market. The UI system pays for a fraction of the worker's salary when a worker becomes unemployed, stabilizing the flow of income (and consumption) for workers who are laid off from their jobs.[15]

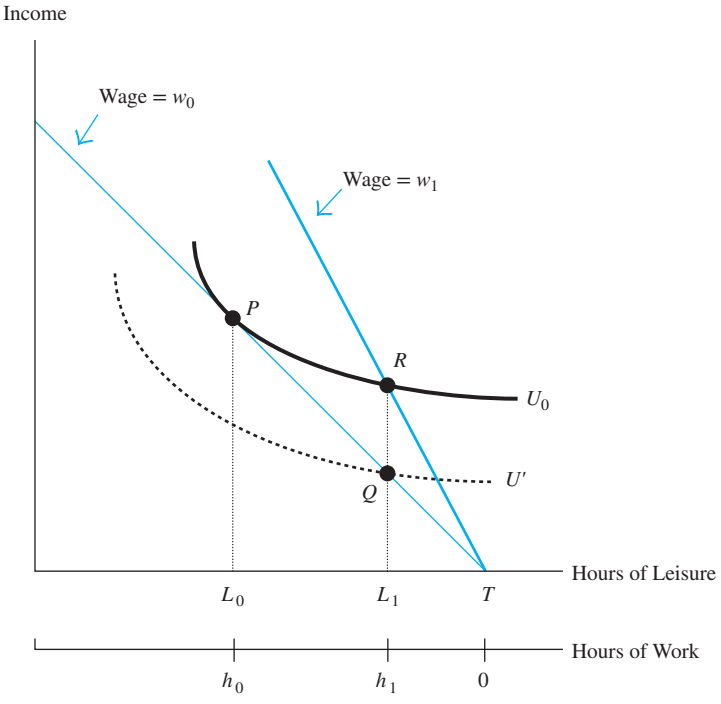
The income-stabilization justification for the UI program, however, seems less appealing if the labor market, through compensating wage differentials, *already* compensates workers with high layoff probabilities. As Adam Smith first argued over two centuries ago, the "constancy or inconstancy of employment" will generate compensating wage differentials.

[13] Robert E. B. Lucas, "The Distribution of Job Characteristics," *Review of Economics and Statistics* 56 (November 1974): 530–540; and Robert E. B. Lucas, "Hedonic Wage Equations and the Psychic Return to Schooling," *American Economic Review* 67 (September 1977): 549–558.

[14] Greg Duncan and Bertil Holmlund, "Was Adam Smith Right after All? Another Test of the Theory of Compensating Differentials," *Journal of Labor Economics* 1 (October 1983): 366–379; and Ernesto Villanueva, "Estimating Compensating Wage Differentials Using Voluntary Job Changes: Evidence from Germany," *Industrial and Labor Relations Review* 60 (July 2007): 544–561.

[15] Jonathan Gruber, "The Consumption Smoothing Benefits of Unemployment Insurance," *American Economic Review* 87 (March 1997): 182–205.

FIGURE 5-9 Layoffs and Compensating Differentials
At point *P,* a worker maximizes utility by working h_0 hours at a wage of w_0 dollars. An alternative job offers the worker a seasonal schedule, where the wage is the same but she only works h_1 hours. The worker is worse off in the seasonal job (her utility declines from U_0 to U' utils). If the seasonal job is to attract any workers, it must raise the wage to w_1 so that workers will be indifferent between the two jobs.



To illustrate, suppose a utility-maximizing worker has a job where she works h_0 hours per year at a wage rate of w_0 dollars. Figure 5-9 shows the situation using the neoclassical model of labor–leisure choice presented in the chapter on labor supply. Utility maximization occurs when the indifference curve is tangent to the budget line at point *P.* The worker gets U_0 utils.

The worker receives an outside job offer. In this alternative job, the worker will continue to receive a wage of w_0, but she will not need to work as many hours. Because of *perfectly predictable* layoffs (perhaps due to seasonal factors like the retooling of an auto factory prior to the beginning of a new model year), the worker has to work only h_1 hours per year. This alternative employment package places the worker at point Q on the original budget line (working h_1 hours at a wage of w_0), and moves the worker to the lower indifference curve U'.

The worker will not accept this offer because it yields less utility than the current job. In order to attract the worker, therefore, a job that offers only h_1 hours of work must offer a higher wage. The new steeper budget line crosses the original indifference curve at point *R.* The worker would be indifferent between an offer of h_0 hours of work at a wage of w_0 (point *P*) and an offer of h_1 hours at a wage of w_1 (point *R*). When layoffs are perfectly

predictable, therefore, a job with reduced work hours will have to compensate its workers by offering a higher wage.[16]

There is evidence that the labor market indeed provides compensating differentials to workers at the risk of layoff.[17] But the size of the compensating differential depends on whether workers are already covered by the UI system. In a sense, the UI system may have replaced one insurance system (determined by the labor market) with another that is taxpayer financed.

Compensating Differentials and Income Taxes

The theory of compensating differentials emphasizes the notion that the total value of a particular job offer equals the sum of the wage and the value of the job's amenities. Income taxes are usually only levied on the cash income the worker receives, so that the positive value of favorable amenities is nontaxable income.

There probably are many jobs, perhaps located in different occupations or locations that offer the same *total* compensation. But these jobs differ in how the compensation is actually paid. The wage might be a large fraction of total compensation in some jobs, but the amenities might matter much more in other jobs.

It is easy to see how a worker might react to an increase in the marginal income tax rate. The worker can receive the same total compensation in different jobs, but the income tax penalizes cash payments relative to amenities. An increase in the tax rate creates an incentive for the worker to switch to jobs where nontaxable benefits make up an ever-larger fraction of the total package.

There is evidence that workers indeed respond to higher income tax rates in this fashion.[18] By tracking a worker's occupational choice over the life cycle, we can see that workers respond to increases in the federal income tax rate by correspondingly changing their occupations—that is, by switching to occupations where a greater fraction of the compensation is paid in terms of "good" amenities.

The link between compensating differentials and income taxes has a second important implication. Note that "bad" amenities, such as a higher probability of injury, typically generate a compensating differential that *raises* the job's wage rate, exposing much of the compensation in those jobs to the income tax. An increase in the tax rate would then force

[16] Of course, the timing and duration of many layoffs are very hard to predict. But even if workers do not know when they will be laid off, the competitive market will still compensate workers who have a high probability of being laid off; see John Abowd and Orley Ashenfelter, "Anticipated Unemployment, Temporary Layoffs, and Compensating Wage Differentials," in Sherwin Rosen, editor, *Studies in Labor Markets,* Chicago: University of Chicago Press, 1981.

[17] James Adams, "Permanent Differences in Unemployment and Permanent Wage Differentials," *Quarterly Journal of Economics* 100 (February 1985): 29–56; Enrico Moretti, "Do Wages Compensate for Risk of Unemployment? Parametric and Semiparametric Evidence from Seasonal Jobs," *Journal of Risk and Uncertainty* 20 (January 2000): 45–66; and Susan Averett, Howard Bodenhorn, and Justas Staisiunas, "Unemployment Risk and Compensating Differentials in New Jersey Manufacturing," *Economic Inquiry* 43 (October 2005): 734–749.

[18] David Powell and Hui Shan, "Income Taxes, Compensating Differentials, and Occupational Choice: How Taxes Distort the Wage-Amenity Decision, *American Economic Journal: Economic Policy* 4 (February 2012): 224–247.

risky firms to redouble their efforts to attract and retain workers, leading to an even larger compensating differential between risky and safe firms.[19]

Compensating Differentials and HIV

The rapid growth of Acquired Immunodeficiency Syndrome (AIDS) created the most serious health crisis of the modern world. AIDS occurs when a person is infected with the human immunodeficiency virus (HIV), a virus that is transmitted by blood-to-blood or sexual contact. By 2016, nearly 37 million people were infected with HIV worldwide. Even though the first case of AIDS was only diagnosed in 1981, a number of studies have documented that the fear of HIV infection has created sizable compensating differentials in many labor markets.

Sonagachi is the red-light district of Calcutta. Workers in this district, located near Calcutta University, have been plying their trade for more than 150 years.[20] The price that the sex workers can charge, of course, depends on the characteristics associated with the transaction, including the physical attributes of the establishment (such as air conditioning and the amount of privacy) and the physical attributes of the sex worker (such as age and beauty).

In September 1992, the All India Institute of Public Health and Hygiene began to educate Sonagachi's sex workers about HIV and AIDS. Prior to this effort, the sex workers had practically no knowledge of the virus, of how it was transmitted, or of how safe sex practices could reduce the risk of transmission. By November 1993, roughly half of the sex workers had received this valuable information.

After learning about HIV, some of the female sex workers chose to practice safe sex and began to demand that customers use condoms. Many men, however, prefer not to use condoms, implying that the typical man was not willing to pay as much to a sex worker who demands the use of a condom as to one who will offer unprotected sex. Inevitably, compensating differentials arose in the Sonagachi marketplace. Sex workers engaged in unprotected sex charged more to compensate for the additional risk, and they attracted male clients who were willing to pay to avoid using a condom. The compensating differential was substantial: Sex workers who did not practice safe sex charged more than twice as much as those who did.

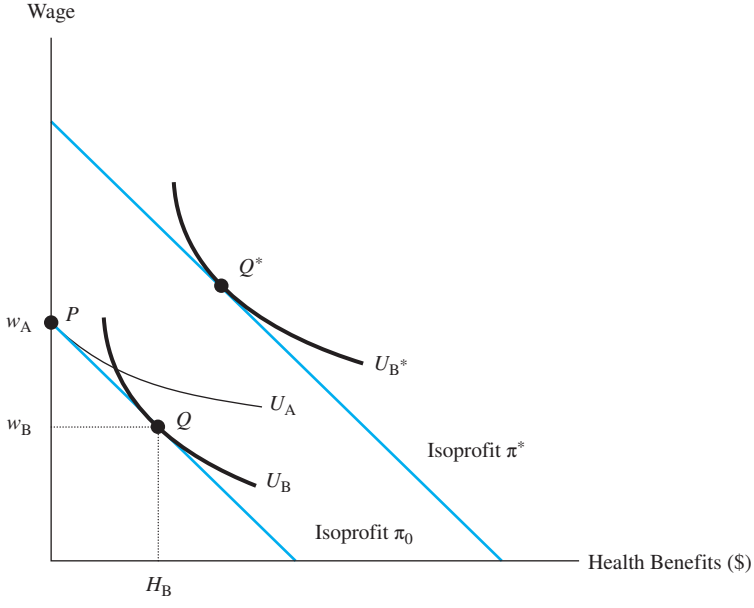
5-6 Policy Application: Health Insurance and the Labor Market

In the United States, employers typically provide health insurance coverage as a fringe benefit to a large fraction of the workforce. The value of this benefit suggests the existence of a sizable compensating differential between wages and the availability of employer-provided insurance.

[19] David Powell, "Compensating Differentials and Income Taxes: Are the Wages of Dangerous Jobs More Responsive to Tax Changes than the Wages of Safe Jobs?" *Journal of Human Resources* 47 (Fall 2012): 1023–1054.

[20] Vijayendra Rao, Indrani Gupta, Michael Lokshin, and Smarajit Jana, "Sex Workers and the Cost of Safe Sex: The Compensating Differential for Condom Use among Calcutta Prostitutes," *Journal of Development Economics* 71 (August 2003): 585–603; see also Paul Gertler, Manisha Shah, and Stefano M. Bartozzi, "Risky Business: The Market for Unprotected Commercial Sex," *Journal of Political Economy* 113 (June 2005): 518–550.

FIGURE 5-10 Health Benefits and Compensating Differentials
Workers A and B have the same earnings potential and face the same zero-profit isoprofit curve summarizing the compensation packages offered by firms. Worker A chooses a package with a high wage and no health insurance benefits. Worker B chooses a package with wage w_B and health benefits H_B. The observed data identify the trade-off between job benefits and wages. Workers B and B* have different earnings potential, so their job packages lie on different isoprofit curves. Their choices generate a positive correlation between wages and health benefits. The observed data do not identify the trade-off between wages and health benefits.



Suppose that all workers view employer-provided health insurance as a "good." The worker's indifference curve relating wages and health insurance would then have the usual downward-sloping convex shape, as shown in Figure 5-10. Worker A has the flat indifference curve U_A, implying that she does not attach much value to being covered by health insurance. She is willing to give up health insurance benefits for a relatively small increase in her wage. Worker B's indifference curve U_B is steeper, implying that she attaches a high value to the employer-provided insurance.

The firm's isoprofit curve is also downward sloping. For a given level of profits, the firm can provide a package consisting of high wages and little health insurance coverage, or of low wages and a generous insurance program. The isoprofit curve drawn in the figure, π_0, represents the zero-profit isoprofit curve for the labor market that includes workers A and B. For simplicity, the isoprofit curve is drawn as a line.

Some workers, like A, choose the corner solution at point *P,* indicating that they would rather work at a job that did not provide health insurance at all, but instead get paid a very high wage. In contrast, worker B chooses point *Q,* and she splits her total compensation between a wage of w_B and health insurance valued at H_B dollars. The data that we would observe in this labor market consists of the compensation packages of the two workers. These data trace out the isoprofit curve, giving the trade-off implied by the compensating

differential model: The earnings that worker B must forego in order to get a slightly better health insurance benefit.

But many of the attempts at documenting this trade-off do not find a negative correlation between wages and employer-provided health insurance. Instead, they find a *positive* correlation.[21] To explain this apparent contradiction of the theory, it has been argued that the workers who have health insurance differ in important ways from the workers who do not.

Suppose, for example, that some workers are very able and have a high earnings potential; other workers are not as able and have less earnings potential. The isoprofit curve π_0 applies to the labor market for low-ability workers. A different (and higher) isoprofit curve would exist for workers who are more able; for a given level of health benefits, the firm can pay more productive workers a higher wage and still have zero profits.

The isoprofit curve π^* in Figure 5-10 is the zero-profit isoprofit curve that summarizes the compensation packages available to high-ability worker B*. This worker chooses the package at point Q^*. Because of her high earnings potential, worker B* can choose a package that offers both higher wages and more generous health benefits. If we were to correlate the observed data on wages and health insurance benefits for workers B and B*, the correlation would be positive because high-wage workers also have more generous health benefits, but this positive correlation provides no information whatsoever about compensating differentials.

One potential solution is to find an instrument that "nudges" a particular worker along the relevant isocost curve, making that worker exogenously choose a compensation package that offers less health insurance. We could then measure what happened to the worker's wage and calculate the compensating differential.

One such instrument is suggested by the way employer-provided insurance works in the United States.[22] Typically, the insurance covers not only the worker (say, the husband), but also his wife and children. As a result, only one of the spouses needs to be covered by employer-provided insurance to obtain coverage for the entire family. A wife whose husband already has employer-provided insurance has much more flexibility in the labor market. She can choose jobs that offer very little health insurance (or none at all) without putting household members in jeopardy.

Consider the link between wages and health insurance coverage in a sample of married women. A variable indicating if the husband has health insurance coverage is a valid instrument if it affects the wife's choice of health insurance coverage (it nudges the wife's choice of a compensation package along the isoprofit curve) *and* if it does not affect the wife's earnings potential (the wife's ability is not correlated with whether the husband has health insurance).

The evidence shows that women whose husbands have employer-sponsored insurance are indeed less likely to work in jobs that provide health insurance. The probability that a wife already covered by her husband's insurance obtains her own insurance is 15.5 percentage points lower than that of a wife whose husband does not have insurance.

[21] Janet Currie and Brigitte C. Madrian, "Health, Health Insurance, and the Labor Market," in Orley C. Ashenfelter and David Card, editors, *Handbook of Labor Economics,* vol. 3C, Amsterdam: Elsevier, 1999, pp. 3309–3415.

[22] Craig A. Olson, "Do Workers Accept Lower Wages in Exchange for Health Benefits?" *Journal of Labor Economics* 20 (April 2002, part 2): S91–S114.

At the same time, women married to men who have health insurance earn 2.6 percent more than women married to men who do not have health insurance.

These statistics suggest that a 15.5-percentage-point drop in the probability of having your own employer-provided insurance is associated with a wage increase of 2.6 percent. The method of instrumental variables then implies that the estimate of the trade-off is given by the ratio of $2.6 \div (-15.5) = -0.168$. Women who choose jobs that offer employer-sponsored insurance earn 16.8 percent less than they would have earned had they chosen a job that did not offer those benefits.

This statistic measures the compensating differential only if the variable indicating whether the husband has health insurance is a valid instrument. In other words, the husband's health insurance coverage affects the probability that the wife has her own employer-provided insurance *and* does not affect the wife's earnings potential. However, one can plausibly argue that high-wage men (who are more likely to have generous health insurance coverage) are more likely to marry high-wage women. This example illustrates the importance of determining whether the instruments that are widely used in empirical work in labor economics indeed satisfy the assumptions that underlie the method of instrumental variables.

Summary

- The worker's reservation price gives the wage increase that will persuade the worker to accept a job with an unpleasant characteristic, such as a higher probability of injury.
- The worker will switch to a riskier job if the market-compensating wage differential exceeds the worker's reservation price.
- Firms choose whether to offer a risky environment or a safe environment to their workers. Firms that offer a risky environment must pay higher wages; firms that offer a safe environment must invest in safety. The firm offers whichever environment is more profitable.
- The compensating wage differential observed in the labor market is the dollar amount required to convince the marginal worker (that is, the last worker hired) to move to the riskier job.
- If a few workers enjoy working in jobs that have a high probability of injury and if those jobs demand relatively few workers, the market wage differential may go the "wrong" way. Risky jobs will pay lower wages than safe jobs.
- There is a "marriage" of workers and firms in the labor market. Workers who dislike particular job characteristics match with firms that do not offer those characteristics; workers who like the characteristics match with firms that provide them. This matching generates the hedonic wage function, relating wages to job characteristics.
- The value of a statistical life can be calculated from the slope of the hedonic wage function that relates the wage to the probability of fatal injury on the job.
- Workers with high earnings potential are likely to earn more and to have more generous job benefits. This positive correlation generates an ability bias that makes it difficult to find evidence that fringe benefits generate compensating wage differentials.

Key Concepts

compensating wage differential, *171*
hedonic wage function, *182*
isoprofit curve, *180*
reservation price, *174*
value of a statistical life, *184*

Review Questions

1. Suppose there are two types of jobs in the labor market: "safe" jobs and "risky" jobs. Describe how the worker decides whether to accept a safe job (where she cannot be injured) or a risky job (where she will certainly be injured).
2. Describe how the firm decides whether to offer a safe working environment or a risky environment.
3. How is the market-compensating wage differential between safe jobs and risky jobs determined? Which type of job will offer a higher wage?
4. Describe how workers and firms "marry" each other in the labor market when there are many types of jobs offering various levels of risk to their workers.
5. What is the hedonic wage function? What does the slope of the hedonic wage function measure?
6. How do we calculate the value of a statistical life?
7. What is the impact of health and safety regulations on the utility of workers and on the profits of firms?
8. Show that the competitive labor market compensates workers for the probability that they will be laid off.
9. Explain how the method of instrumental variables can be used to estimate the compensating differential associated with employer-provided health benefits.

Problems

5-1. Politicians who support the green movement often argue that it is profitable for firms to pursue a strategy that is "environmentally friendly" (for example, by building factories that do not pollute), because workers will be willing to work in environmentally friendly factories at a lower wage rate. Evaluate the validity of this claim.

5-2. Consider the demand for and supply of risky jobs.

(a) Derive the algebra that leads from equations (5-4) and (5-5) to equation (5-6).

(b) Describe why the supply curve in Figure 5-2 is upward sloping. How does your explanation incorporate θ? Why?

(c) Using a graph similar to Figure 5-2, demonstrate how the number of dirty jobs changes as technological advances allow the cost of making worksites cleaner to fall for all firms.

5-3. Suppose there are 100 workers in the economy in which all workers must choose to work a risky or a safe job. Worker 1's reservation price for accepting the risky job is \$1; worker 2's reservation price is \$2, and so on. Because of technological reasons, there are only 10 risky jobs.

(a) What is the equilibrium wage differential between safe and risky jobs? Which workers will be employed at the risky firm?

(b) Suppose now that an advertising campaign, paid for by the employers who offer risky jobs, stresses the excitement associated with "the thrill of injury," and this campaign changes the attitudes of the work force toward being employed in a risky job. Worker 1 now has a reservation price of −$10 (that is, she is willing to pay $10 for the right to work in the risky job); worker 2's reservation price is −$9, and so on. There are still only 10 risky jobs. What is the new equilibrium wage differential?

5-4. Suppose all workers have the same preferences represented by

$$U = \sqrt{w} - 2x,$$

where w is the wage and x is the proportion of the firm's air that is composed of toxic pollutants. There are only two types of jobs in the economy, a clean job ($x = 0$) and a dirty job ($x = 1$). Let w_0 be the wage paid by the clean job and w_1 be the wage paid for doing the dirty job. If the clean job pays $16 per hour, what is the wage in dirty jobs? What is the compensating wage differential?

5-5. Suppose a drop in the compensating wage differential between risky jobs and safe jobs has been observed. Two explanations have been put forward:

- Engineering advances have made it less costly to create a safe working environment.
- The phenomenal success of a new action serial "Die On The Job!" has imbued millions of viewers with a romantic perception of work-related risks.

Using supply and demand diagrams show how each of the two developments can explain the drop in the compensating wage differential. Can information on the number of workers employed in the risky occupation help determine which explanation is more plausible?

5-6. Consider a competitive economy that has four different jobs that vary by their wage and risk level. The table below describes each of the four jobs.

Job	Risk (r)	Wage (w)
A	1/5	$3
B	1/4	$12
C	1/3	$23
D	1/2	$25

All workers are equally productive, but workers vary in their preferences. Consider a worker who values his wage and the risk level according to the following utility function:

$$u(w, r) = w + \frac{1}{r^2}.$$

Where does the worker choose to work? Suppose the government regulated the workplace and required all jobs to have a risk factor of 1/5 (that is, all jobs become A jobs). What wage would the worker now need to earn in the A job to be equally happy following the regulation?

5-7. AB Consulting and DF Partners are two identical consulting firms in all aspects except that AB Consulting fires all new hires who don't bring in at least $5 million in revenue during their 4-year probationary term while DF Partners fires all new hires who don't bring in at least $2 million in revenue during their 4-year probationary term.

(a) Assuming no worker likes to take on the risk of being fired, what would you expect salaries to look like across the two firms? That is, how do you expect the compensating differential to appear?

(b) Suppose rather than seeing what you predicted in part (a), it terms out that salaries are the same in both firms. Provide a few explanations as to why this might be the case.

5-8. The EPA wants to investigate the value workers place on being able to work in "clean" mines over "dirty" mines. The EPA conducts a study and finds the average annual wage in clean mines to be $42,250 and the average annual wage in dirty mines to be $47,250.

(a) According to the EPA, how much does the average worker value working in a clean mine?

(b) Suppose the EPA could mandate that all dirty mines become clean mines and that all workers who were in a dirty mine must therefore accept a $5,000 pay decrease. Are these workers helped by the intervention, hurt by the intervention, or indifferent to the intervention?

5-9. There are two types of farming tractors on the market, the FT250 and the FT500. The only difference between the two is that the FT250 is more prone to accidents than the FT500. Over their lifetime, one in ten FT250s is expected to result in an accident, as compared to one in twenty-five FT500s. Further, one in one-thousand FT250s is expected to result in a fatal accident, as compared to only one in five-thousand FT500s. The FT250 sells for $125,000 while the FT500 sells for $137,000. At these prices, 2,000 of each model are purchased each year. What is the statistical value farmers place on avoiding a tractor accident? What is the statistical value of a life of a farmer?

5-10. Consider the labor market for public school teachers. Teachers have preferences over their salary, amenities, and school characteristics.

(a) One would reasonably expect that high-crime school districts pay higher wages than low-crime school districts. But the data consistently reveal that high-crime school districts pay lower wages than low-crime school districts. Why? (Hint: in many cities the primary source of funding for teacher salaries is local property taxes.)

(b) Does your discussion suggest anything about the relation between teacher salaries and school quality?

5-11. (a) On a graph with the probability of injury on the x-axis and the wage level on the y-axis plot two indifference curves, labeled U_A and U_B, so that the person associated with U_A is less willing to take on risk relative to the person associated with U_B. Include an arrow on the graph showing which direction is associated with higher levels of utility. Explain what it is about the indifference curves that reveals person *A* is less willing to take on risk relative to person *B*.

(b) Consider a third person who doesn't care about the risk associated with the job. That is, he doesn't seek to limit risk or to expose himself to risk. On a new

graph, draw several of this person's indifference curves. Include an arrow on the graph showing which direction is associated with higher levels of utility.

(c) Consider a wage-risk equilibrium that is characterized by an upward-sloping hedonic wage function. Now suppose there is a government campaign that successfully alters people's perception of risk. In particular, each worker adjusts her preferences so that she now needs to be more highly compensated to take on risk. Discuss, and show on a single graph, how the government's campaign affects indifference curves, isoprofit lines, the equilibrium hedonic wage function, and the distribution of workers to firms.

5-12. Suppose everyone is highly productive, college educated, hard-working, etc. People still differ in their preferences for jobs—while some would prefer to be doctors than lawyers, others prefer to be lawyers than doctors, and so on—and everyone prefers to be a professional to being a trash collector, but as usual preferences vary across individuals. In order for this economy to function at all, someone needs to choose to be the trash collector. Who will be the trash collector, and in general terms how much will the job of trash collector pay?

5-13. Consider two identical jobs, but some jobs are located in Ashton while others are located in Benton. Everyone prefers working in Ashton, but the degree of this preference varies across people. In particular, the preference (or reservation price) is distributed uniformly from \$0 to \$5. Thus, if the Benton wage is \$2 more than the Ashton wage, then 40 percent (or two-fifths) of the worker population will choose to work in Benton. Labor supply is perfectly inelastic, but firms compete for labor. There are a total of 25,000 workers to be distributed between the two cities. Demand for labor in both locations is described by the following inverse labor demand functions:

$$\text{Ashton: } w_A = 20 - 0.0024\,E_A.$$
$$\text{Benton: } w_B = 20 - 0.0004\,E_B.$$

Solve for the labor market equilibrium by finding the number of workers employed in both cities, the wage paid in both cities, and the equilibrium wage differential.

5-14. U.S. Trucking pays its drivers \$40,000 per year, while American Trucking pays its drivers \$38,000 per year. For both firms, truck drivers average 240,000 miles per year. Truck driving jobs are the same regardless of which firm one works for, except that U.S. Trucking gives each of its trucks safety inspection every 50,000 miles while American Trucking gives each of its trucks a safety inspection every 36,000 miles. This difference in safety inspection rates results in a different rate of fatal accidents between the two companies. In particular, one driver for U.S Trucking dies in an accident every 24 million miles while one driver for American Trucking dies in an accident every 30 million miles. What is the value of a trucker's life implied by the compensating differential between the two firms?

5-15. When trying to quantify the compensating differential associated with a desirable fringe benefit such as health insurance, it is important to try to collect data on an equally productive set of workers. Why? Is it also true that it is important to try to collect data on an equally productive set of workers when trying to quantify the compensating differential associated with a firm characteristic that is disliked by most workers (for example, exposure to risk of injury on the job)?

Selected Readings

John Abowd and Orley Ashenfelter, "Anticipated Unemployment, Temporary Layoffs, and Compensating Wage Differentials," in Sherwin Rosen, editor, *Studies in Labor Markets.* Chicago: University of Chicago Press, 1981.

Orley S. Ashenfelter, "Measuring the Value of a Statistical Life: Problems and Prospects," IZA Discussion Paper No. 1911, January 2006.

Orley Ashenfelter and Michael Greenstone, "Using Mandated Speed Limits to Measure the Value of a Statistical Life," *Journal of Political Economy* 112 (February 2004): S226–S267.

Charles Brown, "Equalizing Differences in the Labor Market," *Quarterly Journal of Economics* 94 (February 1980): 113–134.

Craig A. Olson, "Do Workers Accept Lower Wages in Exchange for Health Benefits?" *Journal of Labor Economics* 20 (April 2002, part 2): S91–S114.

Vijayendra Rao, Indrani Gupta, Michael Lokshin, and Smarajit Jana, "Sex Workers and the Cost of Safe Sex: The Compensating Differential for Condom Use among Calcutta Prostitutes," *Journal of Development Economics* 71 (August 2003): 585–603

Richard Thaler and Sherwin Rosen, "The Value of Saving a Life: Evidence from the Labor Market," in Nestor Terleckyj, editor, *Household Production and Consumption.* New York: Columbia University Press, 1976, pp. 265–298.

W. Kip Viscusi, "The Value of Risks to Life and Health," *Journal of Economic Literature* 31 (December 1993): 1912–1946.

Chapter 6

Education

If you think education's expensive, try ignorance!

—Derek Bok

The theory of compensating differentials suggests that wages vary among workers because jobs are different. Wages will also vary because workers are different. We each bring into the labor market a unique set of abilities and acquired skills, or **human capital.** Some workers are research biologists while others are musicians. This chapter begins the discussion of how we choose the particular set of skills that we offer to employers and how our choices affect the evolution of earnings over the working life.

We acquire most of our human capital by investing in school and in formal and informal on-the-job training programs. This chapter focuses on how we decide how much schooling to get, and how that decision affects our earnings. The next chapter examines the investments in the postschool period, and shows how our cumulative decisions help determine the earnings distribution and our place in it.

The skills we acquire in school are an increasingly important component of our stock of knowledge. In 1940, 75.5 percent of adults in the United States had not graduated from high school, and only 4.6 percent had a college degree. By 2017, fewer than 10 percent of adults did not have a high school diploma, and more than a third had at least a college degree.

Why do some workers obtain a lot of schooling and others drop out at an early age? Workers who invest in schooling are willing to give up earnings today in return for higher earnings in the future. For example, we earn a relatively low wage while we attend college. However, we expect to be rewarded by higher earnings later on as we collect the returns on that investment. The trade-off between lower earnings today and higher earnings tomorrow, as well as the financial and institutional constraints that limit access to educational institutions, determines the distribution of educational attainment in the population.

We also will discuss whether the money spent on education is a good investment. In particular, how does the rate of return to school compare with the rate of return on other investments? Putting aside our own personal interest in knowing whether we are getting a good deal out of our college education, the rate of return plays an important role in many policy discussions. It is often argued, for instance, that subsidizing investments to education is the surest way to improve the economic well-being of low-income workers.

Our analysis will assume that the worker chooses the level of human capital investments that maximizes the present value of lifetime earnings. This approach to the study of the determinants of the earnings distribution differs fundamentally from alternative approaches that view a worker's wage as determined by luck and other random factors. These random events might include whether we happen to meet an aging billionaire on the way to work or whether we are having breakfast at a Hollywood diner when an influential agent walks in. The human capital approach does not deny that luck, looks, and being in the right place at the right time influence a worker's earnings. But it does emphasize the idea that our educational and training decisions play a crucial role in the determination of earnings.

6-1 Education in the Labor Market: Some Stylized Facts

Table 6-1 summarizes the distribution of education in the U.S. population. Although there are only slight differences in educational attainment between men and women, there are sizable differences among racial and ethnic groups. By 2017, only about 5 percent of white and Asian–American workers did not have a high school diploma, as opposed to 8 percent of black workers and 25 percent of Hispanics. Similarly, more than half of Asian Americans had at least a college diploma, as compared to 40 percent of white workers, 20 percent of African Americans, and 16 percent of Hispanics.

These differences in educational attainment are important because, as Table 6-2 shows, education is strongly correlated with labor force participation rates, unemployment rates, and earnings. The labor force participation rate of persons who lack a high school diploma is only 60 percent, as compared to 85 percent for college graduates. Similarly, the unemployment rate of high school dropouts during the strong economic recovery of 2017 was 8.4 percent, but it was negligible (2.3 percent) for college graduates. Finally, high school dropouts earn $31,000 annually, but college graduates earn $81,000.

The data also indicate that education has a substantial beneficial impact on the labor market experiences of minorities. For example, the unemployment rate of black high

TABLE 6-1 Educational Attainment of U.S. Population, 2017 (Persons Aged 25 and Over)

Source: U.S. Bureau of Labor Statistics, *Annual Social and Economic Supplement of the Current Population Surveys,* March 2017.

	Highest Grade Completed (Percentage of Population in Education Category)					
	Less Than High School (%)	High School Graduates (%)	Some College (%)	Associate Degree (%)	Bachelor's Degree (%)	Advanced Group: Degree (%)
All Persons	8.1	29.1	16.5	10.9	22.5	13.0
Gender						
Male	8.7	31.8	16.5	9.7	21.4	12.0
Female	7.5	26.5	16.5	12.0	23.6	13.9
Race/Ethnicity						
White	4.8	27.5	16.4	11.8	25.3	14.3
Black	7.6	34.7	21.1	11.0	10.3	9.3
Hispanic	24.7	35.6	15.3	8.4	11.5	4.5
Asian	5.5	19.2	9.5	6.8	32.0	27.0

TABLE 6-2 Labor Market Characteristics, by Education Group, 2017 (Persons Aged 25–64)

Source: U.S. Bureau of Labor Statistics, *Annual Social and Economic Supplement of the Current Population Surveys,* March 2017.

		Less Than High School	High School Graduates	Some College	College Graduates
All workers	Labor force participation rate	59.6	71.8	77.7	85.1
	Unemployment rate	8.4	5.3	3.7	2.3
	Annual earnings (in $1,000)	31.1	39.4	46.0	81.3
Gender					
Men	Labor force participation rate	72.1	79.1	82.5	90.4
	Unemployment rate	8.3	5.2	4.0	2.3
	Annual earnings (in $1,000)	36.7	46.7	54.6	98.6
Women	Labor force participation rate	45.6	63.3	73.5	80.5
	Unemployment rate	8.5	5.4	3.5	2.3
	Annual earnings (in $1,000)	21.8	29.3	37.4	64.7
Race/Ethnicity					
White	Labor force participation rate	56.3	72.5	77.6	85.7
	Unemployment rate	8.1	4.5	3.2	2.0
	Annual earnings (in $1,000)	32.1	42.2	48.9	83.7
Black	Labor force participation rate	47.0	66.7	77.2	85.2
	Unemployment rate	19.4	8.8	5.9	3.2
	Annual earnings (in $1,000)	26.0	32.2	38.7	67.3
Hispanic	Labor force participation rate	66.5	73.8	80.0	84.7
	Unemployment rate	6.8	5.4	4.1	3.5
	Annual earnings (in $1,000)	31.3	37.8	40.4	64.0
Asian	Labor force participation rate	57.2	73.1	77.8	80.6
	Unemployment rate	2.9	3.2	2.6	3.1
	Annual earnings (in $1,000)	31.3	33.3	45.3	89.5

school dropouts is 19.4 percent, as compared to 8.8 percent for black high school graduates and 3.2 percent for black college graduates. Similarly, Hispanic high school dropouts earn only $31,000 as compared to $64,000 for Hispanic college graduates.[1]

Although there are sizable differences in labor market outcomes between men and women and among race and ethnic groups—and these differences will be examined in the chapter on labor market discrimination—we first investigate a different lesson from the data: Education seems to play a crucial role in improving labor market outcomes for all workers.

[1] A detailed analysis of differences in educational attainment among race/ethnic groups is given by Stephen V. Cameron and James J. Heckman, "The Dynamics of Educational Attainment for Black, Hispanic, and White Males," *Journal of Political Economy* 109 (June 2001): 455–499. It is well known that the effects of education extend far beyond the labor market. See, for example, Damon Clark and Heather Royer, "The Effect of Education on Adult Mortality and Health: Evidence from Britain," *American Economic Review* 103 (October 2013): 2087–2120; and Giorgio Brunello, Daniele Fabbri, and Margherita Fort, "The Causal Effect of Education on Body Mass: Evidence from Europe," *Journal of Labor Economics* 31 (January 2013): 195–223.

6-2 Present Value

Any study of an investment decision—whether it is an investment in physical or in human capital—must contrast expenditures and receipts incurred at different times. In other words, an investor must be able to calculate the returns to the investment by comparing the current cost to the future returns. For reasons that will become obvious momentarily, the value of a dollar received today is not the same as the value of a dollar received tomorrow. The notion of **present value** allows us to compare dollars spent and received in different time periods.

Suppose somebody gives you a choice between two offers: You can have either \$100 today or \$100 next year. Which offer would you take?

A little reflection should convince you that \$100 today is better than \$100 next year. After all, if you receive \$100 today, you can invest it, and you will then have $\$100 \times (1 + 0.05)$ dollars next year (or \$105), assuming that the rate of interest equals 5 percent. Note, moreover, that receiving \$95.24 today (or $\$100 \div 1.05$) would be worth \$100 next year. Hence, the present value (or the current-dollar value) of receiving \$100 next year is only \$95.24. In general, the present value of a payment of, say, y dollars next year is

$$PV = \frac{y}{1 + r} \tag{6-1}$$

where r is the rate of interest, which is also called the **rate of discount.** The quantity PV tells us how much needs to be invested today in order to have y dollars next year. Put differently, a future payment of y dollars is discounted to make it comparable to current dollars.

The discussion clearly suggests that receiving y dollars 2 years from now is not equivalent to receiving y dollars today or even to receiving y dollars next year. A payment of \$100 today would be worth $\$100 \times (1 + 0.05) \times (1 + 0.05)$ 2 years from now. Hence, the present value of receiving y dollars 2 years from now is

$$PV = \frac{y}{(1 + r)^2} \tag{6-2}$$

By arguing along similar lines, we can conclude that the present value of a payment of y dollars received t years from now equals

$$PV = \frac{y}{(1 + r)^t} \tag{6-3}$$

These formulas are extremely useful when we study decisions that involve expenditures made or dollars received at different time periods because they allow us to state the value of these expenditures and receipts in terms of today's dollars.

6-3 The Schooling Model

More education is associated with lower unemployment rates and higher earnings. So why don't all workers have doctorates or professional degrees? In other words, what motivates some students to stay in school for over 20 years while others drop out before they finish high school?

We begin our analysis of this important question by assuming that students acquire the education level that maximizes the present value of lifetime earnings. Education, therefore, is valued only because it increases earnings. A college education obviously affects utility

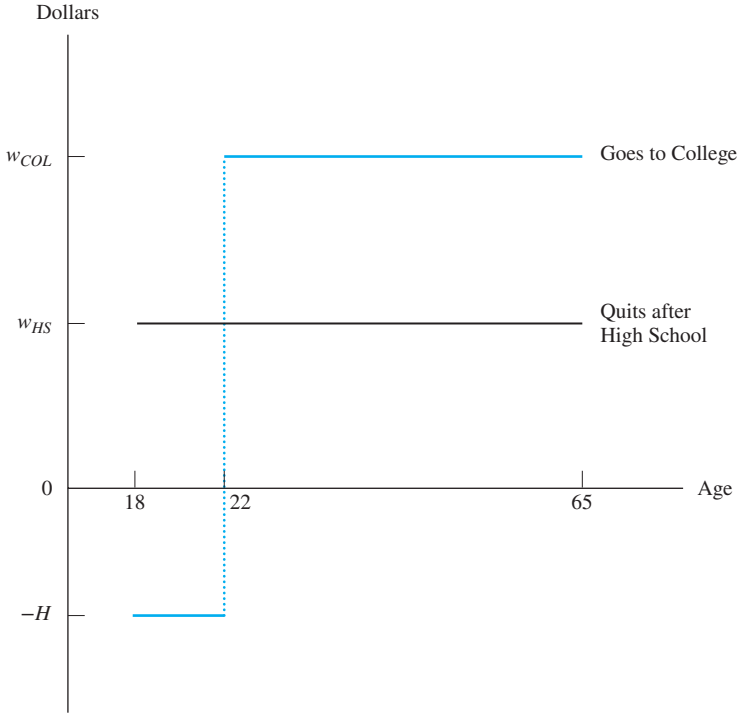
in many other ways. It teaches the student how to read and appreciate Nietzsche and how to work out complex mathematical equations; it even reduces the cost of entering the "marriage market" by facilitating contact with a large number of potential mates. Important though they may be, we ignore these side effects and concentrate exclusively on the monetary rewards of an education.

Consider the situation faced by an 18-year-old man who just graduated from high school and is contemplating whether to enter the labor market or attend college and delay labor market entry by four more years.[2] Suppose there is no on-the-job training, so that schooling is the only form of human capital. And suppose that the skills learned in school do not depreciate over time. These assumptions imply that the worker's productivity does not change once he leaves school, so that real earnings (that is, earnings after adjusting for inflation) are constant over the working life.

Figure 6-1 illustrates the trade-off. The figure shows the **age–earnings profile** (that is, the wage path over the life cycle) associated with each alternative. A high school

FIGURE 6-1 Potential Earnings Streams Faced by a High School Graduate

A student who quits school after getting his high school diploma can earn w_{HS} from age 18 until retirement. If he goes to college, he forgoes these earnings and incurs a cost of H dollars for 4 years and then earns w_{COL} until retirement.



[2] The schooling model originates in Jacob Mincer, "Investment in Human Capital and Personal Income Distribution," *Journal of Political Economy* 66 (August 1958): 281–302. An even earlier study that anticipated many of the central concepts is given by Milton Friedman and Simon Kuznets, *Income from Independent Professional Practice,* Princeton, NJ: Princeton University Press, 1954.

graduate earns w_{HS} dollars annually until retirement age, which occurs when the worker turns 65. If the student chooses to attend college, however, he does not work while in school and incurs an expense of H dollars to cover tuition, books, and fees. After graduation, he earns w_{COL} dollars annually until retirement.

Going to college is costly in two different ways. A year in college is a year spent not working (or at least a year spent at a low-wage part-time job), so that a college education forces the student to forgo some earnings. This is the **opportunity cost** of going to school—the cost of not pursuing the best alternative. The opportunity cost in Figure 6-1 is w_{HS} dollars for each year the student goes to college. The student also has out-of-pocket expenses of H dollars for tuition, books, and other fees.

Because college has no intrinsic value to the student, employers who wish to attract a highly educated (and presumably more productive) worker will have to offer higher wages, so that $w_{COL} > w_{HS}$. In a sense, the high wage paid to workers with more schooling is a compensating differential that compensates workers for their training costs. If college graduates earned less than high school graduates, no one would bother to get a college education.

Present Value of Age–Earnings Profiles

The present value of the earnings stream if the student gets only a high school education is

$$PV_{HS} = w_{HS} + \frac{w_{HS}}{(1+r)} + \frac{w_{HS}}{(1+r)^2} + \cdots + \frac{w_{HS}}{(1+r)^{46}} \tag{6-4}$$

where r gives the student's rate of discount. There are 47 terms in this sum, one term for each year that elapses between the ages of 18 and 64.

The present value of the earnings stream if the student gets a college diploma is

$$PV_{COL} = -H - \frac{H}{(1+r)} - \frac{H}{(1+r)^2} - \frac{H}{(1+r)^3} + \frac{w_{COL}}{(1+r)^4} \cdots + \frac{w_{COL}}{(1+r)^{46}} \tag{6-5}$$

The first four terms in this sum give the present value of the cost of a college education, and the remaining 43 terms give the present value of lifetime earnings in the postcollege period.

A student's schooling choice maximizes the present value of lifetime earnings. He attends college if

$$PV_{COL} > PV_{HS} \tag{6-6}$$

Let's illustrate with a numerical example. Suppose a person lives only two periods and chooses from two schooling options. He can choose not to attend school at all, in which case he earns $20,000 in each period. The present value of earnings is

$$PV_0 = 20{,}000 + \frac{20{,}000}{1+r} \tag{6-7}$$

He can also choose to go to school in the first period, incur out-of-pockets costs of $5,000, and enter the labor market in the second period, earning $47,500. The present value of this earnings stream is

$$PV_1 = -5{,}000 + \frac{47{,}500}{1+r} \tag{6-8}$$

Suppose that the rate of discount is 5 percent. It is easy to calculate that $PV_0 = \$39{,}048$ and that $PV_1 = \$40{,}238$. The student, therefore, chooses to attend school. Note, however, that if the rate of discount were 15 percent, $PV_0 = \$37{,}391$, $PV_1 = \$36{,}304$, and he would not go to school.

As the example shows, the rate of discount r plays a key role in determining whether a student goes to school or not. He goes to school if the rate of discount is 5 percent but does not if the rate of discount is 15 percent. A student who has a high discount rate attaches a low value to future earnings opportunities; he discounts the future "too much." Because the returns to an investment in education are collected in the far-off future, students with high discount rates acquire less schooling.

It is sometimes assumed that a person's rate of discount equals the market rate of interest, the rate at which funds deposited in financial institutions grow over time. After all, the discounting of future earnings in the present value calculations arises partly because a dollar received this year can be invested and is worth more than a dollar received next year.

But the rate of discount also depends on how we feel about giving up some of today's consumption in return for future rewards—or our rate of "time preference." Casual observation (and psychological experiments) suggests that we differ in our rate of time preference. Some of us are present-oriented and some of us are not. Students who are present-oriented have a high discount rate and are less likely to stay in school.[3]

The Wage–Schooling Locus

The rule that a student chooses the level of schooling that maximizes the present value of earnings obviously generalizes to situations when there are more than two options. The student would then calculate the present value associated with each schooling level (for example, 1 year of schooling, 2 years of schooling, and so on) and choose the quantity that maximizes the present value of the earnings stream.

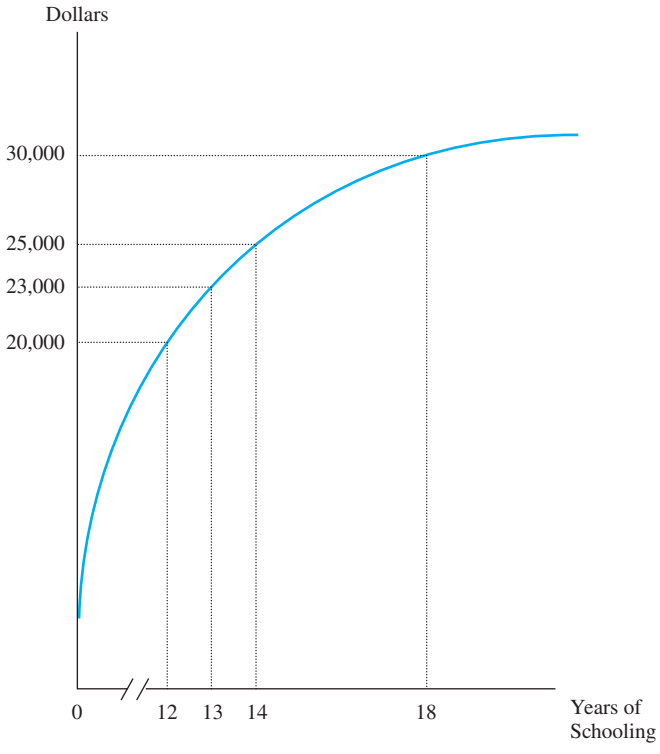
There is, however, a different way of formulating this problem that provides an intuitive "stopping rule."[4] This stopping rule tells us when it is optimal to quit school and enter the labor market. The alternative approach is useful because it also suggests a way for estimating the rate of return to education.

Consider initially a labor market where a worker's productivity depends only on how much education he has. Figure 6-2 illustrates the **wage–schooling locus,** which gives the salary that employers are willing to pay for every level of schooling. If the worker has a high school diploma, his annual salary is $20,000; if he has 18 years of schooling, his annual salary rises to $30,000. The wage–schooling locus is market determined. In other words, the salary for each level of schooling is determined by the intersection of the supply of workers with that particular schooling and the demand for those workers. From the worker's point of view, the salary associated with each level of schooling is a constant.

[3] Gary S. Becker and Casey B. Mulligan, "The Endogenous Determination of Time Preference," *Quarterly Journal of Economics* 112 (August 1997): 729–758; and Emily C. Lawrance, "Poverty and the Rate of Time Preference: Evidence from Panel Data," *Journal of Political Economy* 99 (February 1991): 54–77.

[4] Sherwin Rosen, "Human Capital: A Survey of Empirical Research," *Research in Labor Economics* 1 (1977): 3–39; and David Card, "The Causal Effect of Education on Earnings," in Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics,* vol. 3A, Amsterdam: Elsevier, 1999, pp. 1801–1863.

FIGURE 6-2 The Wage–Schooling Locus
The wage–schooling locus gives the salary that a particular worker would earn if he has a particular level of schooling. If the worker is a high school graduate, he earns $20,000 annually. If he has 1 year of college, he earns $23,000.



The wage–schooling locus in Figure 6-2 has three important properties:

1. *The wage–schooling locus is upward sloping.* Workers who have more education must earn more as long as educational decisions are motivated only by financial gains. To attract educated workers, employers must compensate those workers for the cost of acquiring that education.
2. *The slope of the wage–schooling locus tells us by how much a worker's earnings would increase if he were to obtain one more year of schooling.* The slope of the wage–schooling locus, therefore, is closely related to "the rate of return" to school.
3. *The wage–schooling locus is concave.* The law of diminishing returns also applies to human capital accumulation. Each extra year of schooling generates less incremental knowledge, so that the monetary gain from each additional year of schooling declines as more schooling is acquired.

The Marginal Rate of Return to School

The slope of the wage–schooling locus (or $\Delta w/\Delta s$) tells us by how much earnings increase if a student stays in school one more year. In Figure 6-2, for example, the first year of

college increases annual earnings by $3,000. The percentage change in earnings from getting this additional year of schooling is 15 percent (or $^{3,000}/_{20,000} \times 100$).

The percent change in earnings resulting from one more year of school will sometimes measure the percent increase in earnings per dollar spent in educational investments. As a result, we refer to the *percent* change in earnings from one more year of school as the marginal **rate of return to school.** In particular, suppose that the *only* costs incurred in going to college are forgone earnings. The high school graduate delaying his entry into the labor market by 1 year gives up $20,000. This investment increases his future earnings by $3,000 annually, yielding an annual 15 percent rate of return for the $20,000 investment associated with the first year of college.

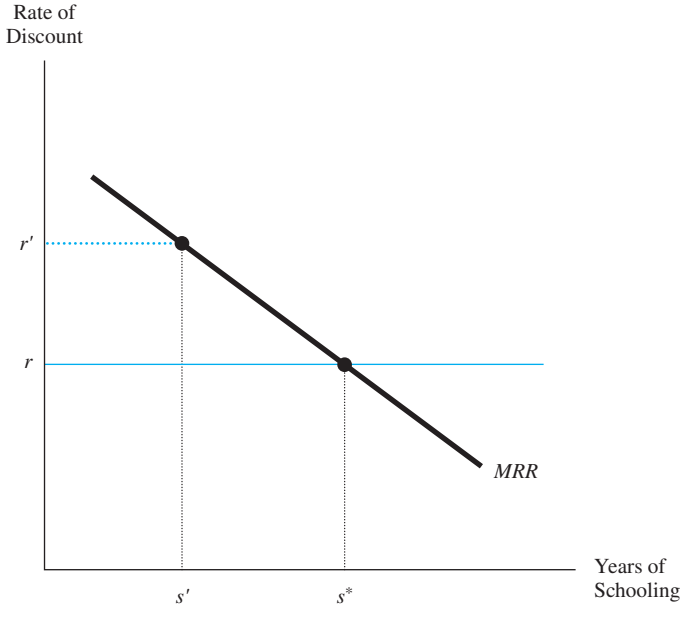
Because the wage–schooling locus is concave, the marginal rate of return *must* decline as a student gets more schooling. For example, the marginal rate of return of the second year of college is only 8.7 percent (a $2,000 return on a $23,000 investment). Each additional year of schooling generates a smaller salary increase because of the law of diminishing returns, and it costs more to stay in school the more educated the worker. The marginal rate of return curve, therefore, is a declining function of the level of schooling, as illustrated by the *MRR* curve in Figure 6-3.

The Stopping Rule, or When Should I Quit School?

Suppose that the student has a rate of discount r that is constant; it is independent of how much schooling he gets. The rate of discount curve, therefore, is perfectly elastic, as illustrated in Figure 6-3.

FIGURE 6-3 The Schooling Decision

The *MRR* schedule gives the marginal rate of return to school, or the percentage increase in earnings resulting from an additional year of school. A student maximizes the present value of lifetime earnings by equating the marginal rate of return with the rate of discount. A worker with discount rate r goes to school for s^* years.



Which level of schooling should a student choose? It turns out that the intersection of the *MRR* curve and the horizontal rate of discount curve determines the optimal level of schooling, or s^* years in the figure. In other words, the stopping rule that tells the student when to quit school is

Stop schooling when the marginal rate of return to school $= r$ **(6-9)**

This stopping rule maximizes the student's present value of earnings. To see why, suppose that his rate of discount equals the market rate of interest offered by financial institutions. Would it be optimal for the student to quit school after completing only s' years in Figure 6-3? If he were to stay in school for an additional year, he would forgo, say, w' dollars in earnings, and the rate of return to this investment equals r'. His alternative would be to quit school, work, and deposit the w' dollars in a bank that offers a rate of return of only r. Because education yields a higher rate of return, the student increases the present value of earnings by staying in school.

Conversely, suppose that the student gets more than s^* years of school. Figure 6-3 then shows that the marginal rate of return to this "excess" schooling is less than the market rate of interest, so that the extra years of schooling are not profitable.

Equation (6-9), the stopping rule for schooling investments, describes a general property of optimal investment decisions. The wealth-maximizing student who must decide when to quit school faces the same economic tradeoff as the owner of a forest who must decide when to cut down a tree. The longer the tree is in the ground, the bigger it gets and the more lumber and revenue it generates. But there are forgone profits (and maintenance costs) associated with keeping the tree in the ground. The tree should be cut down when the rate of return on investing in the tree equals the rate of return on alternative investments.

It is important to emphasize that the schooling decision is influenced by many factors (such as chance encounters with influential teachers or friends), and not just the dollar value of the earnings stream. There is also a lot of uncertainty about the rewards to schooling. Economic and social conditions change in unpredictable ways, and it is very difficult to forecast how these shocks shift the rewards to particular skills and careers. This uncertainty will surely play a role in our human capital decisions—just like the uncertainty in financial markets affects the type of financial portfolio that maximizes our wealth.[5]

6-4 Education and Earnings

The schooling model summarized in Figure 6-3 tells us how a particular student decides how much schooling to acquire and where the student ends up in the earnings distribution after he enters the workforce. Students who get more schooling earn more (although they also give up more). The model isolates two key factors that lead different students to obtain different levels of schooling and, hence, to have different earnings: They either have different rates of discount or face different marginal rate of return curves.

[5] Joseph G. Altonji, "The Demand for and Return to Education When Outcomes Are Uncertain," *Journal of Labor Economics* 11 (January 1993): 48–83.

Differences in the Rate of Discount

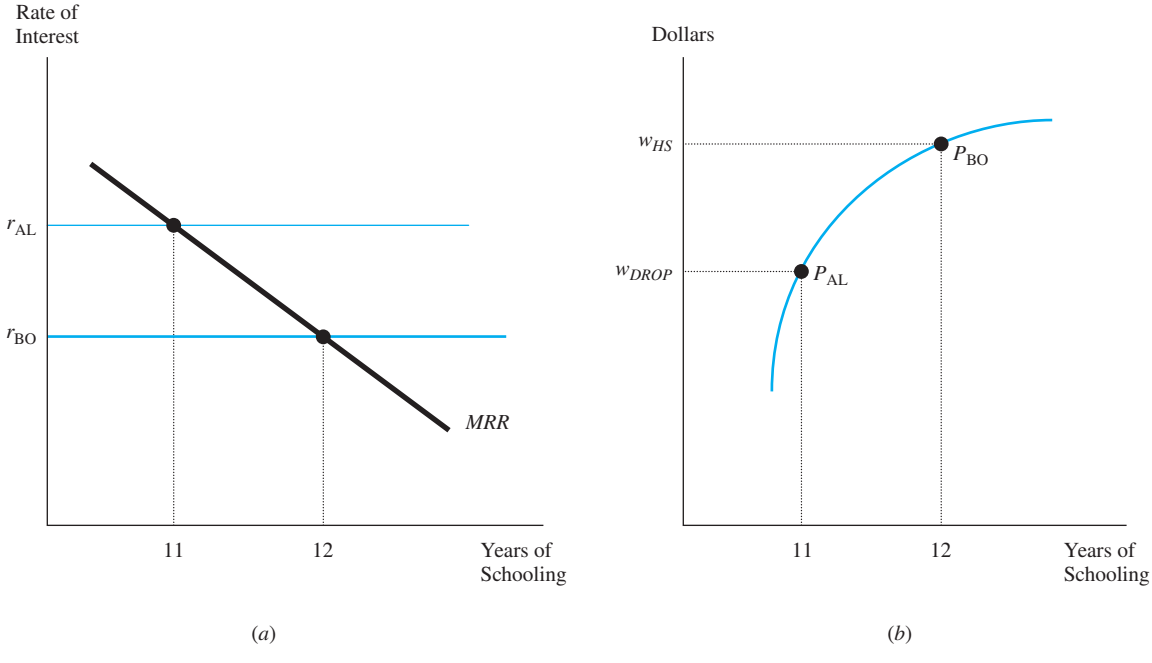
Consider a labor market with two workers who differ *only* in their discount rates, as illustrated in Figure 6-4*a*. Al's discount rate is r_{AL} and Bo's lower discount rate is r_{BO}. The figure shows that Al, who has a higher discount rate, dropped out of high school and got only 11 years of education; Bo got a high school diploma. Students who discount future earnings heavily do not stay in school long because they are present oriented.

Figure 6-4*b* shows the implications of these choices for the observed earnings distribution in the postschool period. We assumed that both workers face the same marginal rate of return curve. The derivation of the marginal rate of return curve implies that the two workers face the same wage–schooling locus. The different decisions of the two workers, therefore, place them at different points of the common locus. Al ends up at point P_{AL}, going to school for 11 years and earning w_{DROP}; Bo ends up at P_{BO}, going to school for 12 years and earning w_{HS}. Note that by connecting points P_{AL} and P_{BO}, we trace out the common wage–schooling locus faced by all workers. In addition, note that the wage gap between Al and Bo lets us estimate the rate of return to the 12th grade, the percent change in earnings that a worker would experience by going from the 11th to the 12th grade.

Estimates of the rate of return to school play a crucial role in many policy discussions. Consider, for example, the impact of a proposed law requiring all students to graduate from

FIGURE 6-4 Schooling and Earnings When Workers Have Different Rates of Discount

Al has a higher rate of discount (r_{AL}) than Bo (r_{BO}), so that Bo graduates from high school but Al drops out. Al chooses point P_{AL} on the wage–schooling locus and Bo chooses point P_{BO}. The observed data on wages and schooling trace out the common wage–schooling locus of the workers.



high school. By how much would this policy increase the earnings of workers who are now high school dropouts?

In effect, this policy "injects" Al with one more year of schooling. The wage–schooling locus in Figure 6-4 shows that Al's earnings would increase by $(w_{HS} - w_{DROP})$ if the law went into effect. A compulsory high school diploma nudges the worker along the *observed* wage–schooling locus.

As long as workers differ only in their discount rates, therefore, we can calculate the marginal rate of return to school from the wage differential between two workers who differ in their educational attainment. We can then correctly predict by how much earnings would increase if we pursued particular policies that injected targeted workers with more education.[6]

Differences in Ability

Up to this point, we have assumed that a worker's productivity is determined only by his educational attainment. However, workers with the same educational attainment may differ in their innate ability. And a higher level of ability will shift *up* the wage–schooling locus, as employers reward the more productive workers with a higher wage. It is much more difficult to estimate the rate of return to school when the wage–schooling locus differs across workers.

How does a shift in the wage–schooling locus affect the marginal rate of return curve? On the one hand, a more able worker earns more and must give up more to obtain an additional year of schooling. The higher cost of an additional year of schooling implies that the *MRR* curve would shift to the left. On the other hand, a more able worker may gain more from an additional year of schooling, so that more ability also makes the wage–schooling locus steeper. The steeper slope implies that the curve would shift to the right. It is often assumed that higher ability shifts the *MRR* curve to the right, so that the earnings gain from an additional year of schooling outweighs the higher cost.

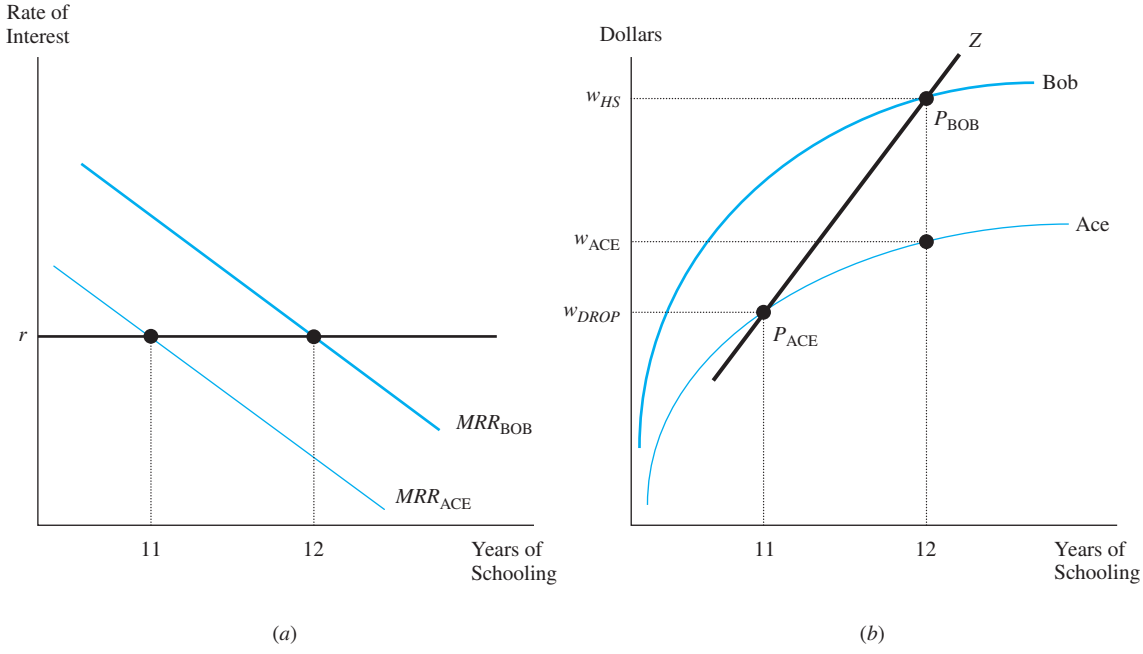
To isolate the impact of ability on the school decision, suppose that Al and Bob have the same rate of discount, but that Bob is more able. As illustrated in Figure 6-5*a,* Bob's *MRR* curve lies to the right of Ace's. Because Bob gets more from an additional year of schooling, Bob stays in school longer (12 years versus 11 years).

Figure 6-5*b* illustrates the impact of these decisions on postschool earnings. Note that Bob's wage–schooling locus lies above and is steeper than Ace's because Bob is more able. Bob chooses point P_{BOB} on *his* wage–schooling locus; Bob gets 12 years of schooling and earns w_{HS} dollars. Ace chooses point P_{ACE} on *his* wage–schooling locus; Ace goes to school for 11 years and earns w_{DROP} dollars.

[6] The model can be extended to examine how credit constraints, student aid, and other financial resources affect the education decision. The relaxation of financial constraints is often interpreted as a decrease in the rate of discount; the additional wealth may make students less present oriented or may allow students to borrow money (to finance their education) at a lower interest rate. See Thomas J. Kane, "College Entry by Blacks since 1970: The Role of College Costs, Family Background, and the Returns to Education," *Journal of Political Economy* 102 (October 1994): 878–911; and Stephen V. Cameron and Christopher Taber, "Estimation of Educational Borrowing Constraints Using Returns to Schooling," *Journal of Political Economy* 112 (February 2004): 132–182.

FIGURE 6-5 Schooling and Earnings When Workers Have Different Ability

Ace and Bob have the same discount rate *r,* but the workers face a different wage–schooling locus. Ace drops out of high school and Bob gets a high school diploma. The wage differential between Bob and Ace (or $w_{HS} - w_{DROP}$) arises both because Bob goes to school for one more year and because Bob is more able. This wage differential does not tell us by how much Ace's earnings would rise if he were to complete high school (or $w_{\text{ACE}} - w_{DROP}$).



The data at our disposal typically report a worker's education and earnings, but do not report his ability. Innate ability, after all, is seldom observed. The observed data, therefore, only allows us to connect points P_{ACE} and P_{BOB} in the figure and trace out the line labeled *Z.* This line does *not* coincide with either Ace's or Bob's wage–schooling locus. As a result, the observed data on earnings and schooling do not allow us to estimate the rate of return to school.

Suppose again that the government proposes a law requiring all persons to complete high school. To determine the economic impact of the proposed legislation, we need to know by how much Ace's earnings would increase if he were injected with one more year of schooling. The available data in line *Z* tell us that a high school graduate earns w_{HS} and that a high school dropout earns w_{DROP}. But the line connects points on different wage–schooling curves and provides no information whatsoever about the wage increase that a particular worker would get if he or she were to obtain additional schooling. If the law goes into effect, Ace's earnings would increase only from w_{DROP} to w_{ACE}, which is much less than what a high school graduate like Bob now earns (w_{HS}).

Put differently, the wage gap between Ace and Bob arises for two reasons. Bob has more schooling than Ace and is getting the returns to additional schooling. But Bob also earns more because Bob is more able. The wage differential between the two workers,

Theory at Work

DESTINY AT AGE 6

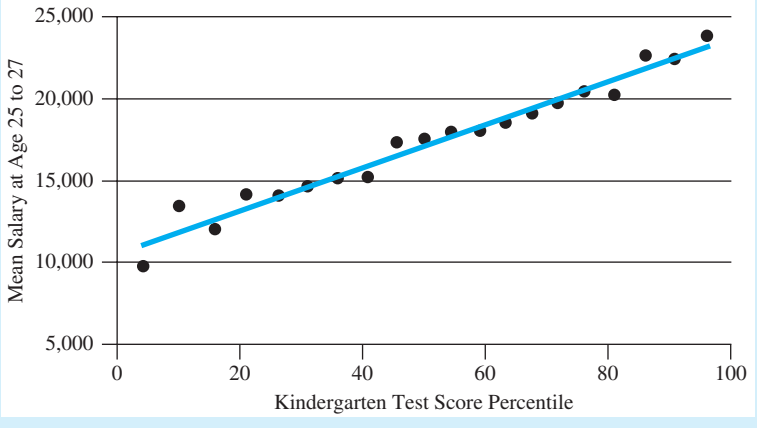
Between 1985 and 1989, 79 schools in Tennessee participated in an experiment that has greatly increased our understanding of what improves children's outcomes in schools. Project STAR randomly assigned more than 11,000 students and their teachers to different classrooms in grades K-3. Some students, for instance, were assigned to small classes, while others were assigned to large classes. At the end of the school year, all of the kindergarten students in STAR were given a grade-appropriate Stanford Achievement Test to measure their performance in math and reading.

The data collected as part of the experiment has made it possible to track these children over time and observe how they did after they entered the labor market. Remarkably, the test scores from kindergarten are highly correlated with *adult* socioeconomic outcomes.

Suppose we divide the distribution of kindergarten scores into 20 groups, representing 20 quantiles of the test score distribution. The available data lets us calculate the mean earnings for each of these groups when the children reached their mid-20s. The figure below shows a strong positive correlation between adult earnings and kindergarten test scores.

Before one concludes that a person's life earnings are predetermined at age 6, it is crucial to note what this correlation does *not* show. There is a huge amount of dispersion in outcomes within each of the 20 quantiles. Some of the kids who scored poorly in the test will do poorly as young adults, but some of those kids will do quite well. And the same kind of dispersion exists for kids who had a high score in the test. The figure "washes out" this variation. In fact, the dispersion in test scores among young children only explains 5 percent of the earnings dispersion among young adults. Nevertheless, it seems remarkable that the scores from a test given in kindergarten play even this small role 20 years later.

Source: Raj Chetty, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan, "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," *Quarterly Journal of Economics* 126 (November 2011): 1593–1660.



therefore, incorporates the impact of both education and ability on earnings and cannot be used to calculate the marginal rate of return to school.

Ability Bias

The discussion teaches us an important lesson: Earnings differentials across workers do not estimate the rate of return to school if there are unobserved ability differences.

The correlation between schooling and earnings is contaminated by ability differences and does not answer the question that initially motivated our analysis: By how much would the earnings of a particular worker increase if he were to obtain more schooling?

Why should we care about this type of **ability bias?** Suppose that an advocate observes that high school graduates earn $15,000 more per year than high school dropouts. He uses this statistic to convince policymakers that funding programs that encourage students to complete high school would increase the average wage of high school dropouts by $15,000. This earnings gain could imply that the program "funds itself" (presumably from higher tax revenues, lower expenditures on social assistance programs, and so forth).

We now know that the advocate's argument is fatally flawed. He is assuming that high school graduates and high school dropouts face the same wage–schooling locus and that one can "fix" the earnings disadvantage of dropouts by injecting them with more education. But it could be that high school graduates faced a higher wage–schooling locus. Encouraging high school dropouts to complete their high school education would not lead to a $15,000 increase in their earnings, and it would be much more difficult to argue that the program pays its way.

6-5 Estimating the Rate of Return to School

The typical method for estimating the rate of return uses data on the earnings and schooling of different workers and estimates the percentage wage differential associated with one more year of schooling—after adjusting the data for differences in other worker characteristics. The regression model is

$$\log w_i = a + b\, s_i + \text{Other variables} \qquad \textbf{(6-10)}$$

where w_i gives the wage of worker i and s_i gives the number of years of schooling acquired by the worker. The coefficient b gives the percent wage differential between two workers who differ by 1 year of schooling and is typically interpreted as the rate of return to school.[7] The regression model typical adds other variables that are known to affect a worker's earnings, including age, race, and region of residence. There are hundreds of studies that estimate variants of this regression model in the United States and in other countries. The evidence suggests that the rate of return to school in the United States was probably around 9 percent in the past two decades, so that schooling seems to be a good investment.[8]

Despite the very large number of studies that use this empirical approach, the schooling model tells us that the interpretation of the coefficient b as the rate of return to school may

[7] The coefficient measures the slope $\Delta \log w/\Delta s$. The change in the log of a variable is approximately equal to a percent change in that variable, so that b gives the percent change in earnings resulting from one more year of schooling.

[8] The classic references include Gary S. Becker and Barry R. Chiswick, "Education and the Distribution of Earnings," *American Economic Review* 56 (May 1966): 358–369; Jacob Mincer, *Schooling, Experience, and Earnings,* New York: Columbia University, 1974; and Giora Hanoch, "An Economic Analysis of Earning and Schooling," *Journal of Human Resources* 2 (Summer 1967): 310–329. International estimates of the rate of return are reported in Philip Trostel, Ian Walker, and Paul Woolley, "Estimates of the Economic Return to Schooling for 28 Countries," *Labour Economics* 9 (February 2002): 1–16.

be wrong. The percent wage differential between two workers who differ in their educational attainment can be interpreted as the rate of return to school only if there is no ability bias.[9] A more general regression model would allow for the possibility that ability shifts the wage–schooling locus and is given by

$$\log w_i = b\, s_i + \alpha A_i + \text{Other variables} \qquad \textbf{(6-11)}$$

where A_i measures the ability of worker i. Unfortunately, the large-scale data sets that are typically used to estimate the regression, such as the Current Population Surveys, do not contain any measures of a worker's ability. A few smaller, specialized data sets report the worker's IQ or a score on some aptitude test. It is doubtful, however, that these test scores correctly measure a worker's innate productive capacity or even attempt to capture all the aspects of "ability" that matter in the labor market.

Twin Studies

Several studies have pursued a clever way out of the problem raised by unobserved ability differences among workers. The ability bias would disappear if we could compare the earnings of two workers who we know have the same ability but have different educational attainment. The comparison of the earnings of a pair of identical twins, who have the same DNA, provides a setting that seems to satisfy these restrictions.

Suppose that we have a sample of identical twins where each twin reports both earnings and years of schooling. Because the twins in pair j are genetically identical, let's assume they have the same ability A. We can then difference the regression model in equation (6-11) within each pair of twins and obtain

$$\Delta \log w_j = b\, \Delta s_j + \text{Other variables} \qquad \textbf{(6-12)}$$

The regression in equation (6-12) correlates the difference in earnings within a pair of identical twins ($\Delta \log w_j$) with the difference in schooling for that pair (Δs_j). Because the twins in the pair have the same ability, the variable A cancels out when we difference the data within each pair. The regression coefficient b is then not affected by ability bias and should estimate the true rate of return to education.

Although the approach is intuitively appealing, the evidence is mixed.[10] Some early studies reported that the rate of return to school estimated in a sample of identical twins was about 3 percent, which is much lower than the rate of return typically estimated in studies that do not adjust for ability bias. More recent studies, however, find that using data on identical twins raises the rate of return to school to about 15 percent, far higher than conventional estimates.

The method of using identical twins to rid the analysis of ability bias raises an important question: If identical twins are really identical in all important ways, why exactly do

[9] Zvi Griliches, "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica* 45 (January 1977): 1–22; and Card, "The Causal Effect of Education on Earnings."

[10] Paul Taubman, "Earnings, Education, Genetics, and Environment," *Journal of Human Resources* 11 (Fall 1976): 447–461; Orley C. Ashenfelter and Alan B. Krueger, "Estimates of the Economic Return to Schooling from a New Sample of Twins," *American Economic Review* 84 (December 1994): 1157–1173; and Orley Ashenfelter and Cecilia Rouse, "Income, Schooling, and Ability: Evidence from a New Sample of Identical Twins," *Quarterly Journal of Economics* 113 (February 1998): 253–284.

they have different levels of schooling? Might the difference be attributable to environmental differences in the womb, and might these differences have an impact on the twins' eventual earnings potential?

When twins are asked why they have different schooling, for example, they sometimes reply that they had different career interests or that "one was better at books." These responses suggest that there may be productivity differences even among genetically identical twins.[11] The possibility that identical twins may not really be identical suggests that we may not be able to use the earnings gap between identical twins to calculate the "true" rate of return to school.

Instrumental Variables

Many government policies generate instruments that allow us to compare the earnings of equally able workers. One particularly famous example is provided by compulsory schooling legislation. These laws typically force students to remain in school until they reach a predetermined age, such as 16 or 17.

In the United States, children typically are not allowed to enter the first grade unless they are 6 years old by, say, August 31 of the entry year. Children born late in the year miss the deadline and are older when they start school than children born early in the year. A compulsory schooling age of 16 would then imply that children born in the latter months of the year can drop out even though they have acquired less education than children born in the early months. This variation serves as an instrument that nudges some persons along a particular wage–schooling locus and that can be used to estimate the rate of return to school.[12]

To illustrate, suppose the school year starts in September and that the school district requires that a child entering the first grade be 6 years old as of August 31. Now look at the situation of Alan and Josh, who were born a day apart in 2000. Alan was born on August 31 and Josh on September 1. Because Alan was born just before the cutoff, he can enter first grade immediately after he turns 6 on August 31, 2006. But because Josh missed the cutoff, he is only 5 years old on August 31, 2006, and must wait until the September 2007 term to start the first grade.

[11] Ashenfelter and Rouse, "Income, Schooling, and Ability: Evidence from a New Sample of Identical Twins," 273; and Orjan Sandewall, David Cesarini and Magnus Johannesson, "The Co-Twin Methodology and Returns to Schooling—Testing a Critical Assumption," *Labour Economics* 26 (January 2014): 1–10.

[12] Joshua Angrist and Alan B. Krueger, "Does Compulsory Schooling Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106 (November 1991): 979–1014. A critical appraisal of this approach is given by John Bound, David A. Jaeger, and Regina Baker, "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak," *Journal of the American Statistical Association* 90 (June 1995): 443–450. For an application of the basic framework to German and British data, see Jörn-Steffen Pischke and Till von Wachter, "Zero Returns to Compulsory Schooling in Germany: Evidence and Interpretation," *Review of Economics and Statistics* 90 (August 2008): 592–598; and Philip Oreopoulos, "Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter," *American Economic Review* 96 (March 2006): 152–175.

	Born on:	Can enter first grade in:	Can drop out on:	Completed years of school at time of dropping out:
Alan	August 31, 2000	September 2006 term	August 31, 2016	10 years
Josh	September 1, 2000	September 2007 term	September 1, 2016	9 years

The compulsory schooling age is 16, so that Alan and Josh will be able to drop out in 2016 at almost exactly the same time. But Alan entered school a year earlier, so he will have completed 10 years of school as of the legal dropout date, while Josh will have completed only 9 years. The school entry age, combined with the compulsory schooling legislation, combine to create an instrumental variable. If we assume that the average ability of children born on August 31 is the same as that of children born on September 1, the legislation nudges some persons along the same wage–schooling locus, creating variation in schooling and earnings.

Put differently, the biological accident of being born just before midnight on August 31 means that the child will be required to be in school for a longer period than a comparable child born in the early morning hours of September 1. The wage gap between the two groups of children, therefore, measures the true rate of return to school because there should be no ability difference between them (on average). The only reason that earnings differ is because those born in late August have slightly more schooling than those born in early September. If one controls for ability bias in this fashion, the estimated rate of return to school is about 7.5 percent.

Another clever example of how government policies create instrumental variables that allow us to estimate the rate of return exploits the 1968 student riots that brought France to a standstill.[13] In May 1968, after months of simmering conflict between students and university administrators, the administrators decided to close the University of Nanterre in Paris on May 2. The resulting protests expanded to other university towns in France and eventually brought the workers out into the streets. Roughly, 10 million workers (or two-thirds of the French workforce) joined the general strike in support of the students.

Because these events took place at the end of the school year, an important component of the negotiations between the students and the government involved the issue of how to deal with the university exams that determine the academic future of French students. One particularly important exam is the *baccalauréat,* an exam that effectively signals the successful completion of a secondary education and opens the door to higher education. Normally, the *baccalauréat* requires several days of written and oral exams. In 1968, however, French authorities acquiesced to a revised *baccalauréat* that only involved oral exams and took place in 1 day.

Because of the relaxed requirements, a relatively large number of the affected students obtained their *baccalauréat.* The higher pass rate allowed a large fraction of French students *in that age cohort* to continue their education. The 1968 riots, in effect, created a valid instrument. It is unlikely that the average ability of the 1968 cohort differs from that

[13] Eric Maurin and Sandra McNally, "Vive la Revolution! Long-Term Educational Returns of 1968 to the Angry Students," *Journal of Labor Economics* 26 (January 2008): 1–33.

of adjacent cohorts. But *la revolution* nudged that cohort along the wage–schooling locus, giving them more schooling and higher earnings.

Roughly, about 20 percent of the cohort obtained higher degrees as compared to 17 percent of the adjacent cohorts. Moreover, the earnings of the cohort affected by the 1968 riots were around 3 percent higher. The implied rate of return to school was around 14 percent.

6-6 Policy Application: School Construction in Indonesia

Many studies find that the wage gap between highly educated and less-educated workers in developing countries is even higher than the gap in industrialized economies. It is tempting to infer from these findings that developing labor markets offer a high rate of return to school and that these high rates of return justify sizable investments in the education infrastructure. As we have seen, however, these wage gaps need not suggest that increasing schooling opportunities for a wide segment of the population would substantially improve the earnings of those workers.

In Indonesia, children typically go to school between the ages of 7 and 12. In 1973, the Indonesian government launched a major school construction program (INPRES) designed to increase the enrollment of children in disadvantaged areas.[14] By 1978–1979, more than 61,000 new primary schools had been built, approximately two schools per 1,000 children. This program cost nearly $950 million (2017 U.S. dollars), representing 1.5 percent of Indonesian GDP as of 1973. As a way of grasping the scale, a similar commitment by the United States today (in terms of GDP share) would cost $275 billion. The results of the construction program were immediate: Enrollment rates among Indonesian children aged 7–12 rose from 69 percent in 1973 to 83 percent by 1978.

A well-known study used data drawn from the Indonesian labor market in 1995 (two decades after the school construction) to determine if the huge investment increased the educational attainment and earnings of the targeted Indonesians and to calculate the rate of return to school. As noted above, the program attempted to equalize education opportunities across the various regions of Indonesia, building more schools in those parts of Indonesia that had relatively low enrollment rates. Table 6-3 shows how education and earnings were affected for persons residing in two different parts of Indonesia—the "high-construction" area, where many new schools were built, and the "low-construction" area, where few schools were built. About one more school per 1,000 children was built in the high-construction area than in the low-construction area.

The table reports the outcomes experienced by two different demographic groups: Persons who were 2–6 years old and 12–17 years old as of 1974. The younger group was affected by the construction program. Those boys and girls were about to enter school as the construction program began, and they form the treatment group. The older children form the control group; they were past the school-going age and their educational attainment should not be affected by the presence of more schools.

[14] The discussion is based on Esther Duflo, "Schooling and Labor Market Consequences of School Construction in Indonesia," *American Economic Review* 91 (September 2001): 795–813.

TABLE 6-3 The Impact of School Construction on Education and Wages in Indonesia

Source: Duflo, "Schooling and Labor Market Consequences of School Construction in Indonesia."

	Years of Education			Log Wages		
	Persons Aged 12–17 in 1974	Persons Aged 2–6 in 1974	Difference	Persons Aged 12–17 in 1974	Persons Aged 2–6 in 1974	Difference
High-construction area	8.02	8.49	0.47	6.87	6.61	−0.26
Low-construction area	9.40	9.76	0.36	7.02	6.73	−0.29
Difference-in-differences	—	—	**0.11**	—	—	**0.03**

We can use the difference-in-differences methodology to calculate the impact of the program on the targeted population. In the low-construction area, educational attainment rose by 0.36 year between the younger and older groups, while in the high-construction area, the relative educational attainment rose by 0.47 year. The additional construction, therefore, increased educational attainment by 0.11 year. By using a similar approach, the table shows that the earnings of the younger group rose by 3 percent.

We can now apply the method of instrumental variables to calculate the rate of return to school in Indonesia. The instrument is school construction. This variable clearly nudged some students along the wage–schooling locus. The instrument is valid if students in the high-construction areas have the same ability as those in the low-construction areas and if the older group of students has the same ability as the younger group. An additional 0.11 year of schooling increased earnings by 3 percent. This implies that each additional year of school increased earnings by 27 percent (or $0.03 \div 0.11$). The rate of return to school in Indonesia, therefore, seems quite high, justifying the sizable cost of the school construction program. In fact, a more thorough analysis of the data, which controls for many of the other factors that also affected trends in educational attainment and wages in Indonesia, suggests that the rate of return to school may be as high as 10 percent.

6-7 Policy Application: The Education Production Function

Spending on public education in the United States increased dramatically in recent decades. In 1980, per-student expenditures at the primary and secondary levels was about $7,500 (in 2016 dollars). By 2013, the per-student cost was over $13,300, implying a total expenditure of over $600 billion.[15] Not surprisingly, this has led some observers to ask if "throwing money" at the school system improves student outcomes. Put differently, does student achievement or their eventual labor market performance improve because of increases in teacher salaries or reductions in the pupil/teacher ratio?

These policy questions have motivated a lot of research that attempts to estimate the **education production function**

$$Y = f(x_1, x_2, x_3, \ldots) \qquad \textbf{(6-13)}$$

[15] National Center for Education Statistics, *Digest of Education Statistics,* Tables 236.10 and 236.55.

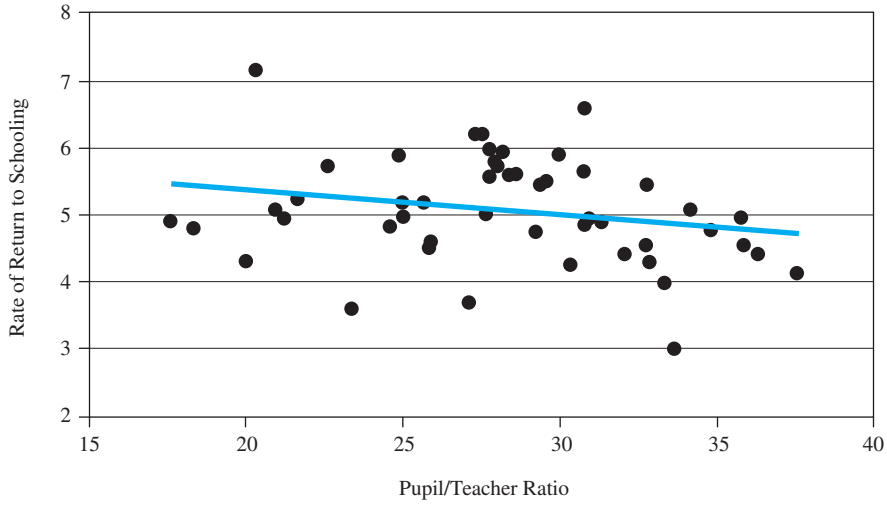
where Y is some measure of student achievement, such as a score in a standardized test or the earnings of the student after he enters the labor market. The x's represent the inputs in production, including the student/pupil ratio, the educational credentials of teachers, and the socioeconomic characteristics of the students' peers.

Prior to the 1990s, the consensus was that high levels of school expenditures had little impact on student outcomes. As an influential survey concluded, "There appears to be no strong or systematic relationship between school expenditures and student performance."[16] There has been an explosion of recent research, however, that pays a lot of attention to the data issues involved in estimating the education production function. This new research sometimes suggests that expenditures in the school system today might indeed improve the outcomes of students later in life.

Figure 6-6 illustrates some of this evidence.[17] The decennial census data reports the current earnings of workers born in a particular state r. We can then run the regression model in equation (6-10) using this subsample of workers. The regression coefficient presumably measures the rate of return to school in state r. By replicating the exercise for each of the 50 states, we get a complete picture of how the rate of return differs across states, allowing us to determine if the variation is related to differences in school quality. The figure shows the relation between the estimated rate of return and the average pupil/teacher

FIGURE 6-6 The Rate of Return and School Quality

Source: David Card and Alan B. Krueger, "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy* 100 (February 1992), Tables 1 and 2. The data in the graphs refer to the rate of return to school and the school quality variables for the cohort of persons born in 1920–1929.



[16] Eric A. Hanushek, "The Economics of Schooling: Production and Efficiency in the Public Schools," *Journal of Economic Literature* 24 (September 1986): 1141–1177.

[17] David Card and Alan B. Krueger, "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy* 100 (February 1992): 1–40.

ratio in the state at the time the students went to school. Children born in states that offered better schools, at least as measured by a small class size, had a higher rate of return to school after they entered the labor market.[18]

Much of the recent research moves beyond estimating correlations by exploiting natural experiments that change class size to see if the eventual outcomes of students are affected by random variation in the classroom environment. In effect, these studies are searching for an instrumental variable that affects class size but does not affect other outcomes directly. One well-known study used an instrument based on the interpretation of the Talmud by the twelfth-century rabbinic scholar Maimonides.[19] According to Maimonides's rule, "Twenty-five children may be put in charge of one teacher. If the number in the class exceed twenty-five but is not more than forty, he should have an assistant to help with the instruction. If there are more than forty, two teachers must be appointed."

The Israeli school system uses Maimonides's rule to allocate students to classes. The maximum class size is 40. According to Maimonides's rule, class size increases with enrollment until 40 pupils are enrolled. An extra student, however, implies that class size drops sharply to 20.5. Because there is little reason to suspect that the shift from a class size of 40 to one of 20.5 has anything to do with the underlying ability of the students, Maimonides's rule provides a valid instrument—it shifts class size without affecting any other variables. The analysis of the outcomes experienced by Israeli students suggests a negative relation between class size and test scores for fourth and fifth graders.[20]

A few studies examine data from real-world experiments where some children are purposefully assigned to small classes and some children are not. Beginning in 1985, the Tennessee Student/Teacher Achievement Ratio (STAR) experiment randomly assigned kindergarten students *and* their teachers to small classes (with a pupil/teacher ratio of 13–17) or to larger classes (with a ratio of 22–25 students). After the initial assignment, students remained in the same class type for 4 years. Between 6,000 and 7,000 students were involved in this experiment. Some evaluations of the STAR data indicate that students assigned to the small classes scored higher in achievement tests than those assigned to the larger classes. Further, the students assigned to the small classes were more likely to go to college.[21] Other evaluations of the same data, however, find little relation between class size and student achievement.[22]

[18] See also Julian R. Betts, "Does School Quality Matter? Evidence from the National Longitudinal Survey," Review of Economics and Statistics 77 (May 1995): 231–250; and James J. Heckman, Anne S. Layne-Farrar, and Petra E. Todd, "Does Measured School Quality Really Matter," in Gary Burtless, editor, *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success,* Washington, D.C.: The Brookings Institution, 1996.

[19] Joshua D. Angrist and Victor Lavy, "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics* 114 (May 1999): 533–575.

[20] A related study in the U.S. context, however, finds no relation between class size and student achievement; see Carolyn M. Hoxby, "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *Quarterly Journal of Economics* 115 (November 2000): 1239–1285.

[21] Alan B. Krueger, "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics* 114 (May 1999): 497–532; and Susan Dynarski, Joshua M. Hyman, and Diane Whitmore Schanzenbach, "Experimental Evidence on the Effect of Childhood Investment on Postsecondary Attainment and Degree Completion," *Journal of Policy Analysis and Management* 32 (September 2013): 692–717.

[22] Eric A. Hanushek, "Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects," *Educational Evaluation and Policy Analysis* 21 (Summer 1999): 143–163.

Theory at Work

BOOKER T. WASHINGTON AND JULIUS ROSENWALD

Black children in the rural south faced extremely limited opportunities to go to school in the early part of the twentieth century, with an inadequate and decaying educational infrastructure in terms of school buildings, classrooms, and teaching personnel. Not surprisingly, there was a three-year gap in the educational attainment of black and white children.

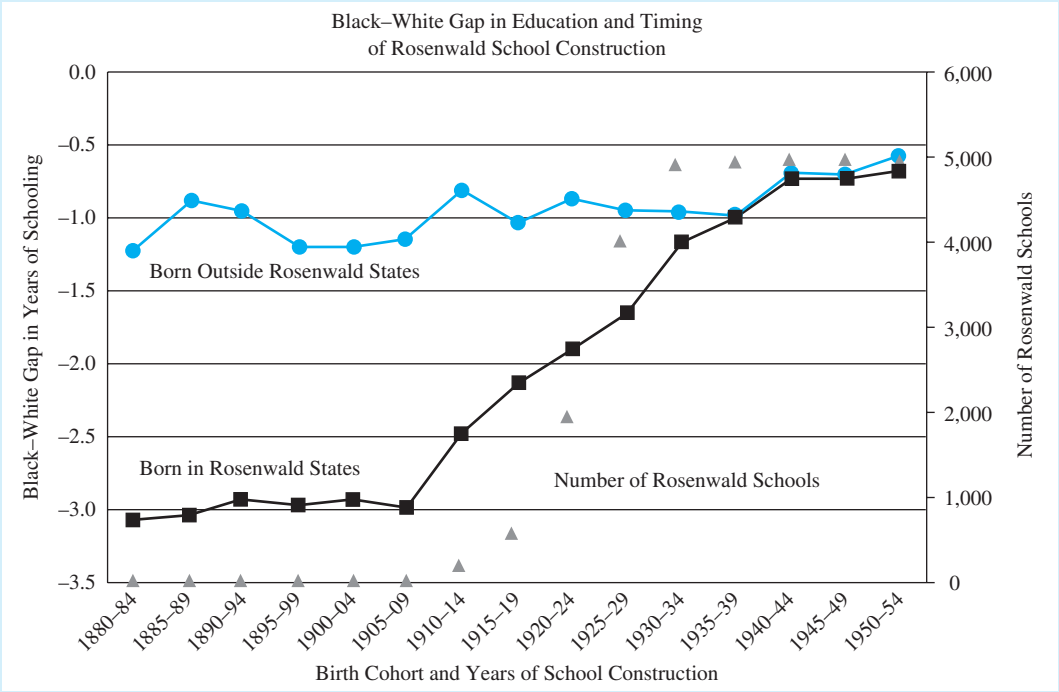
Frustrated by the lack of progress, Booker T. Washington, who led the Tuskegee Institute in Alabama, sought out a number of northern philanthropists to privately fund the building of schools in the rural south. One of these philanthropists was Julian Rosenwald, a Chicago area businessman who played a central role in the founding of the Sears-Roebuck Company. The first six schools were built near Tuskagee in 1913. By 1920, 716 schools had been built in 11 southern states. And nearly 5,000 new schools had been built by the time this remarkable initiative concluded in 1932.

The Rosenwald program did more than just build schools. It also improved the quality of the physical and human capital available for the education of black children. The schools had proper lighting and sanitation; they were equipped with the books, chairs, and desks one would expect to find in a successful educational environment; and there were related initiatives that set minimum teacher salaries and that provided training programs to recruit and prepare teachers.

This program led to rapid growth in the relative educational attainment of black children. As shown in the figure below, the black–white education gap did not narrow among children born in states that did not have Rosenwald schools. But the gap narrowed by 3 years among children born in the states covered by the Rosenwald program, and the timing of the narrowing coincided with the building of the Rosenwald schools.

By using detailed information on the location and timing of the Rosenwald schools, it is possible to identify the impact on the affected schoolchildren. Black children who had direct access to a Rosenwald school completed about a year of school more than black children who did not. A comparison of the eventual salary gains for the Rosenwald children implies a rate of return to school of about 7–9 percent.

Source: Daniel Aaronson and Bhashkar Mazumder, "The Impact of Rosenwald Schools on Black Achievement," *Journal of Political Economy* 119 (October 2011): 821–888.



Finally, there has been some interest in determining if attending an "elite" college or university, which presumably offers a higher quality of education, affects earnings. The problem with comparing the earnings of students who attend selective institutions with the earnings of those who do not is that there may be underlying ability differences between the two groups. Any resulting wage gap may have little to do with the "value added" by the selective institution and may simply reflect a preexisting ability gap between the two groups of students.

One way of minimizing the ability bias is to compare students who attend highly selective schools with students who were accepted by those institutions but decided to go to a less-selective college.[23] These two groups of students presumably have the same underlying ability; they were all accepted by the same selective schools, after all. Interestingly, this comparison reveals that selective schools provide no value-added: Students who graduate from selective schools earn no more than students who were accepted by those schools but decided to go elsewhere.

6-8 Do Workers Maximize Lifetime Earnings?

The schooling model provides the conceptual framework that allows us to estimate the rate of return to school. It specifies the conditions under which percent wage differentials among workers who differ in their education can be interpreted as a rate of return. This calculation, however, does *not* test the theory. Rather, it uses the theory to interpret earnings differences in a particular way.

We still want to determine if the schooling model provides a useful "story" of how workers go about the business of deciding how much schooling to get. The model assumes that persons choose the level of schooling that maximizes the present value of lifetime earnings. It would be easy to test the validity of this assumption if we could observe the age–earnings profile of a particular worker both if he were to go to college and if he were to stop after high school. We would calculate the present value for each schooling option, compare the numbers, and see if the worker chose the one with the largest present value.

This simple test, however, can *never* be conducted. The reason is both trivial (because it is painfully obvious) and profound (because it raises a number of conceptual questions that have yet to be adequately resolved). *Once a worker makes a particular choice, we can only observe the earnings stream associated with that choice.* Consider the group of workers who go to college. We observe their life cycle earnings after college graduation, but we will never observe what they would have earned had they not attended college. Similarly, consider the group of workers who quit after completing high school. We observe their earnings stream after high school graduation, but we will never observe what they would have earned had they gone on to college.

It is tempting to propose a simple solution to this problem. Even though we can never observe how much a worker who quits after completing high school would have earned if he had attended college, we do observe the earnings of those who did attend college.

[23] Stacy Berg Dale and Alan B. Krueger, "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables," *Quarterly Journal of Economics* 117 (November 2002): 1491–528. See also Dan A. Black and Jeffrey A. Smith, "Estimating the Returns to College Quality with Multiple Proxies for Quality," *Journal of Labor Economics* 24 (July 2006): 701–728.

We could then predict the high school graduate's earnings had he attended college by using the data on what college graduates actually earn. Similarly, even though we do not observe how much college graduates would have earned had they stopped after high school, we do observe the earnings of high school graduates. We could then predict the college graduate's earnings (had he not attended college) from the salary data for high school graduates.

The schooling model implies that this exercise is valid only if college graduates and high school graduates have the same wage–schooling locus. The exercise is invalid if there are ability differences between the two groups. The observed wage differential between college graduates and high school graduates then reflects not only the returns to college but also the ability differential between the groups.

A Numerical Example

To illustrate, let's work through a simple numerical example with two workers, Willie and Wendy. Willie is particularly adept at "blue-collar" work, and this type of work requires no schooling. Wendy is particularly adept at "white-collar" work, and this type of work requires 1 year of schooling. There are two periods in the life cycle. If a person does not go to school, he works in the blue-collar job in both periods. If the person goes to school, he goes to school in the first period and works in the white-collar job in the second.

The wage–schooling locus for each worker is

Worker	Earnings in Blue-Collar Job	Earnings in White-Collar Job
Willie	$20,000	$40,000
Wendy	$15,000	$41,000

Because Willie is better at doing blue-collar work, he earns more at the blue-collar job ($20,000) than Wendy would ($15,000). And because Wendy is better at white-collar work, she earns more in the white-collar job than Willie would.

Both Willie and Wendy have a discount rate of 10 percent. Each calculates the present value of lifetime earnings for each schooling option and chooses the one that has the highest present value. Let's also assume that net income is zero during the schooling period if the worker chooses to go to school. The present values of Willie's alternative earnings streams are

$$\text{Willie's present value if he does not go to school} = 20{,}000 + \frac{20{,}000}{1.1} = \$38{,}182 \qquad \textbf{(6-14)}$$

$$\text{Willie's present value if he goes to school} = 0 + \frac{40{,}000}{1.1} = \$38{,}364 \qquad \textbf{(6-15)}$$

If Willie chooses the level of schooling that maximizes the present value of earnings, he will not go to school and become a blue-collar worker.

The present values of Wendy's alternative earnings streams are

$$\begin{matrix}\text{Wendy's present value if}\\ \text{she does not go to school}\end{matrix} = 15{,}000 + \frac{15{,}000}{1.1} = \$28{,}636 \qquad \textbf{(6-16)}$$

$$\text{Wendy's present value if she goes to school} = 0 + \frac{41{,}000}{1.1} = \$37{,}273 \qquad \textbf{(6-17)}$$

If Wendy maximizes the present value of earnings, she will go to school in the first period and work at the white-collar job in the second.

What data do we observe in the labor market? We observe the earnings of blue-collar workers who do not go to school (like Willie). The present value of their earnings is $38,182. We also observe the earnings of white-collar workers who do go to school (like Wendy). The present value of their earnings is $37,273.

A comparison of these two numbers seems to suggest that Wendy made a terrible mistake. In our example, persons who go to school earn less over their lifetime than persons who do not. But this is an irrelevant comparison; both Willie and Wendy, in fact, made the right choice. The problem arises because we are comparing the earnings of two very different groups of workers. This comparison is akin to comparing apples and oranges and is contaminated by **selection bias,** the fact that workers self-select themselves into jobs for which they are best suited. The selection bias leads to an incorrect rejection of the income-maximization hypothesis.

Selection Bias Corrections

Given the importance of selection bias in the labor market—we all self-select ourselves into those job locations, occupations, and career paths that best suit us—it is not surprising that a lot of attention has been devoted to this problem. The effort has led to statistical techniques, known as "selection bias corrections," that allow us to purge the data of the bias.[24] Specifically, the methodology lets us correctly predict what a high school graduate would have earned had he attended college, and what a college graduate would have earned had he quit school after getting a high school diploma.

A well-known study uses the selection bias correction to estimate the life cycle earnings profiles associated with each of two alternatives (going to college or quitting after high school) for a large number of workers.[25] The empirical analysis confirms the basic hypothesis of the theory: On average, workers chose the schooling option that maximized the present value of lifetime earnings. Interestingly, the evidence indicates that when both a high school graduate and a college graduate are placed in the type of job that high school graduates typically fill, the high school graduate would be more productive. Conversely, if both high school graduates and college graduates were placed in jobs typically filled by college graduates, the college graduate would be more productive.

This result suggests that the notion that there is only one type of ability that inevitably leads to higher earnings does not correctly describe how workers differ in the labor market. There exist various types of abilities, and each of us may be particularly adept at doing some things and quite inept at doing others. Some persons have a knack for doing work that is best learned in college, whereas others have a knack for work that is best learned elsewhere.

[24] James J. Heckman, "Sample Selection Bias as a Specification Error," *Econometrica* 47 (January 1979): 153–162.

[25] Robert J. Willis and Sherwin Rosen, "Education and Self-Selection," *Journal of Political Economy* 87 (October 1979 Supplement): S7–S36. See also Lawrence W. Kenny, Lung-Fei Lee, G. S. Maddala, and R. P. Trost, "Returns to College Education: An Investigation of Self-Selection Bias Based on the Project Talent Data," *International Economic Review* 20 (October 1979): 775–789; and John Garen, "The Returns to Schooling: A Selectivity-Bias Approach with a Continuous Choice Variable," *Econometrica* 52 (September 1984): 1199–1218.

6-9 Signaling

The schooling model is based on the idea that education increases a worker's productivity and that this increase in productivity raises wages. An alternative argument is that education does not increase the worker's productivity at all, but that "sheepskin" levels of education (such as a high school or college diploma) signal a worker's innate ability to potential employers.[26] Education then increases earnings not because it increases productivity, but because it signals that the worker is cut out for "smart" work. Obviously, education can play this signaling role only when it is difficult for employers to observe the worker's ability directly. The firm would not have to rely on third-party certifications if it was easy to determine if a particular worker was qualified.

To illustrate, let's work through a simple numerical example. There are two types of workers, low-productivity and high-productivity workers. The productivity distribution in the population is given by

Type of Worker	Proportion of Population	Present Value of Lifetime Productivity
Low-productivity	q	$200,000
High-productivity	$1 - q$	300,000

A worker is randomly assigned to one of the two groups at birth, *and* the worker knows which group he was assigned to. Once the assignment takes place, the worker's productivity is immutable. Crucially, productivity has nothing to do with how much schooling a particular worker gets.

If an employer could determine easily if a job applicant is a high-productivity worker, he would pay the worker $300,000 over the life cycle. After all, if the employer's wage offer did not match the high-productivity applicant's true value, the job applicant would simply go elsewhere, where his high productivity was better appreciated and rewarded. And if the employer could determine easily that the applicant is a low-productivity worker, he would pay the worker only $200,000.

But life is not this easy. Even though a particular worker knows which group he belongs to, it might take a while and be quite costly for the employer to learn that. There is **asymmetric information** in the labor market, where one of the parties in the transaction knows more about the terms of the contract. Moreover, if the employer were to ask an applicant if he is a low- or a high-productivity worker, the applicant (who wants a high salary) will always reply that he is a high-productivity worker. When a job applicant shows up at the firm, therefore, there is a lot of uncertainty about that worker's potential contribution to the firm.

[26] A. Michael Spence, "Job Market Signaling," *Quarterly Journal of Economics* 87 (August 1973): 355–374. See also Kenneth J. Arrow, "Higher Education as a Filter," *Journal of Public Economics* 2 (July 1973): 193–216; and Joseph Stiglitz, "The Theory of Screening, Education, and the Distribution of Income," *American Economic Review* 65 (June 1975): 283–300.

Pooling Equilibrium

Because low-productivity applicants will lie about their productivity, the firm disregards what anyone says about their qualifications. The employer then simply pools all job applicants and treats them identically. The average salary of the workers hired by the firm is

$$\begin{aligned}\text{Average salary} &= (200{,}000 \times q) + [300{,}000 \times (1 - q)] \\ &= 300{,}000 - 100{,}000q\end{aligned} \tag{6-18}$$

The average salary is a weighted average of the workers' productivities, where the weights are the proportions in the population that belong to each productivity group. Because the proportion q is between 0 and 1, the average salary in the **pooling equilibrium** is between $200,000 and $300,000.

Low-productivity workers prefer this equilibrium because they are being pooled with more productive workers, and get their salary pushed up. But neither employers nor high-productivity workers like the pooling equilibrium. Employers are mismatching workers and jobs. Some high-productivity workers are assigned to menial tasks, and low-productivity workers are placed in jobs they are not qualified to perform. The mismatching reduces the firm's efficiency and profits. Similarly, the high-productivity workers see their potential earnings dragged down by the low-productivity workers. High-productivity workers would like to find a way of demonstrating that they truly are more productive.

Separating Equilibrium

High-productivity workers have an incentive to provide *and* firms have an incentive to take into account credible information that can be used to allocate the worker into the correct productivity group. This type of information is called a **signal.** It turns out that an educational diploma or certificate can perform this signaling job and that it can perform the task with absolute precision. *No mismatches occur.*

Suppose a firm chooses the following rule of thumb for allocating workers to the two types of jobs. If a worker has at least $\overline{y}$ years of college, the firm assumes that the worker is a high-productivity worker, allocates him to a job that requires a high level of skills, and pays him a (lifetime) salary of $300,000. If a worker has fewer than $\overline{y}$ years of college, the firm assumes that the worker is a low-productivity worker, allocates him to an unskilled job, and pays him a salary of $200,000.

Because employers are willing to pay more to workers who get at least $\overline{y}$ years of college, all workers will want to get the required college credits. Obtaining these credits, however, is expensive. We assume that obtaining credits is more expensive for less-able workers. In particular, a year's worth of college credits costs $20,000 for a high-productivity worker, but $25,001 for a low-productivity worker.

Obviously, tuition and fees do not differ according to ability, but the real cost of a college credit may be higher for a low-productivity worker. To attain a particular level of achievement, a low-productivity worker may have to study longer and pay for tutors, study guides, and special classes. The assumption that low-productivity workers find it costlier to obtain the signal is the fundamental assumption of the signaling model.

Given the firm's wage offer, workers must now decide how many years of college to get. A **separating equilibrium** occurs when low-productivity workers choose not to get $\overline{y}$ years of schooling and voluntarily signal their low productivity, and high-productivity workers choose to get at least $\overline{y}$ years of schooling and separate themselves from the pack.

Figure 6-7*a* illustrates the firm's wage offer and the cost function facing a low-productivity worker. The wage offer is such that if the worker has fewer than $\overline{y}$ years of college, he earns \$200,000, and if he has $\overline{y}$ or more years, he earns \$300,000. The cost function is upward sloping and has a slope of \$25,001 because each additional year of college costs \$25,001 for a low-productivity worker.

In our numerical example, a worker will always decide either not to go to college at all or to go to college for exactly $\overline{y}$ years. After all, a worker's earnings do not increase if he goes to college more than $\overline{y}$ years, yet it costs \$25,001 to get an additional year's worth of college credits. Similarly, because the worker's lifetime salary is \$200,000 for any level of education between 0 and $\overline{y}$ years, there is no point to getting "just a few" credits.

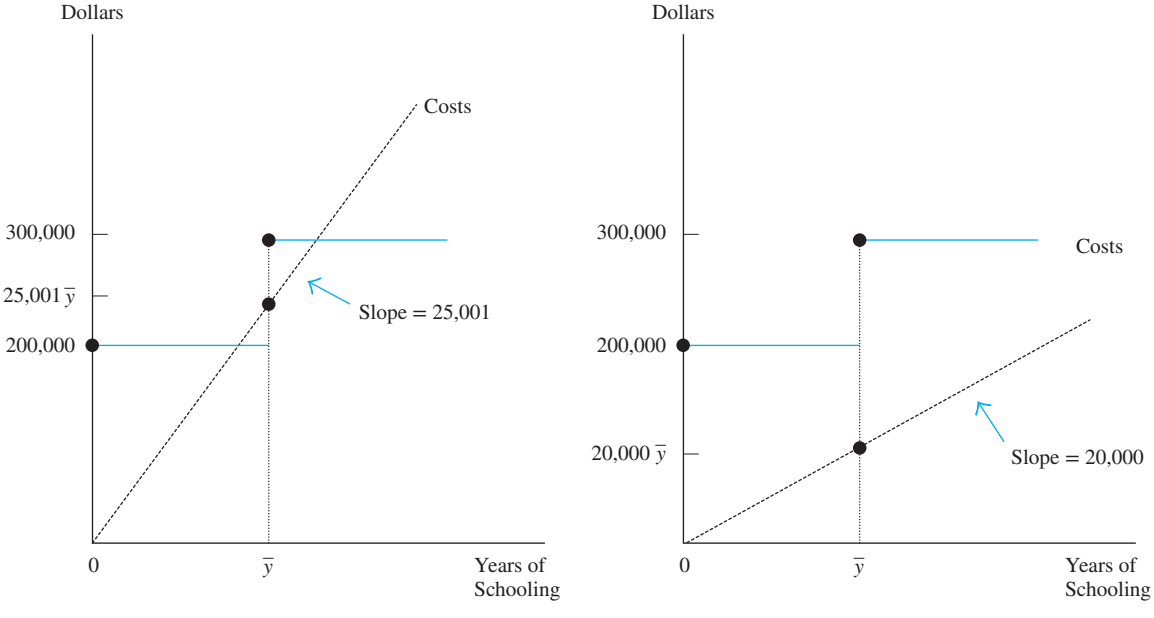
A separating equilibrium requires that low-productivity workers do not go to college at all. This will occur whenever the net return from getting zero years of college exceeds the net return from getting $\overline{y}$ years. Figure 6-7*a* indicates that when a low-productivity worker does not go to college, he takes home \$200,000 (because he does not incur any education costs). If he goes to college $\overline{y}$ years, his net salary is the vertical difference between the \$300,000 wage offer and the cost of going to college for $\overline{y}$ years (which equals \$25,000 × $\overline{y}$).

Therefore, the low-productivity worker will not attend college if

$$\$200{,}000 > \$300{,}000 - (\$25{,}001 \times \overline{y}) \qquad \textbf{(6-19)}$$

FIGURE 6-7 A Separating Equilibrium

Workers get paid \$200,000 if they get less than $\overline{y}$ years of college and \$300,000 if they get at least $\overline{y}$ years. Low-productivity workers find it expensive to invest in college and will not get $\overline{y}$ years. High-productivity workers do get $\overline{y}$ years. The worker's education signals if he is a low-productivity or a high-productivity worker.



(*a*) Low-Productivity Workers

(*b*) High-Productivity Workers

Solving for $\overline{y}$ implies that

$$\overline{y} > 3.999 \qquad \textbf{(6-20)}$$

If the firm uses the rule of thumb that only workers who get more than 3.999 years of college will be considered high-productivity workers, *no* low-productivity worker will ever go to college; it is just too expensive. By choosing not to attend college, low-productivity workers "voluntarily" signal their low productivity and separate themselves out.

A separating equilibrium also requires that high-productivity workers *do* get $\overline{y}$ years of college. Figure 6-7*b* illustrates their decision. The net salary of a high-productivity worker who does not go to college is \$200,000. The net salary if he goes to college for $\overline{y}$ years is the vertical difference between the \$300,000 wage offer and the cost of going to college (or $\$20{,}000 \times \overline{y}$). Therefore, high-productivity workers get $\overline{y}$ years whenever

$$\$200{,}000 < \$300{,}000 - (\$20{,}000 \times \overline{y}) \qquad \textbf{(6-21)}$$

Solving for $\overline{y}$ yields

$$\overline{y} < 5 \qquad \textbf{(6-22)}$$

As long as the firm does not demand "too many" years of higher education (such as a master's degree or Ph.D.), high-productivity workers go to college and voluntarily signal that they are highly productive.

Putting together both conditions implies that low-productivity workers do not go to college and that high-productivity workers do whenever

$$3.999 < \overline{y} < 5 \qquad \textbf{(6-23)}$$

A firm can choose any hiring standard in this range and generate a separating equilibrium. For instance, the firm can announce that workers with more than 4.5 years of college will be considered high-productivity workers, and the two types of workers will sort themselves out accordingly. There seem to be an infinite number of valid thresholds that the firm can use (4 years of college, 4.5 years, 4.666 years, 4.997 years, and so on). Not all of these solutions, however, can survive the competitive pressures of the marketplace. Suppose some firms set 4.333 years as the threshold, while other firms use 4.000 years. High-productivity workers would prefer the firm with the 4.000 threshold because both firms pay the same competitive salary (of \$300,000) and high-productivity workers have nothing to gain by getting more education than the minimum required. The competitive solution, therefore, is the smallest possible threshold, so that using a college diploma (4 years of college) to separate out job applicants generates a separating equilibrium.

The signaling model shows that education can play the role of signaling the worker's innate ability without increasing the worker's productivity. This insight has important policy implications. The human capital model, for example, suggests that human capital investments, such as education, provide a way out of low incomes and poverty. Indeed, the rationale behind government programs that subsidize tuition expenses or that improve the quality of public schools is that these programs increase the human capital stock of the targeted groups. The signaling model says that education does not really increase a worker's

innate productivity. Low-productivity workers remain low-productivity workers despite the billions spent on these government programs.

Regardless of which model is correct, an outsider looking at a particular labor market will observe that more-educated workers earn higher wages. Because both the schooling model and the signaling model predict this positive correlation, it has proven difficult to establish empirically which of the two mechanisms are more important.[27]

The GED as a Signal

One estimate of the signaling value of education exploits the peculiarities of acquiring the General Equivalency Diploma (GED) credential.[28] The GED was first introduced in World War II to help veterans without a high school diploma earn an equivalent credential. It has since become the primary mechanism through which persons who drop out of high school can eventually acquire a high school equivalent credential. About 1 million teenagers drop out of school each year, and a third of them eventually acquire the credential.

To get the GED, the dropout needs to take a battery of exams that test his proficiency in mathematics, writing, social students, science, and literature. The GED is awarded to the dropouts who get a passing grade in the exams.

Interestingly, although the exam is administered nationally by the GED Testing Service, each state can set its own passing grade. There is a lot of variation in this threshold. New York and Florida, for example, set a relatively high passing grade, making it "hard" to pass and get the GED. Other states, including Minnesota and New Hampshire, set a relatively low passing grade, making it "easy" to get the GED.

The interstate variation in the passing grade means that a dropout who scores x points might be awarded the GED credential in one state but not in another, and this creates an interesting natural experiment. Consider a sample of dropouts who took the GED exam and scored exactly x points. In rough terms, the dropouts in this sample are equally able—at least in terms of whatever it is that the GED exam measures. Because of the different passing grades used by different states, however, some of the dropouts were awarded the GED, but others were not. This allows us to determine how earnings react to the receipt of a shiny new credential.

Table 6-4 uses a simple difference-in-differences exercise to establish that the GED has a substantial signaling value. It reports the 1995 annual earnings of different groups of high school dropouts who took the GED exam in 1990. The workers who got a low score and happened to live in states with a high passing grade did not get the GED credential, and earned about $7,800 5 years after the exam. In contrast, the workers who had done just

[27] Kevin Lang and David Kropp, "Human Capital versus Sorting: The Effects of Compulsory Schooling Laws," *Quarterly Journal of Economics* 101 (August 1986): 609–624; Eugene A. Kroch and Kriss Sjoblom, "Schooling as Human Capital or a Signal: Some Evidence," *Journal of Human Resources* 29 (Winter 1994): 156–180; and Kelly Bedard, "Human Capital versus Signaling Models: University Access and High School Dropouts," *Journal of Political Economy* (August 2001): 749–775.

[28] John H. Tyler, Richard J. Murnane, and John B. Willett, "Estimating the Labor Market Signaling Value of the GED," *Quarterly Journal of Economics* 115 (May 2000): 431–468. For evidence on whether a high school diploma has any signaling value, see David A. Jaeger and Marianne E. Page, "Degrees Matter: New Evidence on Sheepskin Effects in the Returns to Education," *Review of Economics and Statistics* 78 (November 1996): 733–740; and Damon Clark and Paco Martorell, "The Signaling Value of a High School Diploma," *Journal of Political Economy* 122 (April 2014): 282–318.

TABLE 6-4 The Signaling Value of the GED

Source: John H. Tyler, Richard J. Murnane, and John B. Willett, "Estimating the Labor Market Signaling Value of the GED," Quarterly Journal of Economics 115 (May 2000): 446.

	Average Salary of Test-Takers by State's Passing Threshold		
Applicant's Test Score	Easy-to-Pass Exam	Hard-to-Pass Exam	Difference
Low	$9,628	$7,849	$1,779
High	$9,981	$9,676	$305
Difference-in-differences			**$1,473**

as badly in the exam, but who happened to live in "easy" states and got the GED credential, earned quite a bit more, about $9,600. The credential, it seemed, increase the earnings of the low-scoring workers by $1,800.

But this earnings gap may partly reflect earnings differences that exist for all workers between the "easy" and the "hard" states. We need a control group to net out this regional wage difference, and a sensible control group might be the dropouts who also took the GED exam at the same time but who got a high score, so they got the GED credential regardless of where they lived. The workers in this group earned about $10,000 regardless of which state they lived in. As the table shows, the interstate earnings gap in this group is only about $300.

If we adjust for this regional wage difference, the difference-in-differences methodology implies that the GED credential increased the earnings of low-scoring workers by about $1,500, or about a 19 percent increase in earnings. In short, the GED credential matters, and it seems to lead to substantially higher earnings simply because of its signaling value.

Private and Social Rates of Return

The fact that education may both increase productivity as well as have a signaling value suggests that the **private rate of return to school,** as measured by the increase in a worker's earnings resulting from an additional year of schooling, may differ substantially from the **social rate of return to school,** as measured by the increase in national income resulting from that same year of education.

Suppose the signaling model is correct and education does not increase productivity. From a worker's point of view, education still has a positive private rate of return. The high-productivity worker gains from signaling that he is highly productive. From a social point of view, however, educational expenditures are wasteful. National income is not increased because the worker's productivity is the same both before and after the investment in education. The social rate of return is zero.

But this conclusion ignores the fact that even if education is only a signal, it still serves the useful role of sorting workers into the right jobs. The employer can use the signal to channel highly productive workers into "skilled" jobs and to channel the less-productive workers into other types of jobs. The mismatching of workers and jobs—for instance, assigning a low-productivity worker to manage software at a nuclear power plant—would surely have a detrimental effect on national income. As a result, education could have a positive social rate of return even if it does not increase a particular worker's human capital.

Summary

- A dollar received today does not have the same value as a dollar received tomorrow. The present value of a dollar received in the future gives the value of that amount in terms of today's dollars.
- The wage–schooling locus gives the salary that a worker earns if he or she completes a particular level of schooling.
- The marginal rate of return to school gives the percent increase in earnings associated with one more year of school.
- Workers choose the point on the wage–schooling locus that maximizes the present value of lifetime earnings. Specifically, workers quit school when the marginal rate of return to school equals the rate of discount.
- When workers differ only in their discount rates, the rate of return to school can be estimated by comparing the earnings of different workers. When workers differ in their innate abilities, the wage differential among workers does not measure the rate of return to school because the wage gap also depends on the unobserved ability differential.
- Workers sort themselves into those occupations for which they are best suited. This self-selection implies that we cannot test the hypothesis that workers choose the schooling level that maximizes the present value of lifetime earnings by comparing the earnings of workers in different education groups.
- Schooling can play a signaling role in the labor market, indicating to employers that the worker carrying the certificate or diploma is a highly productive worker. The signaling value of education can help firms differentiate highly productive workers from less productive workers.
- If education plays only a signaling role, workers with more schooling earn more not because education increases productivity, but because education signals a worker's innate ability.

Key Concepts

ability bias, *215*
age–earnings profile, *205*
asymmetric information, *227*
education production function, *220*
human capital, *201*
opportunity cost, *206*
pooling equilibrium, *228*
present value, *204*
private rate of return to schooling, *232*
rate of discount, *204*
rate of return to schooling, *209*
selection bias, *226*
separating equilibrium, *228*
signal, *228*
social rate of return to schooling, *232*
wage–schooling locus, *207*

Review Questions

1. Discuss how the present value of a future income payment is calculated.
2. Discuss how the wage–schooling locus is determined in the labor market, and why it is upward sloping and concave.
3. Derive the stopping rule for investments in education.

4. Why does the percentage gain in earnings observed when a worker gets one more year of schooling measure the marginal rate of return to education?
5. Discuss how differences in discount rates or in ability across workers lead to differences in earnings and schooling. Under what conditions can the rate of return to school be estimated?
6. Discuss the relationship between ability bias in the estimation of the rate of return to school and selection bias in tests of the hypothesis that workers choose the level of schooling that maximizes the present value of earnings.
7. Discuss how empirical studies estimate the rate of return to school and the methods used to avoid the problem of ability bias.
8. Show how education can signal the worker's innate ability in the labor market. What is a pooled equilibrium? What is a perfectly separating equilibrium?
9. How can we differentiate between the hypothesis that education increases productivity and the hypothesis that education is a signal for the worker's innate ability?

Problems

6-1. Debbie is about to choose a career path. She has narrowed her options to two alternatives. She can become either a marine biologist or a concert pianist. Debbie lives two periods. In the first, she gets an education. In the second, she works in the labor market. If Debbie becomes a marine biologist, she will spend $15,000 on education in the first period and earn $472,000 in the second period. If she becomes a concert pianist, she will spend $40,000 on education in the first period and then earn $500,000 in the second period. Suppose Debbie can lend and borrow money at a 5 percent rate of interest between the two periods. Which career will she pursue? What if she can lend and borrow money at a 15 percent rate of interest? Describe in general terms how Debbie's decision depends on the interest rate.

6-2. Peter lives for three periods. He is currently considering three alternative education-work options. He can start working immediately, earning $100,000 in period 1, $110,000 in period 2 (as his work experience leads to higher productivity), and $90,000 in period 3 (as his skills become obsolete and physical abilities deteriorate). Alternatively, he can spend $50,000 to attend college in period 1 and then earn $180,000 in periods 2 and 3. Finally, he can receive a doctorate degree in period 2 after completing his college education in period 1. This last option will cost him nothing when he is attending graduate school in the second period as his expenses on tuition and books will be covered by a research assistantship. After receiving his doctorate, he will become a professor in a business school and earn $400,000 in period 3. Peter's discount rate is 20 percent per period. What education path maximizes Peter's net present value of his lifetime earnings?

6-3. Jane has 3 years of college, Pam has 2, and Mary has 1. Jane earns $21 per hour, Pam earns $19, and Mary earns $16. The difference in educational attainment is completely due to different discount rates. How much can the available information reveal about each woman's discount rate?

6-4. Suppose the skills acquired in school depreciate over time, perhaps because technological change makes the things learned in school obsolete. What happens to a worker's optimal amount of schooling if the rate of depreciation increases?

6-5. (a) Describe the basic self-selection issue involved whenever discussing the returns to education.

(b) Does the fact that some high school or college dropouts go on to earn vast amounts of money (for example, Bill Gates dropped out of Harvard without ever graduating) contradict the self-selection story?

6-6. Suppose Carl's wage-schooling locus is given by

Years of Schooling	Earnings
9	$18,500
10	$20,350
11	$22,000
12	$23,100
13	$23,900
14	$24,000

Derive the marginal rate of return schedule. When will Carl quit school if his discount rate is 4 percent? What if the discount rate is 9 percent?

6-7. Suppose people with 15 years of schooling average earnings of $60,000 while people with 16 years of education average $66,000.

(a) What is the annual rate of return associated with the 16th year of education?

(b) It is typically thought that this type of calculation of the returns to schooling is biased, because it doesn't take into account innate ability or innate motivation. If this criticism is true, is the actual return to the 16th year of schooling more than or less than your answer in part (a)?

6-8. Suppose there are two types of people: high-ability and low-ability. A particular diploma costs a high-ability person $8,000 and costs a low-ability person $20,000. Firms wish to use education as a screening device where they intend to pay $25,000 to workers without a diploma and $$K$ to those with a diploma. In what range must K be to make this an effective screening device?

6-9. Some economists maintain that the returns to additional years of education are actually quite small but that there is a substantial "sheepskin" effect whereby one receives a higher salary with the successful completion of degrees or the earning of diplomas (that is, sheepskins).

(a) Explain how the sheepskin effect is analogous to a signaling model.

(b) Typically, in the United States, a high school diploma is earned after 12 years of schooling while a college degree is earned after 16 years of school. Graduate degrees are earned with between 2 and 6 years of post-college schooling. Redraw Figure 6-2 under the assumption that there are no returns to years of schooling but there are significant returns to receiving diplomas.

6-10. Consider a model with two periods—the first time period is the four years after high school and the second time period is the next 40 years. A person without a college education receives $120,000 of income during the first period and $1.2 million of income during the second period. A college graduate pays $200,000 during the first period to obtain a college degree and forgoes all earnings but then earns $2 million of income during the second period. Will the individual work or go to college in the first period if her individual rate of return between the two periods is 40 percent?

6-11. One policy objective of the federal government is to provide greater access to college education for those who are less able to afford it. Recently many state governments have passed budgets that have significantly reduced funding for state universities. Using supply and demand analysis, what is the likely effect on the price of a university education to potential students? What does your model predict in terms of the number of people who will complete a university education?

6-12. In 1970, men aged 18–25 were subject to the military draft to serve in the Vietnam War. A man could qualify for a student deferment, however, if he was enrolled in college and made satisfactory progress on obtaining a degree. By 1975, the draft was no longer in existence. The draft did not pertain to women. According to the 2008 edition of the *U.S. Statistical Abstract,* 55.2 percent of male high school graduates enrolled in college in 1970, but only 52.6 percent were enrolled in 1975. Similarly, 48.5 percent of female high school graduates were enrolled in college in 1970, while 49.0 percent were enrolled in 1975. Use women as the control group to estimate (using the difference-in-differences methodology) the effect abolishing the draft had on male college enrollment.

6-13. The textbook discusses in Section 6–5 some strategies for correcting for ability bias when trying to estimate the rate of return to education.

(a) What is the main argument for why using data on identical twins can control for ability bias? What problem arises if most pairs of identical twins pursue different levels of education? What problem arises if most pairs of identical twins pursue the same level of education?

(b) What is the main argument for why using certain birthdates can control for the bias? Do you think this method will be better as identifying the rate of return to different years of high school education or college education? Why?

6-14. A high school graduate has to decide between working and going to college. If he works, he will work for the next 50 years of his life. If he goes to college, he will be in college for 5 years, and then work for 45 years. In this model, the rate of discount that equates the lifetime present value of not going to college and going to college is 8.24 percent when the cost of each year of college is $15,000, each year of non-college work pays $35,000, and each year of post-college work pays $60,000. For each of the parts below, discuss how the rate of discount that equalizes the two options would change and who would make a different schooling decision based on the change. (Extra credit: Use Excel to show that the rate of return to schooling is 8.24 percent in the above case, and solve for the rates of discount associated with each of the parts below.)

(a) Each year of college still costs $15,000 and each year of post-college work still pays $60,000, but each year of non-college work now pays $40,000.

(b) Each year of college still costs $15,000 and each year of non-college work still pays $35,000, but each year of post-college work now pays $80,000.

(c) Each year of non-college work and post-college work still pays $35,000 and $60,000, respectively, but now each year of college costs $35,000.

(d) Each year of college still costs $15,000. The first year of noncollege work pays $35,000 but then increases by 3 percent each year thereafter. The first year of post-college work pays $60,000 but then increases by 5 percent each year thereafter.

6-15. Suppose the decision to acquire schooling depends on three factors–preferences (joy of learning), costs (monetary and psychic), and individual-specific returns to education.

(a) Explain how each of these factors affects one's optimal amount of schooling?

(b) Using these three factors, explain why someone who faces a very steep returns to education function may still opt to obtain very little schooling.

(c) Consider two groups of people—Alphas and Betas. The cost of schooling is the same for each. The average level of schooling and salary for Alpha types is 15 years and $120,000, while the average level of schooling and salary for Beta types is 13 years and $100,000. Why is it that 10 percent, which is calculated as ($120,000 − $100,000)/(15 − 13), is not a good estimate of the annual return to an additional year of education?

Selected Readings

Daniel Aaronson and Bhashkar Mazumder, "The Impact of Rosenwald Schools on Black Achievement," *Journal of Political Economy* 119 (October 2011): 821–888.

Joshua Angrist and Alan B. Krueger, "Does Compulsory Schooling Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106 (November 1991): 979–1014.

Orley C. Ashenfelter and Alan B. Krueger, "Estimates of the Economic Return to Schooling from a New Sample of Twins," *American Economic Review* 84 (December 1994): 1157–1173.

Raj Chetty, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane W. Schanzenbach, and Danny Yagan, "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," *Quarterly Journal of Economics* 126 (November 2011): 1593-1660.

Eric Maurin and Sandra McNally, "Vive la Revolution! Long-Term Educational Returns of 1968 to the Angry Students," *Journal of Labor Economics* 26 (January 2008): 1–33.

David Card and Alan B. Krueger, "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy* 100 (February 1992): 1–40.

Michael Spence, "Job Market Signaling," *Quarterly Journal of Economics* 87 (August 1973): 355–374.

Robert J. Willis and Sherwin Rosen, "Education and Self-Selection," *Journal of Political Economy* 87 (October 1979 Supplement): S7–S36.

Chapter 7

The Wage Distribution

What makes equality such a difficult business is that we only want it with our superiors.

—Henry Becque

The last chapter began our discussion of human capital by examining the worker's decision to acquire an education. Although the skills we learn in school make up an important fraction of our human capital stock, we do not stop accumulating knowledge the day we finally graduate. Instead, we continue to augment our human capital throughout our working lives.

Because of the skills acquired through informal on-the-job training or through formal vocational programs, college graduates in their fifties earn twice as much as college graduates in their twenties. This chapter examines how workers choose a particular path for their postschool investments and investigates how these choices influence the evolution of earnings over the life cycle.

The cumulative decisions we make about how much human capital to acquire, combined with the laws of supply and demand, determine the distribution of earnings in the labor market.[1] Inevitably, there will be some inequality in the allocation of rewards among workers; some workers will command higher earnings than others.

The observed wage distribution reflects two "fundamentals." First, workers differ in their productivity, and these differences arise partly because we acquire different amounts of human capital. The greater the productivity differences, the more unequal the wage distribution. Second, the rate of return to skills will vary across labor markets and over time, responding to changes in the supply and demand for skills. The greater the rate of return to skills, the greater the wage gap between skilled and unskilled workers, and the greater the inequality. This chapter examines the factors that determine the shape of the wage distribution, a distribution that typically exhibits a long tail at the top end. In other words, a few workers get a very large share of the rewards in the labor market.

The shape of the wage distribution changed in historic ways in recent decades, beginning in the 1980s. There was a sizable increase in inequality as the wage gap between high-skilled and low-skilled workers, as well as the wage dispersion within a particular skill group, rose rapidly. The fact that income inequality rose in the United States is

[1] For convenience, the discussion uses the terms *income distribution, earnings distribution,* and *wage distribution* interchangeably.

indisputable, but we have not yet reached a consensus on *why* this happened. It seems that no single culprit can explain the bulk of the increase. Instead, changes in labor market institutions and in economic conditions worked jointly to create a historic shift in how the labor market allocates rewards among workers.

The chapter concludes by showing how wage inequality can persist from generation to generation. Because parents care about the well-being of their children, many parents will make substantial investments in their children's human capital. These investments produce a positive correlation between the earnings of parents and the earnings of children, ensuring that part of the wage inequality observed in today's generation will be preserved into the next.

7-1 Postschool Human Capital Investments

The evolution of wages over the life cycle is described by the age–earnings profiles in Figure 7-1, which report the average weekly earnings of U.S. workers (in 2016 dollars) for various schooling groups at different ages. The figure reveals three important properties of age–earnings profiles:

1. *Highly educated workers earn more than less-educated workers.* We have seen that education increases earnings either because education increases productivity or because education serves as a signal of a worker's innate ability.
2. *Earnings rise over time, but at a decreasing rate.* The wage increase suggests that a worker becomes more productive as she accumulates labor market experience, perhaps because of on-the-job or off-the-job training programs. But the rate of wage growth slows down as workers get older. Younger workers seem to add more to their human capital than older workers.
3. *The age–earnings profiles of different education groups diverge over time.* Earnings increase fastest for the most educated workers. The steeper slope of age–earnings profiles for the most educated suggests a complementarity education and postschool investments. This complementarity might arise if some workers have a knack for acquiring all types of human capital.

7-2 On-the-Job Training

Most workers add to their human capital after they leave school, particularly through on-the-job training (OJT) programs. The diversity of OJT investments is striking: Programmers learn new languages, lawyers get courtroom experience, investment bankers concoct new financial instruments, and politicians learn from failed policies. Evidently, OJT is an important component of a worker's human capital stock, making up at least half of a worker's human capital.[2]

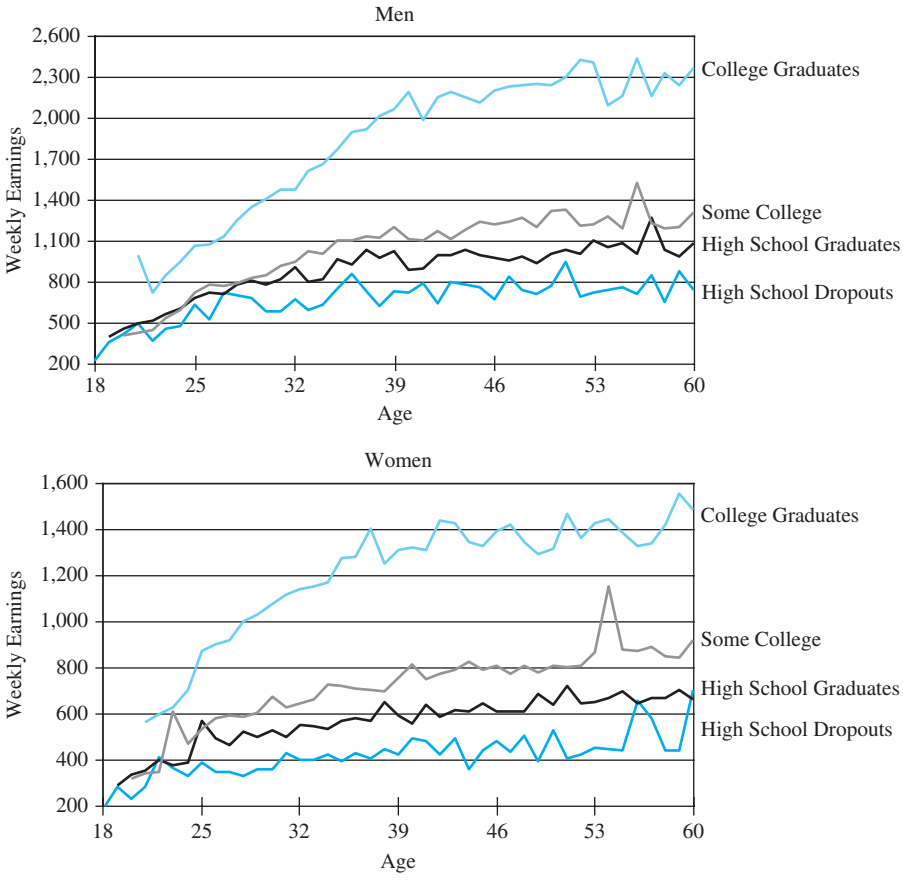
There are two types of OJT: **General training** and **specific training.**[3] General training enhances productivity equally in all firms. These general skills, which include typing,

[2] Jacob Mincer, "On-the-Job Training: Costs, Returns, and Some Implications," *Journal of Political Economy* 70 (October 1962, Part 2): 50–79.

[3] The concepts of general and specific training are due to Gary S. Becker, *Human Capital,* 3rd ed., Chicago: University of Chicago Press, 1993. Becker's framework is the cornerstone of the human capital model and an essential component in the toolkit of modern labor economics.

FIGURE 7-1 Age–Earnings Profiles, 2016

Source: U.S. Bureau of Labor Statistics, *Annual Social and Demographic Supplement of the Current Population Surveys,* pooled 2015–2017.



learning how to drive, or learning how to manipulate a spreadsheet, are found frequently in the labor market. Specific training enhances productivity only in the firm where it is acquired, so that skill is useless once the worker leaves the firm. Examples of specific training also abound in the labor market: Learning how to drive a tank in the army or memorizing the hierarchical nature of a particular organization.[4] In reality, much OJT is a mixture of general and specific training, but the conceptual separation into purely general and purely specific training is useful.

Consider a simple model where the employment relationship between a competitive firm and the worker lasts two periods. The *total* labor costs equal C_1 dollars in the first

[4] Some studies examine what it is about a worker's skills that is specific to certain jobs, occupations, or industries; see Maxim Poletaev and Chris Robinson, "Human Capital Specificity: Evidence from the Dictionary of Occupational Titles and Displaced Worker Surveys, 1984–2000," *Journal of Labor Economics* 26 (July 2008): 387–420; and Daniel Parent, "Industry-Specific Capital and the Wage Profile: Evidence from the National Longitudinal Survey of Youth and the Panel Study of Income Dynamics," *Journal of Labor Economics* 18 (April 2000): 306–323.

period and C_2 in the second. The values of marginal product are VMP_1 and VMP_2, respectively. Finally, let r be the rate of discount. The marginal productivity condition giving the profit-maximizing level of employment for the firm is

$$C_1 + \frac{C_2}{1+r} = VMP_1 + \frac{VMP_2}{1+r} \qquad \textbf{(7-1)}$$

The left-hand side of the equation gives the present value of the cost associated with hiring a worker over the two-period cycle. The right-hand side gives the present value of the worker's contribution to the firm. It is easy to grasp that this equation generalizes the condition that the wage equals the value of marginal product. In a multi-period framework, the analogous condition is that the present value of employment costs equals the present value of the value of marginal product.

Suppose OJT takes place only in the first period. It costs H dollars to train the worker, including the expense of instructor salaries and the purchase of training equipment. The total cost of hiring a worker during the first period is given by the sum of training costs H and the wage paid during the training period, or w_1. This implies that $C_1 = w_1 + H$. Because there is no training in the second period, the total cost of hiring C_2 then only equals the wage w_2. We can rewrite equation (7-1) as

$$H + w_1 + \frac{w_2}{1+r} = VMP_1 + \frac{VMP_2}{1+r} \qquad \textbf{(7-2)}$$

Who Pays for General Training?

Suppose all training is general. In the posttraining period, the worker's value of marginal product increases to VMP_2 in *all* firms. Put differently, the labor market is willing to pay the worker a wage equal to VMP_2. The firm that provided the training must either follow suit or lose the worker. Therefore, the second period wage, w_2, must equal VMP_2. Equation (6-22) simplifies to

$$w_1 = VMP_1 - H \qquad \textbf{(7-3)}$$

The first-period wage equals the value of the worker's initial marginal product minus training costs. In other words, workers pay for general training by accepting a lower "trainee wage" during the training period. In the second period, workers get the returns from the training by receiving a wage that equals the value of their posttraining marginal product. *Competitive firms provide general training only if they do not pay any of the costs.*

There are many examples of workers paying for general training through lower wages. Trainees in formal apprenticeship programs receive low wages during the training period and a higher wage after the training is completed. Similarly, medical residents (even though they already have a medical degree) earn low wages and work long hours during their residency, but their investment is well rewarded once they complete their training.

If a firm were to pay for general training, as some firms claim to do when they reimburse the tuition of employees who enroll in graduate programs, the firm would surely attract a large number of job applicants. Because the firm cannot enslave its workforce, however, the employees would take advantage of the free training opportunities and then run to a firm that offers them a wage commensurate with their newly acquired skills once the training was finished. A firm that paid for general training and did not raise the posttraining

wage would get an oversupply of trainees and the trained workers would then quit. The firm would quickly learn that it can lower the initial wage because there are "too many" trainees, allowing the firm to pass on the training costs to the workers.[5]

Who Pays for Specific Training?

The productivity gains resulting from specific training vanish when the worker leaves the firm. As a result, the worker's alternative wage (that is, the wage that other firms are willing to pay) is *independent* of the training and equals his pretraining productivity. Who then pays for specific training and who collects the returns?

Consider what would happen if the firm paid for specific training. The firm could incur the cost and collect the returns by not changing the wage in the posttraining period, even though the worker's value of marginal product *in this firm* has increased. Because VMP_2 would exceed w_2, there are gains to providing the training. If the worker were to quit in the second period, however, the firm would suffer a capital loss. The firm, therefore, would hesitate to pay for specific training unless it had some assurance that the trained worker will not quit.

Suppose instead that the worker pays for the specific training. Workers would receive a low wage in the training period and higher wages in the posttraining period. But the worker does not have an ironclad assurance that the firm will employ him in the second period. If the worker were to get laid off, he loses his investment because specific training is not portable. The worker, therefore, is not willing to invest in specific training unless he is very confident that he will not be laid off.

Both the firm and the worker, therefore, are reluctant to invest in specific training. The problem arises because there does not exist a legally binding contract that ties workers and firms together "until death do them part."

One way out of this dilemma is to fine-tune the posttraining wage in a way that reduces the possibility of *both* quits and layoffs. Consider an employment contract where the posttraining wage, w_2, is set such that

$$w^* < w_2 < VMP_2 \qquad \textbf{(7-4)}$$

where w^* is the alternative wage. This contract implies that the worker and the firm share the returns from specific training. The worker's posttraining wage w_2 is higher than his productivity elsewhere, but less than his productivity at the current firm. Because the worker is better off at this firm than elsewhere, he has no incentive to quit. And because the firm is better off by employing the worker than by laying him off (that is, the worker gets paid less than his value of marginal product), the firm does not want to let the worker go. By sharing the returns from specific training, both the worker and the firm commit to an employment contract that will not abruptly end in the posttraining period.

[5] Noncompetitive firms may pay for general training under some circumstances; see Daron Acemoglu and Jörn-Steffen Pischke, "The Structure of Wages and Investment in General Training," *Journal of Political Economy* 107 (June 1999): 539–572. Empirical studies of who pays for training are given by John M. Barron, Mark C. Berger, and Dan A. Black, "Do Workers Pay for On-the-Job Training?" *Journal of Human Resources* 34 (Spring 1999): 235–252; and David H. Autor, "Why Do Temporary Help Firms Provide Free General Skills Training?" *Quarterly Journal of Economics* 116 (November 2001): 1409–1448.

If firms and workers do share the returns of specific training, they will also have to share the cost. After all, if firms paid the entire cost of providing specific training and got only part of the returns, they would attract an oversupply of trainees. Therefore, if firms pay, say, 30 percent of the cost of specific training, they also will get 30 percent of the returns. Otherwise, the firm would attract either too few or too many job applicants.[6]

Implications of Specific Training

Specific training breaks the link between the worker's wage and the value of marginal product. During the training period, workers get paid less than their value of marginal product because they are paying part of the training costs. In the post training period, workers get paid less than their marginal product in the firm providing the training but get paid more than their marginal product in other firms.

Workers who have specific training are effectively granted a type of tenure or lifetime contract in the firm. Neither workers nor firms that have invested in specific training want to terminate the employment contract. It might seem surprising to argue that lifetime contracts might be common in labor markets where workers and firms are evidently very mobile, as in the United States. Nevertheless, the evidence indicates that jobs lasting more than 20 years have been the rule rather than the exception even in the United States.[7]

The concept of specific training provides an intuitive explanation of the "last hired, first fired" rule that determines who gets laid off during an economic downturn. Workers who have been with a firm for many years probably have more specific training than newly hired workers. When the demand for the firm's product falls, the price of the product and the worker's value of marginal product also fall. Workers with seniority have a buffer between their value of marginal product and their wage, protecting these senior workers from layoffs. Profit-maximizing firms that want to cut employment, therefore, will prefer to lay off newly hired workers.

Moreover, if a specifically trained worker does get laid off, that worker has little incentive to find alternative employment. The laid-off workers will suffer a capital loss if they change employers. They will prefer to "wait out" the unemployment spell until they are recalled by their former employers. There is, in fact, a very high incidence of **temporary layoffs.** At least 60 percent of the layoffs in the United States end when their former employers recall laid-off workers.[8]

Because specific training "marries" firms and workers, the probability of job separation for a given worker (either through a quit or through a layoff) declines with job seniority. Newly hired workers will have high turnover rates, and more senior workers will have low turnover rates. The negative correlation between job turnover and job seniority would not exist if all trainings were general. General training is portable and can be carried to any firm at any time. As a result, there would be no reason to expect the worker's economic opportunities in the current firm (relative to other firms) to improve over time.

[6] Masanori Hashimoto, "Firm-Specific Human Capital as a Shared Investment," *American Economic Review* 71 (June 1981): 475–482.

[7] Robert E. Hall, "The Importance of Lifetime Jobs in the U.S. Economy," *American Economic Review* 72 (September 1982): 716–724; and Manuelita Ureta, "The Importance of Lifetime Jobs in the U.S. Economy, Revisited," *American Economic Review* 82 (March 1992): 322–335

[8] Martin S. Feldstein, "Temporary Layoffs in the Theory of Unemployment," *Journal of Political Economy* 84 (October 1976): 937–957.

7-3 The Age–Earnings Profile

The shape of the age–earnings profile depends on the timing of human capital investments over the working life.[9] At every age, we will want to invest in human capital up to the point where the marginal revenue of the investment equals the marginal cost of the investment. To describe the timing of human capital acquisitions, we need to know what happens to the marginal revenue and the marginal cost of investments as a worker ages.

To simplify, let's assume that all training is general. Suppose we measure the human capital stock in **efficiency units.** Efficiency units are standardized units of skills. The total human capital stock of a worker equals the total number of efficiency units embodied in that worker. If David has 100 efficiency units and Mac has 50 units, David is equivalent to two Macs—at least in terms of his labor market productivity.

An efficiency unit of human capital can be rented out in the labor market, and the rental rate per efficiency unit is R dollars. The market for efficiency units is competitive. From the perspective of the worker, she can rent out as many efficiency units as she acquires at the same rental price R. Finally, to keep things simple, let's assume there is no depreciation of the human capital stock over time. Therefore, an efficiency unit of human capital generates R dollars per year from the date it is acquired until retirement.

Suppose the worker enters the labor market at age 20 and retires at 65. The marginal revenue of acquiring one efficiency unit of human capital at age 20 is

$$MR_{20} = R + \frac{R}{1+r} + \frac{R}{(1+r)^2} + \frac{R}{(1+r)^3} + \cdots + \frac{R}{(1+r)^{45}} \qquad \textbf{(7-5)}$$

where r is the discount rate. The intuition behind equation (7-5) is easy to understand. If a worker acquires one efficiency unit at age 20, this investment yields a return of R dollars during that first year of work experience. In the second year, the present value of the return to that same efficiency unit is $R/(1 + r)$ dollars; in the third year, the return equals $R/(1 + r)^2$ dollars; and so on. Equation (7-5) simply adds the discounted returns to the efficiency unit over the entire working life.

The curve MR_{20} in Figure 7-2 illustrates the relationship between the marginal revenue of an efficiency unit acquired at age 20 and the number of efficiency units that the worker acquires. Because we assumed that the rental rate R is the same regardless of how much human capital the worker acquires, the marginal revenue curve MR_{20} is horizontal.

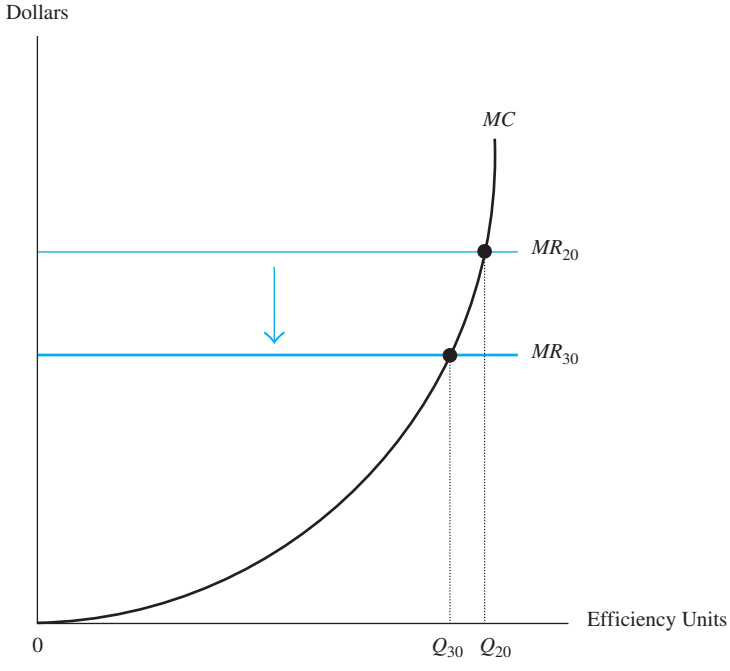
Now suppose that the worker has just turned 30 years old. The marginal revenue of an efficiency unit acquired at age 30 is

$$MR_{30} = R + \frac{R}{1+r} + \frac{R}{(1+r)^2} + \frac{R}{(1+r)^3} + \cdots + \frac{R}{(1+r)^{35}} \qquad \textbf{(7-6)}$$

Equation (7-6) indicates that the marginal revenue of acquiring an efficiency unit at age 30 is the discounted sum of the returns collected at age 30, at age 31, and so on. Because the

[9] Yoram Ben-Porath, "The Production of Human Capital and the Life Cycle of Earnings," *Journal of Political Economy* 75 (August 1967): 352–365; and James J. Heckman, "A Life-Cycle Model of Earnings, Learning, and Consumption," *Journal of Political Economy* 84 (August 1976 Supplement): S11–S46.

FIGURE 7-2 The Acquisition of Human Capital over the Life Cycle
The marginal revenue of an efficiency unit of human capital declines as the worker ages (so that MR_{20}, the marginal revenue of a unit acquired at age 20, lies above MR_{30}). At each age, the worker equates the marginal revenue with the marginal cost, so that more units are acquired when the worker is younger.



worker is now 10 years closer to retirement, the sum in equation (7-6) has 10 fewer terms than the sum in equation (7-5).

By comparing the marginal revenue of acquiring an efficiency unit at ages 20 and 30, we can see that $MR_{20} > MR_{30}$. Figure 7-2 illustrates that the MR_{20} curve lies above the MR_{30} curve. The marginal revenue of human capital investment falls as the worker ages for a simple reason: We do not live forever. Human capital acquired when young can be rented out for a long time, but investments undertaken at older ages will be rented out for shorter periods. The lesson is obvious: Human capital investments are more profitable the earlier they are undertaken.

The number of efficiency units acquired at any age is determined by equating the marginal revenue with the marginal cost of human capital investments. The marginal cost (*MC*) curve, also illustrated in Figure 7-2, has the usual shape: Marginal cost rises as more efficiency units are acquired. The shape of the marginal cost curve is implied by the production function for human capital. The law of diminishing returns guarantees that marginal costs increase at an increasing rate as the worker attempts to acquire more and more human capital.

The intersection of the marginal revenue and marginal cost curves implies that the worker acquires Q_{20} efficiency units at age 20 and Q_{30} units at age 30. The theory implies that the worker acquires fewer efficiency units as he gets older. This prediction helps us

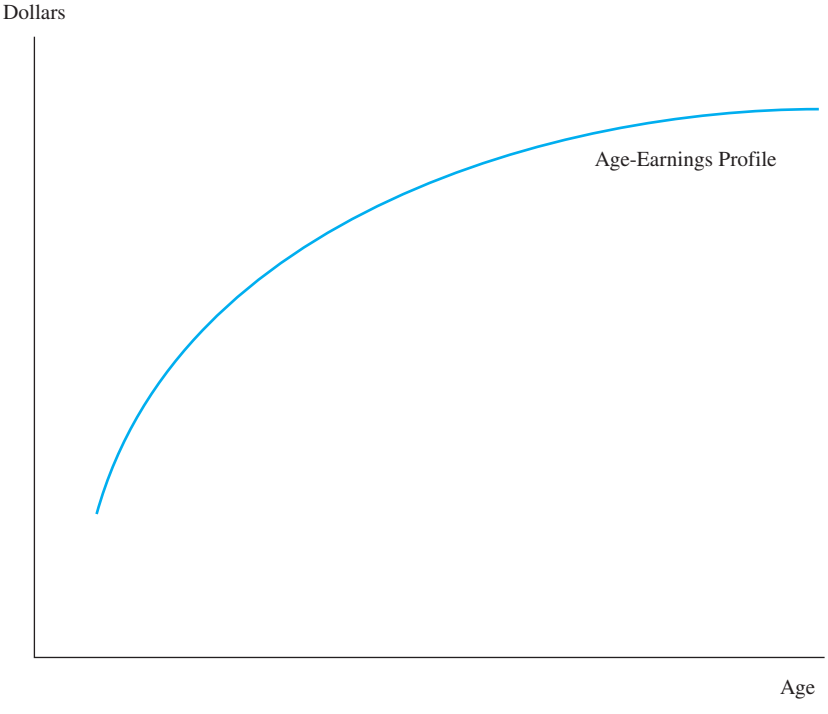
understand why workers typically go to school when young, why this period of specialization is followed by on-the-job training, and why on-the-job training activities taper off as the worker ages. This is the timing of investments that maximizes the present value of lifetime earnings.[10]

Because the worker acquires more human capital when he is young, the worker's age–earnings profile is upward sloping, as illustrated in Figure 7-3. As we have seen, workers pay for on-the-job training through reduced wages. Older workers, therefore, earn more than younger workers because older workers invest less in human capital and are collecting the returns to prior investments.

The optimal timing of investments over the working life also implies that the age–earnings profile is concave so that earnings increase over time but at a decreasing rate. Year-to-year wage

FIGURE 7-3 The Age–Earnings Profile Implied by Human Capital Theory

The age–earnings profile is upward-sloping and concave. Older workers earn more because they invest less in human capital and because they are collecting the returns from earlier investments. The rate of growth of earnings slows down over time because workers accumulate less human capital as they get older.



[10] The discussion assumes that the marginal cost curve is constant over time (that is, it does not shift as the worker ages). It may be that older workers are more efficient at producing human capital, shifting the marginal cost curve down. At the same time, forgone earnings are higher for older workers, and this would shift the marginal cost curve up as the worker ages. The two opposing effects are sometimes assumed to exactly offset each other (the "neutrality assumption"), so that the marginal cost curve is constant. See Yoram Ben-Porath, "The Production of Human Capital over Time," in W. Lee Hansen, editor, *Education, Income, and Human Capital,* New York: Columbia University Press, 1970.

Theory at Work

EARNINGS AND SUBSTANCE ABUSE

The typical discussion of the economic impact of human capital focuses on the benefits that the worker's choices, such as schooling or on-the-job training, impart on productivity and earnings. Many workers, however, also make decisions that have an adverse impact on the value of their human capital stock, such as alcoholism and drug use.

About 5 percent of the U.S. population suffer from alcoholism at any point in time, and nearly 10 percent of the population does so at some point in their lives. There is strong evidence that alcoholics pay a heavy price not only in terms of their health and the well-being of their families but also in the labor market. Among prime-aged workers, alcoholics are 15 percentage points less likely to work and earn 17 percent less than nonalcoholics. And these large gaps persist even if we only look at alcoholics whose health has not yet been impaired.

Drug use is an equally important challenge. By the time workers reach the age of 30, nearly 30 percent have used cocaine and about 3 percent have used it in the past month. Surprisingly, the evidence does not suggest that cocaine users have systematically lower employment rates or wages.

It is important to emphasize that these correlations between substance abuse and labor market characteristics do not necessarily prove that "alcoholism lowers wages" or that "cocaine use does not reduce productivity." The population of substance abusers is self-selected. Perhaps alcoholism does not reduce earnings, but those workers who are less successful in the labor market have a greater chance of becoming alcoholics. Similarly, it may be that only those workers who can handle the adverse consequences of cocaine use, or who can afford to buy cocaine, become habitual users.

Sources: Thomas S. Dee and William N. Evans, "Teen Drinking and Educational Attainment: Evidence from Two-Sample Instrumental Variables Estimates," *Journal of Labor Economics* 21 (January 2003): 178–209; John Muhally and Jody L. Sindelar, "Alcoholism, Work, and Income," *Journal of Labor Economics* 11 (July 1993): 494–520; and Robert Kaestner, "New Estimates of the Effect of Marijuana and Cocaine Use on Wages," *Industrial and Labor Relations Review* 47 (April 1994): 454–470.

growth depends partly on how many additional efficiency units the worker acquires. Because fewer units are acquired as the worker ages, the rate of wage growth declines over time.

The Mincer Earnings Function

The implications of the human capital model for the age–earnings profile have been the subject of extensive analysis. This line of research culminated in the development of the **Mincer earnings function**[11]

$$\log w = as + bt - ct^2 + \text{Other variables} \qquad \textbf{(7-7)}$$

where w is the worker's wage rate, s is the number of years of schooling, t gives the number of years of labor market experience, and t^2 is a quadratic on experience that captures the concavity of the age–earnings profile.

[11] Jacob Mincer, *Schooling, Experience, and Earnings,* New York: Columbia University Press, 1974. This literature is surveyed by Robert J. Willis, "Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Functions," in Orley C. Ashenfelter and Richard Layard, editors, *Handbook of Labor Economics,* vol. 1, Amsterdam: Elsevier, 1986, pp. 525–602; and James J. Heckman, Lance J. Lochner, and Petra E. Todd, "Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond," in Eric Hanushek and Finis Welch, editors, *Handbook of Education Economics,* Amsterdam: Elsevier, 2006.

The coefficients of the Mincer earnings function provide information about human capital investments. For example, the coefficient on years of schooling a estimates the percent increase in earnings resulting from one more year of schooling and is typically interpreted as the rate of return to school. We have seen that this interpretation is correct only when workers do not differ in their unobserved ability. The coefficients on experience and experience squared estimate the rate of growth in earnings resulting from one more year of labor market experience and provide information about the volume of OJT investments. If the worker did not invest in OJT, these coefficients would be zero because there would be no reason for real earnings to change as a worker ages.

Hundreds of studies conclude that the Mincer earnings function provides a reasonably accurate summary of age–earnings profiles in the United States and in many other countries (even in countries with very different labor market institutions). Real-world age–earnings profiles are indeed concave and are higher for more-educated workers. The studies also conclude that the variation in education and labor market experience across workers accounts for about a third of the variation in wage rates. The human capital framework, therefore, is a solid first step in understanding the determinants of the earnings distribution.[12]

7-4 Policy Application: Training Programs

Perhaps the most important policy implication of the human capital model is that the provision of training to low-skill workers may substantially improve their economic well-being. Since the declaration of the War on Poverty in the mid-1960s, a large number of government programs have subsidized training for disadvantaged workers, with federal expenditures on these programs now exceeding $4 billion per year. In view of the large cost, it is not surprising that many studies attempt to determine if the programs do what they are supposed to do—namely, increase the human capital and earnings of the trainees.[13]

The program evaluations raise a number of difficult conceptual issues. It would seem that by comparing the earnings of trainees before and after the "treatment," we could measure the effectiveness of the program. Studies that make this type of before-and-after comparison report that there are some earnings gains. Typically, trainees earn about $300 to $1,500 more per year after the program than before the program.[14]

But this statistic may not be very useful. As in many other contexts in labor economics, the problem of self-selection mars the analysis. Only those workers who have the most to gain from the program and who are most committed to "self-improvement" are likely to

[12] Conflicting evidence on the importance of OJT as a determinant of earnings growth in the post-school period is given by Burkhanettin Burusku, "Training and Lifetime Income," *American Economic Review* 96 (June 2006): 832–846

[13] James J. Heckman, Robert J. LaLonde, and Jeffrey A. Smith, "The Economics and Econometrics of Active Labor Market Programs," in Orley C. Ashenfelter and David Card, editors, *Handbook of Labor Economics,* vol. 3A, Amsterdam: Elsevier, 1999, pp. 1865–2097.

[14] Orley C. Ashenfelter and David Card, "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics* 67 (November 1985): 648–660; and Burt Barnow, "The Impact of CETA Programs on Earnings: A Review of the Literature," *Journal of Human Resources* 22 (Spring 1987): 157–193.

enroll and subject themselves to the treatment. The earnings gain achieved by this nonrandom sample of workers, therefore, tells us something about how the programs affect motivated workers but may say nothing about how the program would affect a randomly chosen low-income worker.

Social Experiments

To avoid this pitfall, there has been a revolutionary shift in the methodology used in program evaluation. Recent evaluations use randomized experiments, akin to the experimental methods used in the physical sciences, to estimate the impact of the program on trainee earnings. In these experiments, potential trainees are randomly assigned to participate in the program. Every other applicant, for instance, may be allocated to the "treatment" group (that is, they are enrolled in the training program), while the remaining applicants form the control group and are administered a placebo (that is, they are not provided the training).

The National Supported Work Demonstration (NSW) provides a good example of such an experiment.[15] The objective of the NSW was to ease the transition of disadvantaged workers into the labor market by exposing them to a work environment where experience and counseling would be provided. Eligible applicants were randomly assigned to one of the two tracks. The lucky "treated" workers received all the benefits provided by the NSW, while those in the control group received none of the benefits and were left on their own. The NSW guaranteed the workers in the treatment group a job for 9–18 months, at which time they had to find regular employment. The program costs about $12,500 per participant (in 1998 dollars).

Table 7-1 summarizes the evidence from an influential evaluation of the program. The typical worker who was treated by the program earned $1,512 annually in the pretraining period and $7,888 after the training. The typical trainee, therefore, experienced a wage gain of almost $6,400.

TABLE 7-1 The Impact of the NSW Program on the Earnings of Trainees (in 1998 Dollars)

Source: Robert J. LaLonde, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76 (September 1986): 604–620, Table 2.

Group	Pretraining Annual Earnings (1975)	Posttraining Annual Earnings (1979)	Difference
Treatment group	1,512	7,888	6,376
Control group	1,481	6,450	4,969
Difference-in-differences	—	—	**1,407**

[15] Robert J. LaLonde, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76 (September 1986): 604–620. Other studies of experimental data include Stephen H. Bell and Larry L. Orr, "Is Subsidized Employment Cost Effective for Welfare Recipients? Experimental Evidence from Seven State Demonstrations," *Journal of Human Resources* 29 (Winter 1994): 42–61; and Alberto Abadie, Joshua D. Angrist, and Guido W. Imbens, "Instrumental Variables Estimates of the Effects of Training on the Quantiles of Trainee Earnings," *Econometrica* 70 (January 2002): 91–117. International evidence is given by Laura Abramovsky, Erich Battistin, Emla Fitzsimons, Alissa Goodman, and Helen Simpson, "Providing Employers with Incentives to Train Low-Skilled Workers: Evidence from the UK Employer Training Pilots," *Journal of Labor Economics* 29 (January 2011): 153–193.

This wage gain, however, does not estimate the impact of the training program because the earnings of trainees would have changed between 1975 and 1979 for other reasons, such as aging and changes in aggregate economic conditions. To isolate the impact of the program, we must net out the influence of these extraneous events on earnings. Workers in the control group earned $1,481 in 1975 and $6,450 in 1979, a gain of almost $5,000. Since earnings would have increased by $5,000 regardless of whether the worker was injected with the training, the true impact of the program is the difference-in-differences, or about $1,400. As noted above, the NSW program costs about $12,500 per participant, implying a rate of return of about 10 percent.

Although the experimental approach has become the standard way of evaluating training programs, the methodology has its detractors.[16] For example, the $1,400 increase in earnings may not be the net gain that would be observed if the programs were made available to the entire low-income population because the treatment and control groups may not define a true experiment. Only persons who were interested in receiving the training bothered to go to the training center and fill out an application. As a result, there is already self-selection in the sample of persons who end up in the treatment group. Moreover, some of the workers allocated to the treatment group may not show up for the training, while workers allocated to the control group may find a way of qualifying for some other program (perhaps by trying out other sites).

7-5 Wage Inequality

We have examined how workers acquire the human capital stock that maximizes the present value of earnings, both in school and through on-the-job training. These decisions inevitably create a lot of wage dispersion across workers.

Figure 7-4 illustrates the distribution of full-time weekly earnings for working men in the United States in 2017. The mean weekly wage was $1,062 and the median was $865. The wage distribution exhibits two important properties. First, there is a lot of inequality. Second, the distribution is not symmetric, with similar tails on both sides. Instead, it is positively skewed, with a long right tail. A **positively skewed income distribution** implies that the bulk of workers earn relatively low wages and that a small number of workers in the upper tail receive a disproportionately large share of the rewards.

As Table 7-2 shows, there are sizable differences in the shape of the income distribution across countries. The top 10 percent of U.S. households get 30 percent of total income. The respective statistic for Belgium is 22 percent; for Germany, 25 percent; and for Guatemala, 42 percent. Similarly, the bottom 10 percent receive only 2 percent of income in the United States. The poorest households receive 4 percent of income in Norway, but only 1 percent in Guatemala.

Modern studies of the shape of the wage distribution use the human capital model as a point of departure. This approach has proved popular because it helps us understand some

[16] James J. Heckman and Jeffrey A. Smith, "Assessing the Case for Social Experiments," *Journal of Economic Perspectives* 9 (Spring 1995): 85–110. An overview of experimental methods in labor economics is given by John A. List and Imran Rasul, "Field Experiments in Labor Economics," in Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics,* Vol. 4a, Amsterdam: Elsevier, 2011, pp. 103–228.

FIGURE 7-4 The Wage Distribution in the United States, 2017

Source: U.S. Bureau of Labor Statistics, *Current Population Survey, Outgoing Rotation Group, 2017.*

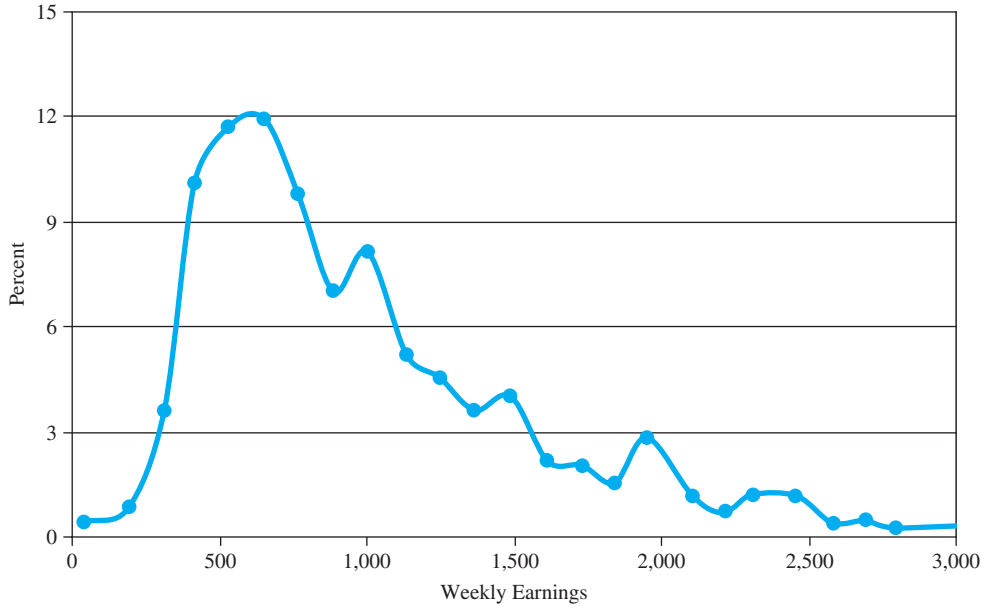


TABLE 7-2 International Differences in the Income Distribution

Source: World Bank, World Development Indicators, 2017. The statistics report the shape of the income distribution as of 2013 for most countries.

Country	Percentage of Total Income Received by Bottom 10% of Households (%)	Percentage of Total Income Received by Top 10% of Households (%)
Australia	3	27
Austria	3	24
Belgium	3	22
Canada	2	25
Chile	2	36
Dominican Republic	2	37
France	3	26
Germany	3	25
Guatemala	1	42
Hungary	3	24
India	4	30
Israel	2	30
Italy	2	26
Mexico	2	39
Norway	4	21
Sweden	3	22
United Kingdom	3	25
United States	2	30

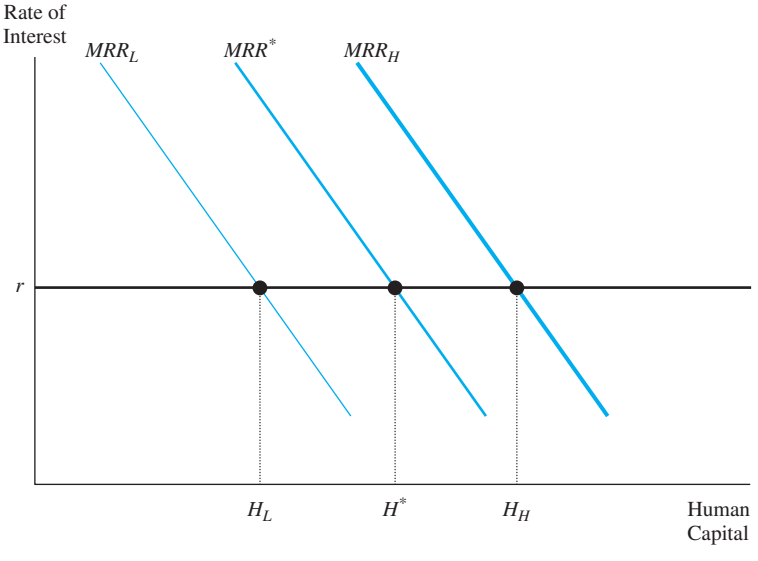
of the key characteristics of real-world wage distributions. In the human capital framework, wage differentials arise not only because some workers accumulate more human capital than others, but also because young workers are accumulating skills (and forgoing earnings) while older workers are collecting the returns from prior investments.

The human capital model implies that the wage distribution will be positively skewed. A worker invests in human capital up to the point where the marginal rate of return to the investment equals the rate of discount. This stopping rule generates a positively skewed wage distribution *even if the distribution of ability in the population is symmetric.* To illustrate, suppose that a third of the workforce is composed of low-ability workers, a third of medium-ability workers, and the remaining third of high-ability workers. Further, suppose all workers have the same rate of discount.

Figure 7-5 shows the investment decision for each of the groups. The curve MRR_L gives the marginal rate of return curve for low-ability workers. This group will acquire H_L efficiency units of human capital. Similarly, MRR^* gives the curve for medium-ability workers, who acquire H^* units; and MRR_H gives the curve for high-ability workers, who acquire H_H units. High-ability workers then earn more than low-ability workers for two distinct reasons. First, high-ability workers would earn more even if all groups acquired the same amount of human capital. Ability increases productivity and earnings. But high-ability workers also earn more because they acquire more human capital. Put differently, the positive correlation between ability and human capital investments "stretches out" wages in the upper tail, creating positive skewness.

FIGURE 7-5 Ability Differences Create Positive Skewness

Low-ability workers face the marginal rate of return curve MRR_L and acquire H_L units of human capital. High-ability workers face the MRR_H curve and acquire H_H units of human capital. High-ability workers earn more than low-ability workers both because they have more ability and because they acquire more human capital. The positive correlation between ability and acquired human capital "stretches out" the wage distribution at the top, creating positive skewness.



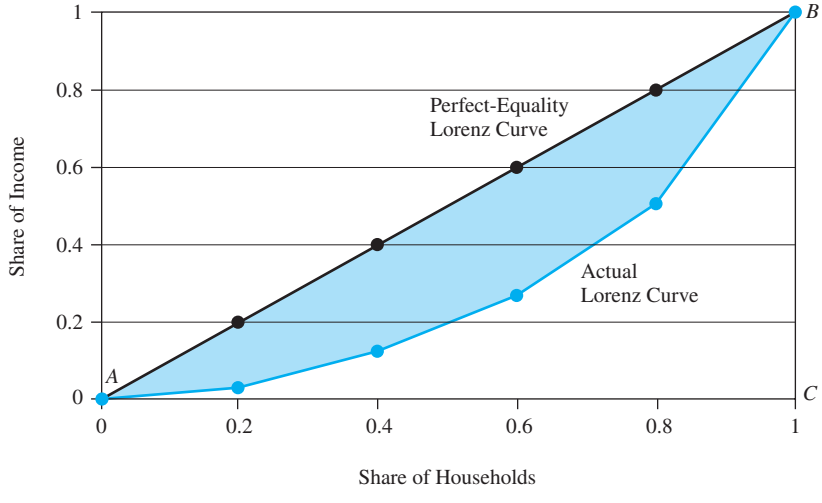
7-6 Measuring Inequality

Measures of income inequality are typically based on calculations of how much income goes to particular segments of the population.[17] Suppose we rank all households according to their income, from lowest to highest. Let's now break the population into five groups of equal size. The first quintile contains the 20 percent of the households with the lowest incomes and the fifth quintile contains the 20 percent with the highest incomes. Let's calculate how much income accrues to households in each quintile. If every household earned the same income, so that there were perfect income equality, it would be the case that 20 percent of the income goes to the first quintile, 20 percent goes to the second quintile, 20 percent goes to the third quintile, and so on.

We can summarize the data graphically by relating the *cumulative* share of income accruing to the various groups. In the case of perfect equality, the result would be the line *AB* in Figure 7-6. This line indicates that 20 percent of total income accrues to the bottom 20 percent of the households; 40 percent of the income accrues to the bottom 40 percent; and 60 percent of the income accrues to the bottom 60 percent. The line *AB* is called a **Lorenz curve.** The Lorenz curve reports the cumulative share of income accruing to the various quintiles. The "perfect-equality" Lorenz curve must be a straight line with a 45° angle.

FIGURE 7-6 The Lorenz Curve and the Gini Coefficient

The straight line AB gives the "perfect-equality" Lorenz curve, indicating that each quintile of households gets 20 percent of aggregate income. The Lorenz curve describing the actual income distribution lies below it. The ratio of the shaded area to the area in the triangle *ABC* gives the Gini coefficient.



[17] Daniel J. Slottje, *The Structure of Earnings and the Measurement of Income Inequality in the U.S.* Amsterdam: Elsevier, 1989.

TABLE 7-3 Household Shares of Aggregate Income, by Quintiles of the Income Distribution, 2010

Source: U.S. Bureau of the Census, Income, Poverty, and Health Insurance Coverage in the United States: 2010, Table 3; http://www.census.gov/prod/2010pubs/p60-238.pdf.

Quintile	Share of Income	Cumulative Share of Income
First	0.034	0.034
Second	0.086	0.120
Third	0.147	0.267
Fourth	0.233	0.500
Fifth	0.500	1.000

Table 7-3 reports the actual distribution of household income in the United States as of 2010. The bottom 20 percent of the households received 3.4 percent of all income and the next quintile received 8.6 percent. The cumulative share received by the bottom two quintiles must then be 12.0 percent. Obviously, the cumulative share received by all quintiles must equal 1.0.

Figure 7-6 also illustrates the Lorenz curve derived from the actual income distribution. The real-world Lorenz curve lies below the perfect-equality curve. In fact, the construction of the Lorenz curve suggests that the more inequality in an income distribution, the further away the actual Lorenz curve will be from the 45° line. Consider a world in which all income accrues to the fifth quintile, the top fifth of the households. In this world of "perfect inequality," the Lorenz curve would look like a mirror image of the letter L; it would be flat along the horizontal axis, so that 0 percent of the income accrues to 80 percent of the households, and then shoots up so that 100 percent of the income accrues to 100 percent of the households.

The intuition behind the construction of the Lorenz curve suggests that the shaded area between the perfect-equality Lorenz curve and the real-world Lorenz curve can be used to measure inequality. In fact, the **Gini coefficient** is defined by

$$\text{Gini coefficient} = \frac{\textit{shaded area}}{\textit{Area of triangle } ABC} \qquad (7\text{-}8)$$

In terms of Figure 7-6, the Gini coefficient is given by the ratio of the shaded area to the area of the triangle *ABC*.[18] The Gini coefficient would then be 0 when the real-world distribution of income exhibits perfect equality and would equal 1 when the real-world distribution exhibits perfect inequality (that is, when all income goes to the highest quintile). By calculating the areas of various triangles and rectangles in Figure 7-6 and then applying equation (7-8), it is easy to show that the Gini coefficient in 2010 was 0.43.

Although an increase in the Gini coefficient represents an increase in inequality, some subtleties are overlooked when the entire shape of the income distribution is summarized in a single number. Consider, for example, a shift in income from the bottom quintile to the top quintile. This shift obviously increases the Gini coefficient. But we can obtain a similar numerical increase in the Gini coefficient by transferring some income from, say, the second and third quintiles to the top quintile. Although the numerical increase in the Gini coefficient is the same, the two redistributions are not identical.

[18] Note that the area of the triangle *ABC* must equal 0.5.

Because of this ambiguity, we often look at additional measures of inequality. Two commonly used measures are the **90–10 wage gap** and the **50–10 wage gap.** The 90–10 wage gap gives the percent wage differential between the worker at the 90th percentile of the income distribution and the worker at the 10th percentile. The 90–10 wage gap gives a sense of the range of the income distribution. The 50–10 wage gap gives the percent wage differential between the worker at the 50th percentile and the worker at the 10th percentile. The 50–10 wage gap gives a sense of the inequality between the "middle class" and low-income workers.

7-7 The Changing Wage Distribution

Many studies show that the U.S. wage distribution changed dramatically in the 1980s and 1990s.[19] One conclusion is inescapable: Wage inequality increased substantially. In particular:

- The wage gap between those at the top of the distribution and those at the bottom widened.
- Wage differentials *across* education groups and *across* age groups widened.
- Wage differentials *within* demographic and skill groups widened. The earnings of workers of the same education, age, sex, occupation, and industry were more dispersed in the mid-1990s than they were in the late 1970s.

This section briefly documents some of these changes in the wage distribution. Figure 7-7*a* shows the trend in the Gini coefficient. The coefficient declined steadily from the 1930s through 1950. It was then relatively stable until about 1970, when it began a dramatic rise.

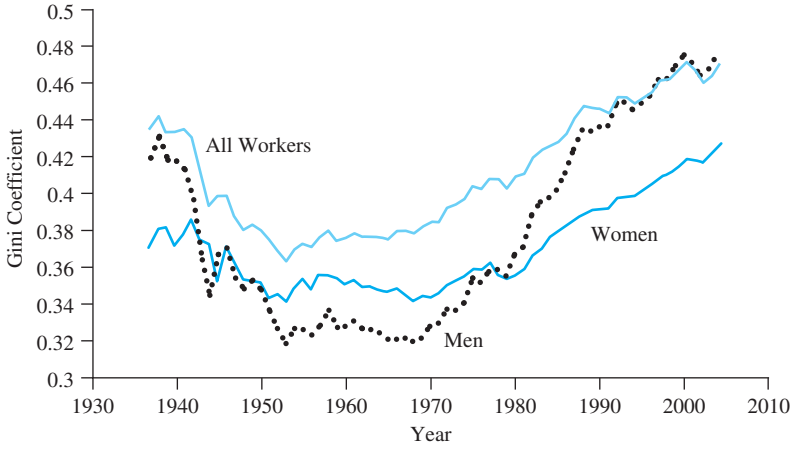
Figure 7-7*b* shows that some of the increase in wage inequality can be directly attributed to a sizable increase in the returns to schooling. The figure illustrates the trend in the percent wage differential between college graduates and high school graduates. This wage gap rose slightly throughout the 1960s until about 1971. It then began to decline until about 1979, when it made "a great U-turn" and began a very rapid rise. In 1979, college graduates earned 47 percent more than high school graduates. By 2001, college graduates earned 90 percent more. If we interpret the wage gap across education groups as a general measure of the rate of return to skills, it seems that structural changes in the U.S. economy may have led to a historic increase in the rewards to skills.

There is also strong evidence that wage inequality increased not only across schooling groups, but also *within* narrowly defined skill groups. Figure 7-7*c* shows the trend in the 90–10 wage gap within a group of workers who have the same age, education, gender,

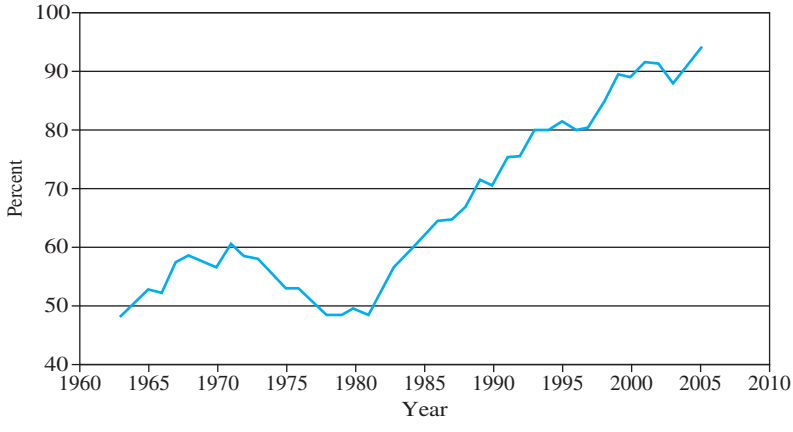
[19] The key studies include Lawrence F. Katz and Kevin M. Murphy, "Changes in Relative Wages, 1963–1987: Supply and Demand Factors," *Quarterly Journal of Economics* 107 (February 1992): 35–78; Kevin M. Murphy and Finis Welch, "The Structure of Wages," *Quarterly Journal of Economics* 107 (February 1992): 285–326; and Chinhui Juhn, Kevin M. Murphy, and Brooks Pierce, "Wage Inequality and the Rise in Returns to Skills," *Journal of Political Economy* 101 (June 1993): 410–442. The literature is surveyed by Lawrence F. Katz and David H. Autor, "Changes in Wage Structure and Earnings Inequality," in Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics,* vol. 3A, Amsterdam: Elsevier, 1999, pp. 1463–1555.

FIGURE 7-7
Earnings Inequality in the United States

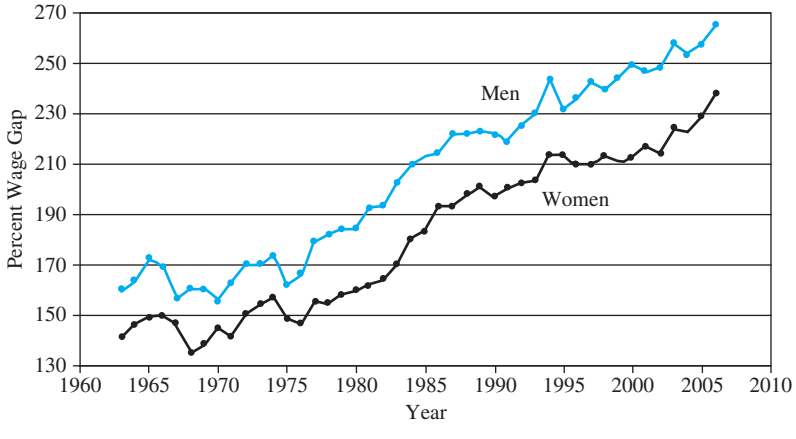
Sources: (*a*) Wojciech Kopczuk, Emmanuel Saez, and Jae Song, "Earnings Inequality and Mobility in the United States from Social Security Data Since 1937," *Quarterly Journal of Economics* 125 (February 2010): 91–128; (*b*) and (*c*) David H. Autor, Lawrence F. Katz, and Melissa S. Kearney, "Trends in U.S. Wage Inequality: Revising the Revisionists," *Review of Economics and Statistics* 90 (May 2008): 300–323.



(*a*) Gini Coefficient



(*b*) Wage Differential between College Graduates and High School Graduates



(*c*) The "Residual" 90-10 Wage Gap

and race. This measure of "residual" wage inequality shows a striking upward trend since the 1970s.[20]

In short, wage inequality began to rise rapidly in the U.S. labor market sometime in the mid-1970s, both within and across skill groups. This fact ranks among the most important economic events of the last half of the twentieth century.

7-8 Policy Application: Why Did Inequality Increase?

Although the increase in wage inequality is well documented, there is still disagreement over *why* this increase took place. Many researchers have attempted to find the smoking gun that would explain the historic change in the wage distribution. The search, however, has not been entirely successful. No single factor seems to be able to explain all—or even most—of the increase. Instead, the increase in inequality seems to have been caused by concurrent changes in economic "fundamentals" and labor market institutions.

Most studies that attempt to explain the rise in inequality use a simple framework that shows how shifts in labor supply and labor demand could have caused such a sizable increase.[21] To simplify, suppose there are only two types of workers in the labor market, skilled and unskilled. Suppose further that the production technology used by the firm can be described by the **Constant Elasticity of Substitution (CES)** production function. Ignoring capital, the CES production function specifies how much output is produced as a function of the number of skilled workers L_S and the number of unskilled workers L_U, and is given by

$$Q = \left[\alpha L_S^{\delta} + (1-\alpha) L_U^{\delta}\right]^{1/\delta} \qquad \textbf{(7-9)}$$

Despite the obvious mathematical complexity of the CES production function, it turns out to be a very useful framework for examining the factors that determine the wage gap between skilled and unskilled workers. By using some calculus, it can be shown that the marginal productivity conditions implied by the CES production function can be rewritten as[22]

$$\log\left(\frac{w_S}{w_U}\right) = b_0 - b_1 \log\left(\frac{L_S}{L_U}\right) \qquad \textbf{(7-10)}$$

Equation (7-10) relates the *relative* wage of skilled workers (or log w_S/w_U) to their relative number (or log L_S/L_U). It gives the **relative demand curve** for skilled workers. The CES production function predicts that the relative demand curve is downward sloping and linear. The greater the relative number of skilled workers, the lower their relative wage.

[20] Wage inequality also increased even within narrowly defined occupation and industry groups; see Pedro Carneiro and Sokbae Lee, "Trends in Quality-Adjusted Skill Premia in the United States, 1960–2000." *American Economic Review* 101 (October 2011): 2309–2349.

[21] See Katz and Murphy, "Changes in Relative Wages, 1963–1987: Supply and Demand Factors"; Murphy and Welch, "The Structure of Wages"; and David Card and Thomas Lemieux, "Can Falling Supply Explain the Rising Return to College for Younger Men," *Quarterly Journal of Economics* 116 (May 2001): 705–746.

[22] The marginal productivity condition for skilled workers is $w_S = \partial Q/\partial L_S = \alpha\delta Q^{1-\delta} L_S^{\delta-1}$. The marginal productivity condition for unskilled workers is $w_U = \partial Q/\partial L_U = (1-\alpha)\delta Q^{1-\delta} L_U^{\delta-1}$. Equation (7-10) follows by taking the ratio of the two marginal productivity conditions, and then taking logs.

FIGURE 7-8 Changes in the Wage Distribution Resulting from Shifts in Supply and Demand
The downward-sloping relative demand curve implies that employers hire relatively fewer skilled workers when their relative wage is high. The perfectly inelastic supply curve indicates that the relative number of skilled workers is fixed. The labor market is in equilibrium at point *A*. A rise in the relative supply of skilled workers to S_1 would lower their relative wage. The relative wage can rise only if there was also a sizable outward shift in the relative demand curve (to point *C*).

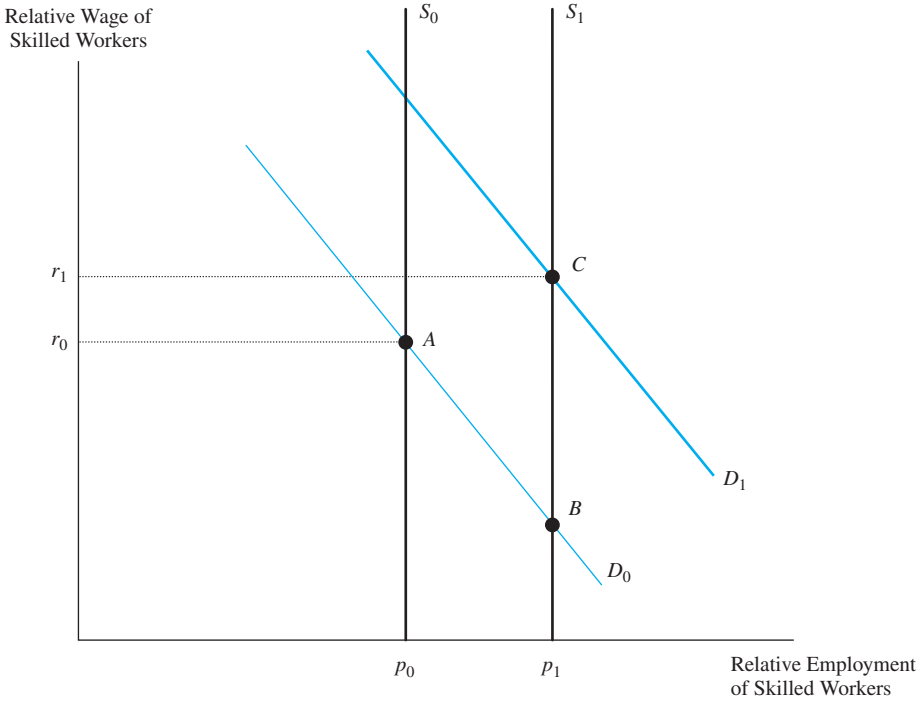


Figure 7-8 illustrates the basic model. For simplicity, suppose that the relative supply of skilled workers is perfectly inelastic; initially at fraction p_0. Put differently, a certain fraction of the workforce is skilled regardless of the percent wage gap between skilled and unskilled workers. Initially, the relative supply and demand curves are given by S_0 and D_0, respectively. The competitive labor market reaches equilibrium at point *A*. In equilibrium, the fraction p_0 of the workforce is skilled and the relative wage of skilled workers is given by r_0.

There are two ways for changing economic conditions to increase the wage gap between skilled and unskilled workers. The first is for the supply curve to shift to the left: A reduction in the relative number of skilled workers would drive up their relative wage. The second is for the demand curve to shift to the right: An increase in the relative demand for skilled workers would drive up their relative wage.

In fact, there was a sizable *increase* in the relative number of skilled workers in the United States in recent decades, shifting the relative supply curve to the right (to S_1). This supply shift should have moved the labor market to point *B, reducing* the relative wage of skilled workers. The type of supply shift that actually occurred, therefore, cannot possibly explain why the relative wage of skilled workers rose so rapidly.

To generate a rise in the relative wage of skilled workers, the model in Figure 7-8 implies that it the relative demand curve must have shifted to the right, to D_1. If the demand shift is sufficiently large, the final equilibrium at point *C* is characterized by an increase in the fraction of skilled workers *and* by a larger wage gap between skilled and unskilled workers.

Any attempt at understanding the rapid rise in the relative wage of skilled workers must then identify factors that increased their relative demand. In a sense, the relative supply and demand curves for skilled workers were in a race in recent years—both curves shifting to the right. The observed rise in wage inequality suggests that the demand shifts "won" the race; the relative demand for skilled workers rose at a faster rate than their relative supply. Although there has been a lot of debate over how to best explain the increase in inequality, the available evidence isolates a few variables that have become the "usual suspects" in any analysis of the changes in the wage distribution.

Supply Shifts

The rise in wage inequality in the 1980s and 1990s cannot be explained by a change in the relative supply of skilled workers. The relative supply of skilled workers went up in those decades, not down. Nevertheless, some of the changes in the wage structure prior to 1980 can probably be attributed to supply shifts.

The labor market entry of the well-educated baby boom cohort in the 1960s and 1970s increased the relative supply of college graduates at the time. The fraction of workers who were college graduates doubled between 1960 and 1980, from 9.3 to 17.9 percent. As Figure 7-7*b* shows, the relative wage of college graduates fell through the 1970s, suggesting that the baby boom supply shock depressed the payoff to a college education throughout that decade.[23]

One recent supply shift that has attracted a lot of attention is the increase in the number of immigrants. The fraction of the population that is foreign-born rose from 4.7 to 12.9 percent between 1970 and 2010. The relative demand curve in equation (7-10) implies that the immigration-induced supply shift would not affect the relative wage of skilled and unskilled workers if immigration was "balanced," in the sense that it had the same skill composition as the native-born workforce. A balanced immigrant flow would not change the relative supply of skilled workers. It turns out, however, that the immigration that actually occurred between 1979 and 1995 increased the supply of high school dropouts by 20.7 percent but increased the supply of workers with at least a high school education by only 4.1 percent.[24]

The immigration supply shift, therefore, greatly increased the relative number of workers at the very bottom of the skill distribution (shifting the supply curve in Figure 7-8 to the *left*). The wage of high school dropouts relative to that of high school graduates fell by 14.9 percent during the 1979–1995 period. The CES framework suggests that perhaps a third of this decline can be traced to the relative increase in low-skill immigration.

[23] Finis Welch, "Effects of Cohort Size on Earnings: The Baby Boom Babies' Financial Bust," *Journal of Political Economy* 87 (October 1979, Part 2): S65–S97; and Richard B. Freeman, *The Overeducated American,* New York: Academic Press, 1976.

[24] George J. Borjas, Richard B. Freeman, and Lawrence F. Katz, "How Much Do Immigration and Trade Affect Labor Market Outcomes?" *Brookings Papers on Economic Activity* (1997): 1–67.

Therefore, shifts in the relative supply curve, including the labor market entry of the college graduates in the baby boom cohort or the increase in the number of unskilled immigrants, can explain some of the changes in the wage distribution. But supply shifts alone cannot explain the overall increase in wage inequality. The number of college graduates relative to the number of high school graduates continued to rise after the 1980s—at the same time that the relative wage of college graduates was rising. Moreover, the rise in wage inequality *within* skill groups may have little to do with immigration. In short, it is not possible to provide a more complete explanation of the increase in wage inequality without resorting to a story where shifts in the relative demand curve play the dominant role.

International Trade

Part of the increase in the relative demand for skilled workers can probably be attributed to the increased internationalization of the economy, which exposes many American workers to competition from foreign workers.[25] Not surprisingly, the United States tends to export different types of goods than it imports. The workers employed in the importing industries tend to be low-skilled, while those employed in the exporting industries tend to be high-skilled.

The internationalization of the American economy—with rising exports and even more rapidly rising imports—would then have a beneficial impact on the relative demand for skilled workers. As foreign consumers increased their demand for the types of goods produced by American workers, the labor demand for skilled workers would rise. And as American consumers increased their demand for foreign goods, domestic firms would demand fewer unskilled workers. The impact of international trade, therefore, can be graphically represented as an outward shift in the relative labor demand curve in Figure 7-8.

Many of the workers hardest hit by imports were employed in manufacturing industries that were highly concentrated, unionized, and paid relatively high wages (such as automobiles and steel).[26] The high degree of concentration suggests that these industries were quite profitable. Because the industries were unionized, the bargaining agreements ensured that the profits were shared between the stockholders and the workers. As foreign competitors entered those markets, part of the "excess" wage paid to American workers was transferred abroad. Moreover, as the targeted industries cut employment, many of the affected workers moved to the nonunion, competitive sectors of the labor market, also pushing down the competitive wage.

Many researchers have attempted to measure the contribution of foreign trade to the changes in the wage distribution. In recent years, a number of influential studies strongly suggest the existence of a "trade effect," particularly in the context of U.S. trade with China. It seems that increased Chinese imports reduced the manufacturing wage in the typical locality by about 1 percent.[27]

[25] Kevin M. Murphy and Finis Welch, "The Role of International Trade in Wage Differentials," in Marvin Kosters, editor, *Workers and Their Wages,* Washington, D.C.: AEI Press, 1991, pp. 39–69.

[26] George J. Borjas and Valerie A. Ramey, "Foreign Competition, Market Power, and Wage Inequality," *Quarterly Journal of Economics* 110 (November 1996): 1075–1110.

[27] David H. Autor, David Dorn, and Gordon H. Hanson, "The China Syndrome: Local Labor Market Effects of Import Competition in the United States," *American Economic Review* 103 (October 2013): 2121–2168; and David Autor, David Dorn, Gordon Hanson, and Jae Song, "Trade Adjustment: Worker Level Evidence," *Quarterly Journal of Economics* 129 (November 2014): 1799–1860.

Skill-Biased Technological Change

Suppose the technological advances of recent decades have created a capital infrastructure that is a good substitute for unskilled workers but complements skilled workers. For example, the introduction of the personal computer into the workplace may have particularly increased the productivity of skilled workers. This type of technological change, called **skill-biased technological change,** would lower the demand for unskilled labor at the same time that it would increase the demand for skilled labor. Skill-biased technological change would then produce an outward shift in the relative labor demand curve illustrated in Figure 7-8.[28]

There is some disagreement over just how much of the increase in wage inequality can be attributed to skill-biased technological change. Some researchers have argued that this hypothesis can explain much of the increase in inequality.[29] But there is no widely accepted measure of skill-biased technological change that one can correlate with the changes in the wage distribution. Some studies, for example, rely on a "residual" method. In other words, let's account for the impact of supply shifts, immigration, international trade, and so on—and attribute whatever is left unexplained to skill-biased technological change. This approach is not completely satisfactory because it is attributing the effects of variables that we have not yet thought of or that are hard to measure to skill-biased technological change.

It may also be hard to reconcile the timing of the increase in wage inequality with the skill-biased technological change hypothesis.[30] Much of the increase in wage inequality occurred in the 1980s, but the information revolution continued (and perhaps accelerated) in the 1990s and beyond.

Moreover, there are data problems with the wage inequality time series that tend to overstate the increase in inequality during the 1990s. Depending on how these problems are handled, it may be that inequality within skill groups might have even *declined* slightly during the 1990s. It would be difficult to explain these shifts in terms of the technological change story unless one is willing to believe that technological change was biased in favor of skilled workers in the 1980s and then biased against them in the 1990s.

[28] Ann P. Bartel and Nachum Sicherman, "Technological Change and Wages: An Interindustry Analysis," *Journal of Political Economy* 107 (April 1999): 285–325; Stephen Machin and John Van Reenen, "Technology and Changes in Skill Structure: Evidence from Seven OECD Countries," *Quarterly Journal of Economics* 113 (November 1998): 1215–1244; and Mark Doms, Timothy Dunne, and Kenneth Troske, "Workers, Wages, and Technology," *Quarterly Journal of Economics* 112 (February 1997): 217–252. The literature is surveyed in Daron Acemoglu, "Technical Change, Inequality, and the Labor Market," *Journal of Economic Literature* 40 (March 2002): 7–72.

[29] John Bound and George Johnson, "Changes in the Structure of Wages in the 1980s: An Evaluation of Alternative Explanations," *American Economic Review* 82 (June 1992): 371–392; and Eli Berman, John Bound, and Zvi Griliches, "Changes in the Demand for Skilled Labor within U.S. Manufacturing Industries: Evidence from the Annual Survey of Manufacturing," *Quarterly Journal of Economics* 109 (May 1994): 367–398.

[30] David Card and John E. DiNardo, "Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles," *Journal of Labor Economics* 20 (October 2002): 733–783; and Thomas Lemieux, "Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill?" *American Economic Review* 96 (June 2006): 461–498.

Theory at Work

COMPUTERS, PENCILS, AND THE WAGE DISTRIBUTION

In 1984, only 25 percent of workers in the United States used a computer at work. By 1997, half used a computer. The widespread adoption of computers in the workplace has been particularly important for highly educated workers. In 1997, 75 percent of college graduates used computers at work, as compared to only 11 percent of high school dropouts.

Workers who use a computer at work earn more than workers who do not. In 1989, the wage differential between the haves and have-nots was around 18 percent. Suppose we interpret this wage differential as the "returns to computer use"—how much a worker's earnings would increase if he or she began using a computer in the workplace. Because skilled workers are much more likely to use a computer at work, the Information Revolution could be a substantial contributor to the increasing wage gap between skilled and unskilled workers. This correlation, in fact, is often cited as an important piece of evidence for the hypothesis that skill-biased technological change played an important role in producing the increased wage inequality observed in the United States in the 1980s and 1990s.

However, the 18 percent wage differential between those who use computers and those who do not may have little to do with the rewards for using a computer in the workplace. Instead, it may just be the case that employers consciously choose the most productive workers to assign computers to. The 18 percent wage gap cannot then be interpreted as the returns to computer use; it is simply measuring the preexisting productivity differential between the two groups of workers.

Some evidence for this alternative interpretation is found in the German labor market, where it turns out that workers who use *pencils* at work earn about 14 percent more than workers who do not. Surely, one would not want to argue that the use of pencils at work—and the wage gap between those who use pencils and those who do not—provides supporting evidence for the skill-biased technological change hypothesis.

Sources: David H. Autor, Lawrence F. Katz, and Alan B. Krueger, "Computing Inequality: How Computers Changed the Labor Market," *Quarterly Journal of Economics* 113 (November 1998): 1169–1213; and John DiNardo and Jörn-Steffen Pischke, "The Returns to Computer Use Revisited: Have Pencils Changed the Wage Structure Too?" *Quarterly Journal of Economics* 112 (February 1997): 291–303.

In short, even though the skill-biased technological change hypothesis has been (and remains) a favored explanation for the increase in inequality, there are a number of questions that have yet to be resolved satisfactorily.

Labor Market Institutions

There has been a steady decline in the importance of unions in the U.S. labor market. In 1973, 24 percent of the workforce was unionized. By 2010, the proportion of workers in unions had fallen to 12 percent.

Unions have traditionally been considered institutions that are very effective in raising the wage of less-skilled workers. A relatively large number of the workers employed in unions do not have college diplomas. Unions have traditionally propped up the earnings of those workers, guaranteeing them a wage premium. In fact, as we will see in the chapter on labor unions, union workers get paid about 15 percent more than nonunion workers, even after adjusting for differences in the skills of those employed in the two sectors.

The weakened power of unions can be interpreted as an outward shift in the relative demand curve for skilled labor in Figure 7-8. Suppose unions provide a safety net for

less-skilled workers. As union power weakens, employers are less willing to hire the same number of less-skilled workers unless they receive a lower wage, thereby increasing the relative wage of skilled workers and shifting up the relative demand curve. Some studies suggest that perhaps 10 percent of the increasing wage gap between college graduates and high school graduates can be attributed to the decline in unions.[31]

One other institutional factor that has traditionally propped up the wage of low-skill workers is the minimum wage. The *nominal* minimum wage remained constant at $3.35 an hour between 1981 and 1989. In constant 1995 dollars, however, the minimum wage declined from $5.62 an hour in 1981 to $4.12 in 1990. If many of the unskilled workers happen to work at minimum-wage jobs, the decline in the real minimum wage will increase the wage gap between skilled and unskilled workers.

We can use a straightforward calculation to estimate the impact of the falling real minimum wage on wage inequality.[32] Specifically, we can create a "counterfactual" wage distribution where the real minimum wage was constant throughout the 1980s and assume that the higher level of the real minimum wage did not generate any additional unemployment—so that the sample of workers remained roughly constant. This calculation shows that the falling minimum wage had a substantial impact on wages at the bottom of the wage distribution. However, because few educated workers get paid the minimum wage, this hypothesis cannot provide a credible explanation for the increase in the wage gap between college graduates and high school graduates or for why wage inequality rose within the sample of educated workers.

And in the End. . .

Each of the usual suspects (that is, changes in labor supply, the deunionization of the labor market, minimum wages, international trade, immigration, and skill-biased technological change) seems to be able to explain some part of the change in the U.S. wage distribution.

The main lesson, however, is that no single "story" can explain the bulk of the changes. Some of the variables (for example, immigration or trade) may explain part of the increasing wage gap between skilled and unskilled workers but fails to explain why inequality increased within skill groups. Similarly, the stability of the nominal minimum wage may explain why the real wage of low-skill workers fell but cannot explain why the real wage of high-skill workers rose. And the leading explanation—skill-biased technological change—is not fully consistent with the timing of the changes in wage inequality.

Any story that we eventually develop must confront an additional empirical puzzle. As Table 7-4 shows, there are sizable international differences in how the wage distribution

[31] John DiNardo, Nicole Fortin, and Thomas Lemieux, "Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semi-Parametric Approach," *Econometrica* 64 (September 1996): 1001–1044; Richard B. Freeman, "How Much Has De-Unionization Contributed to the Rise in Male Earnings Inequality?" in Sheldon Danziger and Peter Gottschalk, editors, *Uneven Tides,* New York: Russell Sage, 1993, pp. 133–163; and David Card, Thomas Lemieux, and Craig W. Riddell, "Unions and Wage Inequality," *Journal of Labor Research* 25 (2004): 519–562.

[32] DiNardo, Fortin, and Lemieux, "Labor Market Institutions and the Distribution of Wages"; David Lee, "Wage Inequality in the United States during the 1980s: Rising Dispersion or Falling Minimum Wage," *Quarterly Journal of Economics* 114 (August 1999): 977–1023; and Coen Teulings, "The Contribution of Minimum Wages to Increasing Wage Inequality," *Economic Journal* 113 (October 2003): 801–833.

TABLE 7-4
International Trends in Wage Inequality for Male Workers (90–10 Percent Wage Gap)

Source: OECD, Employment Outlook, July 1996. Paris: OECD, Table 3.1.

Country	1984	1994
Australia	174.6	194.5
Canada	301.5	278.1
Finland	150.9	153.5
France	232.0	242.1
Germany	138.7	124.8
Italy	129.3	163.8
Japan	177.3	177.3
Netherlands	150.9	158.6
New Zealand	171.8	215.8
Norway	105.4	97.4
Sweden	103.4	120.3
United Kingdom	177.3	222.2
United States	266.9	326.3

evolved in the 1980s and 1990s. The percent wage gap between the 90th percentile and the 10th percentile worker *rose* from 177 to 222 percent in the United Kingdom but *fell* from 139 to 125 percent in Germany. Presumably, skill-biased technological change would have affected most of these economies at the same time, suggesting that the wage distribution of these countries should have changed in roughly similar ways.

But different countries have very different labor market institutions—particularly when it comes to the safety net that protects the economic outcomes of low-skilled workers.[33] It turns out that the various countries also experienced different trends in the unemployment rate. The unemployment rate in the United States declined throughout much of the 1990s—at the same time that the unemployment rate in many western European countries was rising rapidly.

Perhaps the changes in wage inequality and unemployment are reverse sides of the same coin.[34] The same factors that led to greater wage inequality in the United States, where the labor market permits wage dispersion to grow and persist, manifested itself as higher unemployment rates in those countries where the safety net did not let wages change. In short, the labor market in some countries may have responded to the increase in the relative demand for skilled workers by changing quantities (that is, employment). In other countries, the market responded by changing prices (that is, wages).

[33] See Francine D. Blau and Lawrence M. Kahn, "International Differences in Male Wages Inequality: Institutions versus Market Forces," *Journal of Political Economy* 104 (August 1996): 791–837; David Card, Francis Kramarz, and Thomas Lemieux, "Changes in the Relative Structure of Wages and Employment: A Comparison of the United States, Canada, and France," *Canadian Journal of Economics* 32 (August 1999): 843–877; and Christian Dustmann, Johannes Lundsteck, and Uta Schönberg, "Revisiting the German Wage Structure," *Quarterly Journal of Economics* 124 (May 2009): 843–881.

[34] Adrian Wood, "How Trade Hurt Unskilled Workers," *Journal of Economic Perspectives* 9 (Summer 1995): 57–80; and Stephen Nickell and Brian Bell, "Changes in the Distribution of Wages and Unemployment in OECD Countries," *American Economic Review* 86 (May 1996): 302–308.

7-9 Inequality Across Generations

Up to this point, we have analyzed how human capital investments generate wage inequality and how changes in supply and demand can change the wage distribution in significant ways within a very short time period. We now address whether the wage inequality we observed in a particular generation is transmitted to the next generation.

The link between the economic performance of parents and children, which is sometimes called the rate of social mobility, lies at the core of many hotly debated policy questions. Consider, for example, the debate over whether the lack of social mobility might contribute to the creation of a permanent "underclass," or the debate over how government policies affect the persistence of poverty and welfare dependency across generations.

We have assumed that workers invest in their own human capital. In fact, a large part of our human capital was chosen and funded by our parents, so it is useful to think of human capital investments in an intergenerational context. Parents care both about their own well-being and about the well-being of their children. As a result, parents will invest in their children's human capital.

The parental investments help create the link between the skills of parents and the skills of their children. High-income parents will typically invest more in their children, creating a positive correlation between the earnings of parents and children.

Figure 7-9 illustrates various possibilities for the regression line that connects the earnings of the two generations. The slope of this line is often called an **intergenerational correlation.** An intergenerational correlation equal to 1 (as in line *A* in the figure) implies that if the earnings gap between any two parents is $1,000, their children's income will also differ by $1,000. If the correlation was equal to 0.5, a $1,000 earnings gap between the two parents translates to a $500 earnings gap between their children.

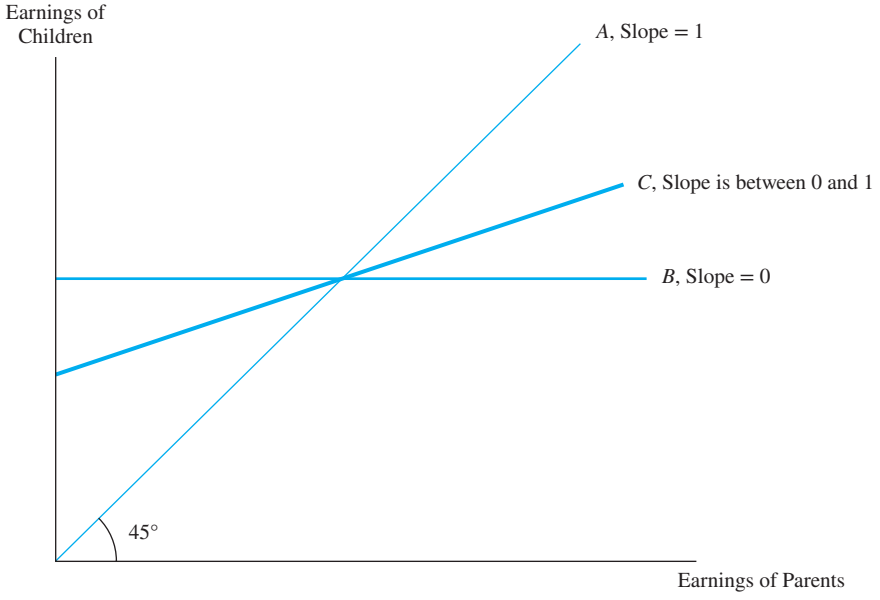
Empirical studies typically find that the intergenerational correlation is less than one, so that earnings differences between any two parents will typically exceed the expected earnings differences found between their children. This attenuation of income differences across generations is known as **regression toward the mean.**

This phenomenon may arise because parents do not devote their entire wealth to investing in their children's human capital—but rather consume some of it themselves. Regression toward the mean may also occur if the parents encounter diminishing returns when they try to invest in their children's human capital—the marginal cost of education would then rise rapidly as parents try to "inject" more schooling in their children. Finally, there is probably some regression toward the mean in ability—it is unlikely that the children of exceptionally bright parents will be just as bright. Note that the closer the intergenerational correlation gets to 0, the more important regression toward the mean becomes. If the intergenerational correlation were equal to 0 (as in line *B* in Figure 7-9), there would be complete regression toward the mean because none of the differences in parental skills are transmitted to their children.

Until the 1990s, it was generally believed that the intergenerational correlation between the earnings of fathers and sons was in the order of 0.2.[35] Put differently, if the wage differential between any two parents is about 50 percent, the wage differential between their

[35] See the discussion in Gary S. Becker and Nigel Tomes, "Human Capital and the Rise and Fall of Families," *Journal of Labor Economics* 4 (July 1986 Supplement): S1–S39.

FIGURE 7-9 The Intergenerational Link in Earnings
The slope of the regression line linking the earnings of the children and the earnings of the parents is called an intergenerational correlation. If the slope is equal to 1, the wage gap between any two parents persists entirely into the next generation and there is no regression toward the mean. If the slope is equal to 0, the wage of the children is independent of the wage of the parents and there is complete regression toward the mean.



children would be about 10 percent (or 50 percent × 0.2). If the intergenerational correlation was constant over time, the wage differential among the grandchildren would then be only 2 percent (or 50 percent × 0.2 × 0.2). An intergenerational correlation of 0.2, therefore, implies that there will be a lot of social mobility in the population. The placement of today's workers in the wage distribution is not a good predictor of where the grandchildren will end up.

More recent research, however, raises serious doubts about the validity of this conclusion.[36] These studies argue that the intergenerational correlation is much higher, perhaps about 0.4. The problem with the earlier evidence is that there is a lot of error in observed measures of parental skills. When workers are asked about the socioeconomic status of their parents, the responses about parental earnings are probably not very precise. This measurement error weakens the estimated correlation between the earnings of parents and children.

It turns out that if we adjust the data for this type of measurement error, the estimated intergenerational correlation jumps to 0.4, so that a 50 percent wage gap between two parents translates into a 20 percent wage gap between the children and an 8 percent wage gap between the grandchildren. Income differences among workers, therefore, are more persistent across generations.

[36] Gary Solon, "Intergenerational Mobility in the Labor Market," in Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics,* vol. 3A, Amsterdam: Elsevier, 1999, pp. 1761–1800.

Summary

- The observed age–earnings profile is upward sloping and concave. Earnings increase over the life cycle, but at a decreasing rate.
- General training is valuable in all firms. Specific training is valuable only in the firm that provides the training. Workers pay for and collect the returns from general training. Workers and firms share both the costs and the returns of specific training.
- The optimal timing of human capital investments over the life cycle implies that the age–earnings profile is upward sloping and concave.
- The positive correlation between human capital investments and ability implies that the wage distribution is positively skewed so that workers in the upper tail of the wage distribution receive a disproportionately large share of national income.
- The Gini coefficient measures the amount of inequality in an income distribution.
- Wage inequality rose rapidly in the 1980s and 1990s. Wage dispersion increased between education and experience groups, as well as within narrowly defined skill groups.
- Some of the changes in the wage structure can be explained by shifts in supply (such as the baby boom or immigration), the increasing globalization of the U.S. economy, institutional changes in the labor market (including the de-unionization of the labor market), and skill-biased technological change. No single variable, however, is a "smoking gun" that explains the bulk of the changes in the wage structure.
- Wage inequality among workers is transmitted from one generation to the next because parents care about the well-being of their children and invest in their children's human capital. But there is regression toward the mean, with the wage gap between any two families narrowing across generations.

Key Concepts

50–10 wage gap, *255*
90–10 wage gap, *255*
Constant elasticity of substitution (CES) production function, *257*
efficiency units, *244*
general training, *239*
Gini coefficient, *254*
intergenerational correlation, *265*
Lorenz curve, *253*
Mincer earnings function, *247*
positively skewed wage distribution, *250*
regression toward the mean, *265*
relative demand curve, *257*
skill-biased technological change, *261*
specific training, *239*
temporary layoffs, *243*

Review Questions

1. Discuss the difference between general training and specific training. Who pays for and collects the returns from each type of training?
2. Discuss the implications of general and specific training for the worker's age–earnings profile.
3. Why are experimental methods now commonly used to evaluate the impact of training programs? Discuss how and under what conditions we can use the results of an experiment to estimate the rate of return to the program.

4. Why is the wage distribution positively skewed?
5. Describe how to calculate a Gini coefficient.
6. Describe how the relative number of skilled workers and the increase in the relative demand for skilled workers helps determine the shape of the wage distribution.
7. Describe the key changes that occurred in the U.S. wage distribution during the 1980s and 1990s.
8. Why did the U.S. wage distribution change so much after 1980?
9. What factors determine how much parents invest in their children's human capital?
10. Discuss why there is regression toward the mean in the correlation between the earnings of parents and children.
11. Discuss the implications of regression toward the mean for the changing shape of the wage distribution across generations.

Problems

7-1. Evaluate the validity of the following claim: The increasing wage gap between highly educated and less educated workers will itself generate shifts in the U.S. labor market over the next decade. As a result of these responses, much of the "excess" gain currently accruing to highly educated workers will soon disappear.

7-2. What effect will each of the following proposed changes have on wage inequality?
   (a) Indexing the minimum wage to inflation.
   (b) Increasing the benefit level paid to welfare recipients.
   (c) Increasing wage subsidies paid to firms that hire low-skill workers.

7-3. From 1970 to 2000, the supply of college graduates to the labor market increased dramatically, while the supply of high school (no college) graduates shrunk. At the same time, the average real wage of college graduates stayed relatively stable, while the average real wage of high school graduates fell. How can these wage patterns be explained?

7-4. (a) Is the presence of an underground economy likely to result in a Gini coefficient that over states or under states poverty?
   (b) Consider a simple economy where 90 percent of citizens report an annual income of $10,000 while the remaining 10 percent report an annual income of $110,000. What is the Gini coefficient associated with this economy?
   (c) Suppose the poorest 90 percent of citizens actually have an income of $15,000 because each receives $5,000 of unreported income from the underground economy. What is the Gini coefficient now?

7-5. Use the two wage ratios for each country in Table 7-4 to calculate each country's percent increase in the 90–10 wage ratio from 1984 to 1994. Which countries experienced a compression in the wage distribution over this time? Which three countries experienced the greatest percent increase in wage dispersion over this time?

7-6. (a) What is the difference between income inequality and wealth inequality?
   (b) Most policies that target inequality either target it at the low end of the income distribution by trying to increase wages of low-income workers, or at the high

end of the income distribution by limiting wages of high-income workers. List a few potential policies of each type.

(c) In your opinion, should the government focus on the low end or the high end? Why?

(d) In order to better understand how sensitive inequality measures are to the choice of measure, provide a graph of an economy with a 90–10 wage gap that is essentially zero but for which the Gini coefficient is close to 1.

7-7. The two points for the international income distributions reported in Table 7-2 for countries in 2013 can be used to make a rough calculation of the Gini coefficient. Use a spreadsheet to estimate the Gini coefficient for each country. Which three countries had the most equal income distribution in 2013? Which three countries had the most unequal income distribution in 2013?

7-8. Most government-provided job training programs are optional to the worker. Describe how the self-selection issue might be used to call into question empirical results suggesting there are large economic benefits to be gained by requiring all workers to receive government-provided job training.

7-9. Before 1990, the 80–50 and the 50–20 log wage gap was higher for women than for men (see Figure 7-7). What are some possible reasons for this?

7-10. Jill is planning the timing of her on-the-job training investments over the life cycle. What happens to Jill's OJT investments if

(a) the market-determined rental rate to an efficiency unit falls?

(b) Jill's discount rate increases?

(c) the government passes legislation delaying the retirement age until age 70.

(d) technological progress is such that much of the OJT acquired at any given age becomes obsolete within the next 10 years.

7-11. Suppose two households earn $40,000 and $56,000 respectively. What is the expected percent difference in wages between the children, grandchildren, and great-grandchildren of the two households if the intergenerational correlation of earnings is 0.2, 0.4, or 0.6?

7-12. Suppose 50 percent of a population all receive an equal share of p percent of the nation's income while the other 50 percent of the population all receive an equal share of $1 - p$ of the nation's income where $0 \leq p \leq 50$.

(a) For any such p, what is the Gini coefficient for the country?

(b) For any such p, what is the 90 – 10 wage gap?

7-13. Consider two developing countries. Country A, though quite poor, uses government resources and international aid to provide public access to quality education. Country B, though also quite poor, is unable to provide quality education for institutional reasons. The distribution of innate ability is identical in the two countries.

(a) Which country is likely to have a more positively skewed income distribution? Why? Plot the hypothetical income distributions for both countries on the same graph.

(b) Which country is more likely to develop faster? Why? Plot the hypothetical income distributions in 20 years for both countries on the same graph.

7-14. Consider an economy with 10,000 individuals. Of them, 5,000 each earn $25,000; 3,000 each earn $40,000; and 2,000 each earn $100,000.
   (a) What is the Gini coefficient for this economy?
   (b) What would the Gini coefficient be if the wealthiest 2,000 individuals were taxed 30% of their income with the proceeds being transferred to the 5,000 poorest individuals?

7-15. Explain why the intergenerational correlation of earnings would likely be higher or lower than average for the following groups and factors in the United States:
   (a) Improved educational outcomes for all populations (for example, minority, low-income, rural).
   (b) The elimination of legacy admits to colleges and universities.
   (c) The implementation of a federal inheritance tax.

Selected Readings

Facundo Alvaredo, Tony Atkinson, Emmanuel Saez, and Thomas Piketty, "The Top 1 Percent in International and Historical Perspective." *Journal of Economic Perspectives* 27 (Summer 2013): 3–20.

David H. Autor, Lawrence F. Katz, and Melissa S. Kearney, "Trends in U.S. Wage Inequality: Revising the Revisionists," *Review of Economics and Statistics* 90 (May 2008): 300–323.

Thomas S. Dee and William N. Evans, "Teen Drinking and Educational Attainment: Evidence from Two-Sample Instrumental Variables Estimates," *Journal of Labor Economics* 21 (January 2003): 178–209.

John DiNardo, Nicole Fortin, and Thomas Lemieux, "Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semi-Parametric Approach," *Econometrica* 64 (September 1996): 1001–1044.

Lawrence F. Katz and Kevin M. Murphy, "Changes in Relative Wages, 1963–1987: Supply and Demand Factors," *Quarterly Journal of Economics* 107 (February 1992): 35–78.

Wojciech Kopczuk, Emmanuel Saez, and Jae Song, "Earnings Inequality and Mobility in the United States from Social Security Data Since 1937," *Quarterly Journal of Economics* 125 (February 2010): 91–128.

Thomas Lemieux, "Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill?" *American Economic Review* 96 (June 2006): 461–498.

Bruce Sacerdote, "How Large Are the Effects from Changes in Family Environment? A Study of Korean American Adoptees," *Quarterly Journal of Economics* 122 (February 2007): 119–157.

Gary Solon, "Intergenerational Income Mobility in the Labor Market," *American Economic Review* 82 (June 1992): 393–408.

Chapter 8

Labor Mobility

Immigration is the sincerest form of flattery.

—*Jack Paar*

Workers are continually searching for higher-paying jobs and firms are continually searching for cheaper workers. In a competitive labor market, these search activities equate the value of marginal product of labor across firms and across labor markets (for workers of given skills). The equilibrium allocation of workers to firms is efficient. No other allocation can increase the value of labor's contribution to national income.

Needless to say, actual labor markets are not quite as neat. Workers often do not know their own skills and abilities and are ill informed about the opportunities available in other jobs or in other labor markets. Firms do not know the true productivity of the workers they hire. As in a marriage, information about the value of the match between the worker and the firm is revealed slowly as both parties learn about each other. Therefore, the existing allocation of workers to firms is not efficient and other allocations are possible that would increase national income.

This chapter examines **labor mobility,** the mechanism that labor markets use to improve the allocation of workers to firms. There is a lot of mobility in the labor market. In fact, it seems as if the U.S. labor market is in constant flux. Nearly 4 percent of workers switch jobs in any given month, 2 percent of the population moves across state lines in a year, and over 1 million immigrants enter the country annually. All of these flavors of labor mobility are driven by the same fundamental factors: Workers want to improve their economic situation and firms want to hire more productive and cheaper workers.

The study of labor mobility helps us address a number of key questions in labor economics: What determines migration? How do the movers differ from the workers who choose to stay? How do movers adapt to the new work environment? And what are the consequences of migration?

8-1 Migration as a Human Capital Investment

In 1932, Nobel Laureate John Hicks proposed that "differences in net economic advantages, chiefly differences in wages, are the main causes of migration."[1] The analysis of migration decisions uses this hypothesis as the point of departure and views the migration of workers

[1] John R. Hicks, *The Theory of Wages,* London: Macmillan, 1932, p. 76; and Larry A. Sjaastad, "The Costs and Returns of Human Migration," *Journal of Political Economy* 70 (October 1962): 80–93.

as a form of human capital investment. Workers calculate the value of the employment opportunities available in each of the alternative destinations, subtract the cost of making the move, and choose whichever option maximizes the net present value of lifetime earnings.

Suppose there are two specific labor markets where the worker can be employed. These markets might be in different cities, in different states, or even in different countries. The worker is currently employed in New York and is thinking about moving to California. The worker, who is 20 years old, now earns w_{20}^{NY} dollars. If he were to move, he would earn w_{20}^{CA} dollars. It costs M dollars to make the move. These migration costs include the airfare and the expense of moving household goods, as well as the dollar value of the "pain and suffering" incurred when one moves away from family, friends, and social networks.

Like all other human capital investments, migration decisions are guided by the comparison of the present value of lifetime earnings in the alternative opportunities. Let PV^{NY} be the present value of the earnings stream if the person stays in New York. This quantity is given by

$$PV^{NY} = w_{20}^{NY} + \frac{w_{21}^{NY}}{(1+r)} + \frac{w_{22}^{NY}}{(1+r)^2} + \cdots \qquad \textbf{(8-1)}$$

where r is the discount rate and the sum in equation (8-1) continues until the worker reaches retirement age. Similarly, the present value of the earning stream if the person moves to California is

$$PV^{CA} = w_{20}^{CA} + \frac{w_{21}^{CA}}{(1+r)} + \frac{w_{22}^{CA}}{(1+r)^2} + \cdots \qquad \textbf{(8-2)}$$

The net gain to migration is then given by

$$\text{Net gain to migration} = PV^{CA} - PV^{NY} - M \qquad \textbf{(8-3)}$$

The worker moves if the net gain is positive. Several obvious and empirically testable propositions follow immediately from this framework:

1. An improvement in the economic opportunities available in the destination increases the net gains from migration and raises the probability that the worker moves.
2. An improvement in the economic opportunities at the current region of residence decreases the net gains from migration and lowers the probability that the worker moves.
3. An increase in migration costs lowers the net gains to migration and reduces the probability of a move.

All these implications deliver the same basic message. Migration occurs when there is a good chance that the worker will recoup his investment.[2]

[2] Although our discussion uses a two-region framework, the same insights follow if the worker is choosing a destination from many alternative regions. He would calculate the present value of earnings in each of the 50 states and would choose the one that maximized the present value of lifetime earnings net of moving costs.

8-2 Internal Migration

Americans are very mobile. Between 2016 and 2017, 2.1 percent of the population moved across counties within the same state, and another 2.1 percent moved across states or out of the country.[3] Many studies examine if the rate of migration between any two places within the United States—called *internal* migration—responds to differences in economic conditions (such as wages and unemployment rates) and to moving costs (typically measured by the distance involved in the move). The evidence suggests that the size and direction of internal migration flows are indeed consistent with the notion that workers migrate in search of better employment opportunities.[4]

These correlations help us understand the direction of some of the major internal migration waves in the United States. Between 1900 and 1960, for example, there was a sizable and steady flow of African–American workers from the rural South to the industrialized cities of the North.[5] In 1900, 90 percent of the African–American population lived in the South; by 1960, the fraction had declined to 60 percent. The size and direction of this migration should not be too surprising. The availability of better employment opportunities in the booming manufacturing sector of northern cities (as well as the discriminatory social and economic disadvantages black workers faced in the South) obviously persuaded many blacks to move north.

There is also strong evidence that the socioeconomic characteristics of workers, particularly their education and age, determine migration propensities. Migration is most common among younger and more-educated workers.

Figure 8-1 shows the relation between age and the probability that a worker will migrate across state lines. This probability declines systematically over the working life. About 4 percent of workers in their twenties move across state lines, but the probability falls to 1 percent by the time they reach their forties. Older workers are less likely to move because migration is a human capital investment, and older workers have a shorter period over which to collect the returns from the investment.

There is also a positive correlation between a worker's educational attainment and the probability of migration. Between 2016 and 2017, workers with a graduate degree had a 2.3 percent probability of moving across state lines, more than double the probability for high school graduates (which was 1.1 percent).

The positive correlation between education and internal migration might arise because highly educated workers may be more efficient at learning about opportunities in alternative labor markets, reducing migration costs. It is also likely that the geographic region that makes up the relevant labor market for highly educated workers is larger than the region that makes up the market for the less educated. Consider, for instance, the labor market

[3] U.S. Bureau of the Census, "Table 1. General Mobility, by Race and Hispanic Origin, Region, Sex, Age, Relationship to Householder, Educational Attainment, Marital Status, Nativity, Tenure, and Poverty Status: 2016 to 2017." www.census.gov/data/tables/2017/demo/geographic-mobility/cps-2017.html.

[4] Michael Greenwood, "Internal Migration in Developed Countries," in Mark R. Rosenzweig and Oded Stark, editors, *Handbook of Population and Family Economics,* vol. 1B, Amsterdam: Elsevier, 1997, pp. 647–720.

[5] Leah P. Boustan, "Competition in the Promised Land: Black Migration and Racial Wage Convergence in the North, 1940–1970," *Journal of Economic History* 69 (September 2009): 755–782.

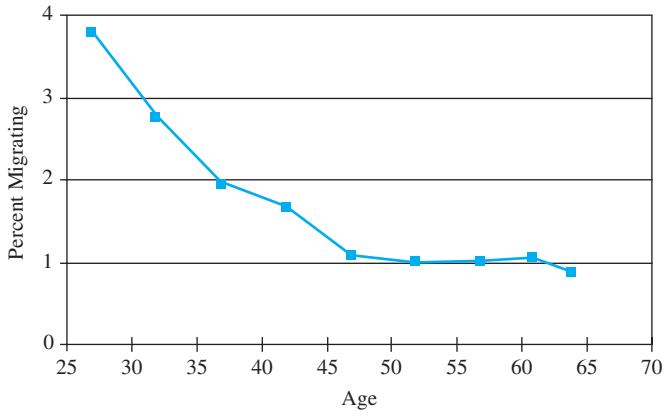


FIGURE 8-1
Probability of Migrating across State Lines between 2016 and 2017, by Age

Source: U.S. Bureau of the Census, "Table 1. General Mobility, by Race and Hispanic Origin, Region, Sex, Age, Relationship to Householder, Educational Attainment, Marital Status, Nativity, Tenure, and Poverty Status: 2016 to 2017."

faced by college professors. There are very few "firms" in any given city and professors' skills are very portable across employers. In effect, college professors sell their skills in a national (and perhaps even an international) labor market.

Return and Repeat Migration

Workers who have just migrated are extremely likely to move back to their original locations (generating **return migration** flows) and are also extremely likely to move onward to still other locations (generating **repeat migration** flows). The probability of a migrant returning to his state of origin within a year is about 13 percent, and the probability of a migrant moving on to yet another location is 15 percent.[6]

Unless economic conditions in the various regions change drastically soon after the move takes place, the high propensity of migrants to move again is *not* consistent with the income-maximization hypothesis. Prior to the initial migration, the worker's cost–benefit calculation indicated that a move from, say, Illinois to Florida maximized his present value of lifetime earnings (net of migration costs). How can a similar calculation made just a few weeks after the move indicate that returning to Illinois or perhaps moving on to Texas maximizes the worker's income?

Two factors can generate return and repeat migration flows. Some of the return moves arise because the worker learned that the initial migration decision was a mistake. A worker contemplating the move from Illinois to Florida faces a lot of uncertainty about conditions in Florida. Once he arrives, he might discover that the available employment opportunities or local amenities are far worse than expected. Return and repeat migration flows arise as workers attempt to correct these errors.

Return or repeat migration might also be the career path that maximizes the present value of lifetime earnings in some occupations, even in the absence of any uncertainty

[6] Julie DaVanzo, "Repeat Migration in the United States: Who Moves Back and Who Moves On?" *Review of Economics and Statistics* 65 (November 1983): 552–559; see also Christian Dustmann, "Return Migration, Wage Differentials, and the Optimal Migration Duration," *European Economic Review* 47 (April 2003): 353–367.

about job opportunities. For example, lawyers who specialize in tax law quickly realize that a brief stint at the Department of the Treasury, the Department of Justice, or the Internal Revenue Service in Washington, D.C., provides them with valuable human capital. This human capital includes intricate knowledge of the tax code and personal connections with government officials. After their government service, the lawyers can return to their home states or move to other areas of the country where their newly acquired skills will be highly rewarded. In effect, the temporary stay of the lawyers in the District of Columbia is one step in the career ladder that maximizes lifetime earnings.[7]

Why Is There So Little Migration?

Even though Americans are very mobile, the rate of internal migration is not sufficiently large to completely equalize wages across regions. The persistence of regional wage differentials raises an important question: Why do more workers *not* take advantage of the higher wage in some regions? The human capital model suggests an obvious answer: Moving costs must be very high. In fact, one can easily apply the model to get a rough idea of their magnitude.

In 2016, the typical tile setter in the construction industry in Puerto Rico earned $28,300; his counterpart in the United States earned $44,770.[8] The typical Puerto Rican tile setter, therefore, could increase his wage by over 50 percent by moving. Because Puerto Ricans are U.S. citizens by birth, there are no legal restrictions limiting their entry into the United States. In fact, the large income gap has motivated around a third of the Puerto Rican population to migrate to the United States since World War II. But, just as important, two-thirds of Puerto Ricans chose *not* to move.

Let w_{PR} be the wage the worker can earn in Puerto Rico and let w_{US} be the wage he can earn in the United States. For simplicity, let's assume these wages are constant over the life cycle. Suppose that the sums in equations (8-1) and (8-2) have many terms, so that the worker lives practically forever. We can then write the discounted present values as[9]

$$PV_{PR} = \frac{(1+r)w_{PR}}{r} \quad \text{and} \quad PV_{US} = \frac{(1+r)w_{US}}{r} \tag{8-4}$$

A Puerto Rican worker is indifferent between moving and staying if the discounted gains from moving are exactly equal to moving costs.

$$\frac{(1+r)(w_{US} - w_{PR})}{r} = M \tag{8-5}$$

To get an idea of how large M must be for a worker to be indifferent, consider the following algebraic rearrangement of equation (8-5): Divide both sides by w_{PR} and let $\pi = M/w_{PR}$.

[7] Sherwin Rosen, "Learning and Experience in the Labor Market," *Journal of Human Resources* 7 (Summer 1972): 326–342.

[8] U.S. Bureau of Labor Statistics, *National Occupational Employment and Wage Estimates,* Washington, D.C.: Occupational Employment Statistics, 2016; available at www.bls.gov/oes/tables.htm. These wage differences do not adjust for price differences. The wage gap remains large even if income is adjusted for differences in purchasing power.

[9] Let $S = 1 + 1/(1+r) + 1/(1+r)^2$ and so on. This implies that $(1+r)S = (1+r) + 1 + 1/(1+r) + 1/(1+r)^2$ and so on. After canceling out many terms, the difference $(1+r)S - S = 1 + r$, so $S = (1+r)/r$.

The variable π gives the fraction of a worker's salary that is spent on moving costs. We can then rewrite the equation as

$$\frac{1+r}{r} \cdot \frac{w_{US} - w_{PR}}{w_{PR}} = \pi \qquad \textbf{(8-6)}$$

The ratio $(w_{US} - w_{PR})/w_{PR}$ is about 0.5, indicating that a tile setter can increase his wage by about 50 percent by migrating to the United States. If the rate of discount is 5 percent, the left-hand side of equation (8-6) takes on the value of 10.5. In other words, migration costs for a worker who is indifferent between migrating or not equal 10.5 times his salary. If the tile setter earns the average wage in Puerto Rico (or $28,300), moving costs are around $300,000.[10]

What exactly is the nature of this expense? It obviously does not represent the cost of transporting the family and household goods to a new location. Instead, the marginal Puerto Rican probably attaches a very high utility value to the social and cultural amenities in his birthplace. Moving costs are likely to be even larger in other contexts—such as international migration, where there are legal restrictions and much greater differences in language and culture. In short, although internal migration improves labor market efficiency, the potential gains are limited by the fact that the flow of migrants is often not sufficiently large.

8-3 Family Migration

The model of internal migration examined the choice made by a single worker as he or she compared employment opportunities across regions and chose the one location that maximized the present value of lifetime earnings. Most migration decisions, however, are not made by single workers, but by families. The decision, therefore, should not be based on whether a particular member of the household is better off by moving, but on whether the family *as a whole* is better off.[11]

Consider a household composed of two persons, a husband and a wife. Let's denote by ΔPV_H the change in the present value of the husband's earnings stream if he were to move geographically (say from New York to Texas). And let ΔPV_W be the change in the present value of the wife's earnings stream if she were to make the same move. Note that ΔPV_H can be interpreted as the husband's gains to migration if he were single and were making the migration decision completely on his own. These gains are called the husband's "private" gains to migration. If the husband were not tied down by family responsibilities, he would migrate if the private gains ΔPV_H were positive. Similarly, the quantity ΔPV_W gives the wife's private gains to migration. And if she were single, she would move when ΔPV_W is positive.

The family unit (that is, the husband and the wife) will move if the *family's* net gains are positive:

$$\Delta PV_H + \Delta PV_W > 0 \qquad \textbf{(8-7)}$$

[10] A more sophisticated calculation of moving costs would allow for worker heterogeneity and for wages to vary over the work life; see John Kennan and James R. Walker, "The Effect of Expected Incomes on Individual Migration Decisions," *Econometrica* 79 (January 2011): 211–251; and Erhan Artuc, Shubham Chaudhuri, and John McLaren, "Trade Shocks and Labor Adjustment: A Structural Empirical Approach," *American Economic Review* 100 (June 2010): 1008–1045.

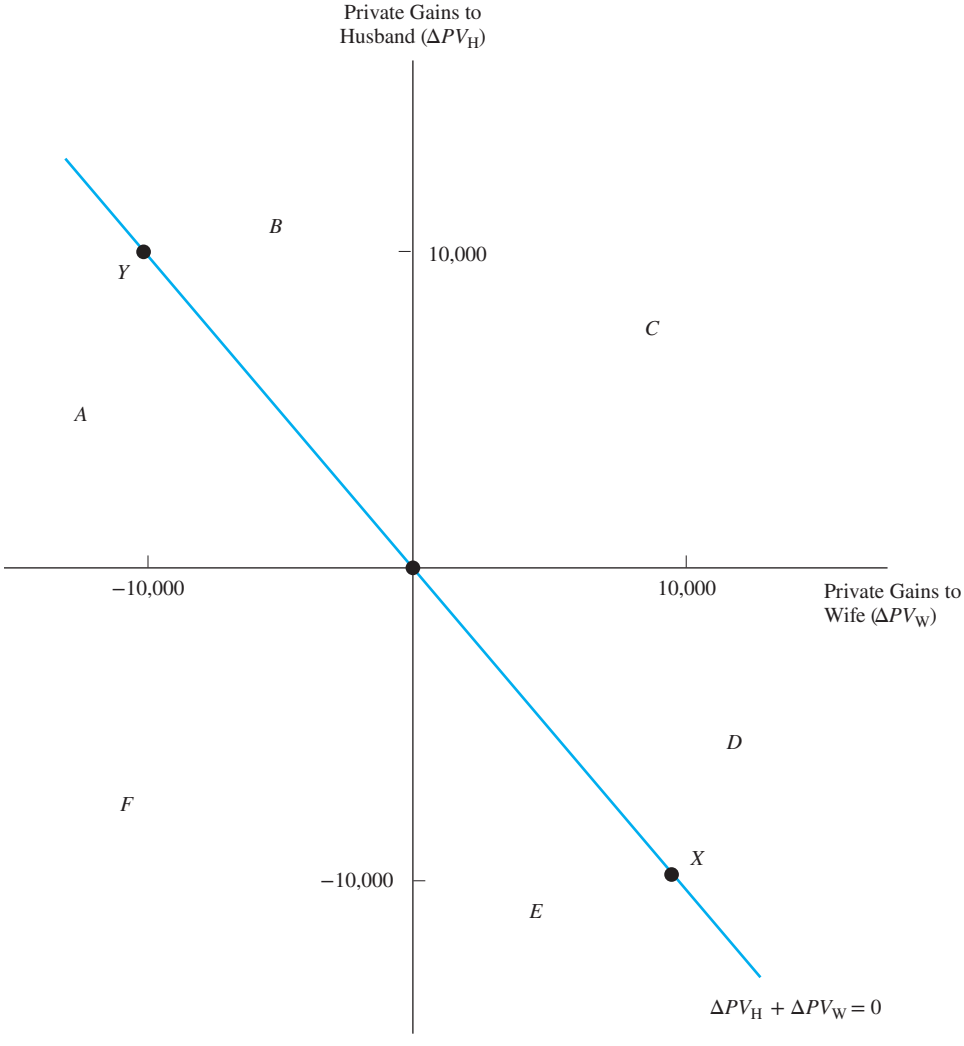
[11] Jacob Mincer, "Family Migration Decisions," *Journal of Political Economy* 86 (October 1978): 749–773.

In other words, the family migrates if the sum of the gains to the husband and to the wife is positive.

Figure 8-2 illustrates the basic idea. The vertical axis in the figure measures the husband's private gains to migration, and the horizontal axis measures the wife's private gains. If the husband were making the migration decision on his own, he would migrate whenever ΔPV_H was positive, which is given by the outcomes that lie above the horizontal axis

FIGURE 8-2 Tied Movers and Tied Stayers

If the husband were single, he migrates if $\Delta PV_H > 0$ (areas *A, B,* and *C*). If the wife were single, she migrates if $\Delta PV_W > 0$ (areas *C, D,* and *E*). The family migrates when the sum of the private gains is positive (areas *B, C,* and *D*). In area *D,* the husband would not move if single but moves as part of the family, making him a tied mover. In area *E,* the wife would move if single but does not move as part of the family, making her a tied stayer.



(or the combination of areas *A, B,* and *C*). Similarly, if the wife were making the migration decision on her own, she would migrate whenever ΔPV_W was positive, which is given by the outcomes to the right of the vertical axis (or areas *C, D,* and *E*).

Let's now examine the family's decision. The 45° downward-sloping line that goes through the origin connects the points where the net gains to the family are zero, or $\Delta PV_H + \Delta PV_W = 0$. The family might have zero gains from migration in a number of ways. For instance, at point *X,* the wife gains $10,000 if she was to move, but the husband loses $10,000. At point *Y,* the husband gains $10,000, but the wife loses $10,000.

The family moves if the sum of the private gains $\Delta PV_H + \Delta PV_W$ is positive. The decision to maximize the *family's* lifetime earnings implies that the family will move whenever the gains lie above the 45° line, or the combination of areas *B, C,* and *D.* The area in which the family wants to move, therefore, does not coincide with the areas indicating what each person in the family would do if he or she were single. In other words, the *optimal decision for the family is not necessarily the same as the optimal choice for a single person.*

Tied Stayers and Tied Movers

To see why the family's incentives to migrate differ from the private incentives of each family member, consider any point in area *E.* In this area, the wife would move on her own, for there are private gains to her move (that is, $\Delta PV_W > 0$). But the husband's loss exceeds her gain (so that $\Delta PV_H + \Delta PV_W < 0$). So it is not optimal for the family to move. The wife is, in effect, a **tied stayer.** She sacrifices the better employment opportunities available elsewhere because her husband is much better off at their current location.

Similarly, consider any point in area *D.* In this area, the husband experiences an income loss if he moves on his own (that is, $\Delta PV_H < 0$). But when he moves as part of a family unit, the wife's gain exceeds the husband's loss, so that $\Delta PV_H + \Delta PV_W > 0$. The husband is a **tied mover.** He follows the wife even though he is better off at his current job.

The analysis of family migration decisions shows that all persons in the family need not have positive private gains from migration. A comparison of the premigration and postmigration earnings of tied movers would indicate that they "lost" from the migration. In fact, some evidence suggests that the postmigration earnings of women are often lower than their premigration earnings.[12]

The rapid rise in the female labor force participation rate implies that both husbands and wives increasingly find themselves in situations in which their private incentives to migrate do not coincide with the family's incentives. Because both the spouses are often looking for work in the same city and sometimes even in the same narrowly defined profession, the chances of finding adequate jobs for the two parties are slim, reducing the likelihood that the family will move.

The increase in the number of two-worker households has given rise to creative labor market arrangements. Employers interested in hiring one of the spouses facilitate the job search process for the other and sometimes even hire both. There also has been an increase in the number of married couples who maintain separate households in different cities, minimizing the financial losses of being tied movers or tied stayers.

[12] Steven H. Sandell, "Women and the Economics of Family Migration," *Review of Economics and Statistics* 59 (November 1977): 406–414.

Theory at Work

POWER COUPLES

The number of "power couples" in the United States, couples in which both the spouses are college graduates, is rising rapidly, from 2 percent in 1940, to 9 percent in 1970, and to 15 percent in 1990. Because highly educated women are more likely to participate in the labor force, power couples are predominantly dual-career couples. In 1990, 73.3 percent of the women in these couples were working.

The fact that both the spouses in a power couple tend to work makes it difficult for both to obtain their "optimal" jobs in the same geographic labor market. Power couples may then have to split and reside in different cities, or one of the spouses will have to accept that he or she is a tied stayer or a tied mover, and work at a job that does not maximize their earnings potential.

Power couples can minimize the problem by settling in parts of the country that provide ample employment opportunities for high-skill workers, such as large metropolitan areas. The diversified labor markets in large cities can potentially provide satisfactory job matches for both the spouses. It turns out that this is precisely what power couples have done in the past few decades. The table below summarizes the evidence.

The proportion of power couples settling in a large metropolitan area rose from 14.6 to 34.8 percent between 1970 and 1990. In contrast, the fraction of "low-power couples" (where neither of the spouse is a college graduate) living in these areas rose only from 8.3 to 20.0 percent. If we treat the locational decisions made by low-power couples as the choice of a control group, the difference-in-differences approach implies that being in a power couple increases the probability of residing in a large metropolitan area by 8.5 percentage points. It seems, therefore, that power couples chose to reduce the cost of being tied stayers or tied movers by settling in different parts of the country.

Sources: Dora L. Costa and Matthew E. Kahn, "Power Couples: Changes in the Locational Choice of the College Educated, 1940–1990," *Quarterly Journal of Economics* 115 (November 2000): 1287–314; Janice Compton and Robert A. Pollak, "Why Are Power Couples Increasingly Concentrated in Large Metropolitan Areas," *Journal of Labor Economics* 25 (July 2007): 475–512.

Percentage of Couples with Working Wives Residing in Large Metropolitan Areas

	1970	1990	Difference
Power couples	14.6	34.8	20.2
Low-power couples	8.3	20.0	11.7
Difference-in-differences	—	—	**8.5**

8-4 The Self-Selection of Migrants

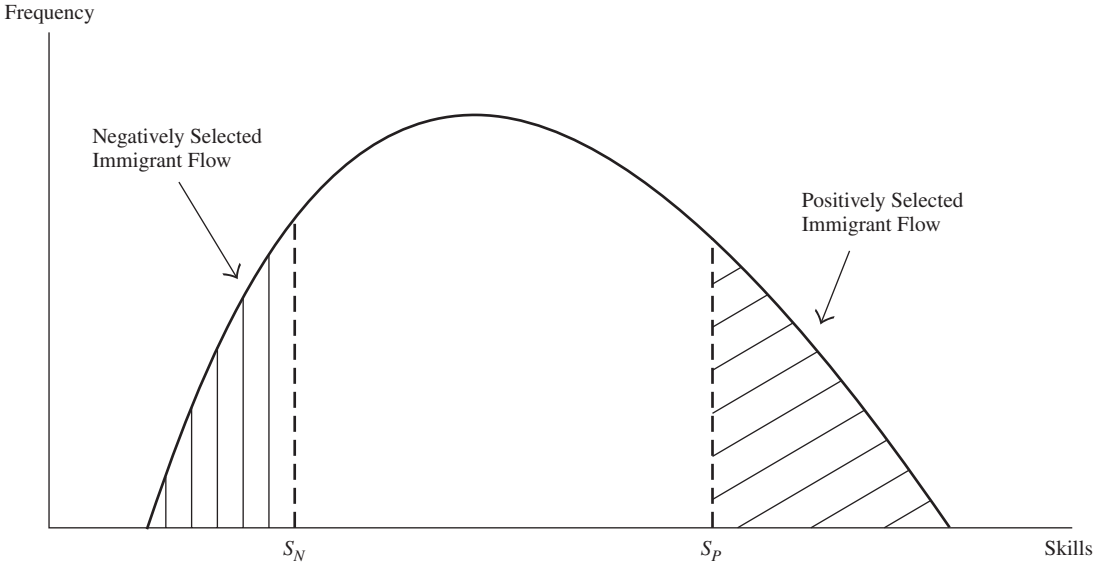
Because of the surge in the number of immigrants in the United States and in many other countries, much of the research interest on labor mobility in recent years has shifted to analyzing the determinants and consequences of international migration.

One frequent observation in immigrant-receiving countries is that the wage of immigrants varies by nationality. In the United States, for example, immigrants from Europe and Canada earned 36 percent more than natives, while those from Mexico earned 46 percent less.[13]

[13] Francine D. Blau and Christopher Mackie, editors, *The Economic and Fiscal Consequences of Immigration,* Washington, D.C.: National Academies Press, 2016, p. 107.

FIGURE 8-3 The Distribution of Skills in the Source Country

The distribution of skills in the source country gives the frequency of workers at each skill level. If immigrants have above-average skills, the immigrant flow is positively selected. If immigrants have below-average skills, the immigrant flow is negatively selected.



Part of the variation probably arises because skills acquired in industrialized labor markets may be more easily transferable to the American setting. The skills sought by firms in industrialized countries likely resemble the skills rewarded by American employers. But there may also be variation because different types of immigrants come from different countries. Migrants are not randomly chosen from the population; they are self-selected. And this self-selection raises one of the most important questions in the study of labor mobility: Which subset of workers finds it worthwhile to move, the most skilled or the least skilled?

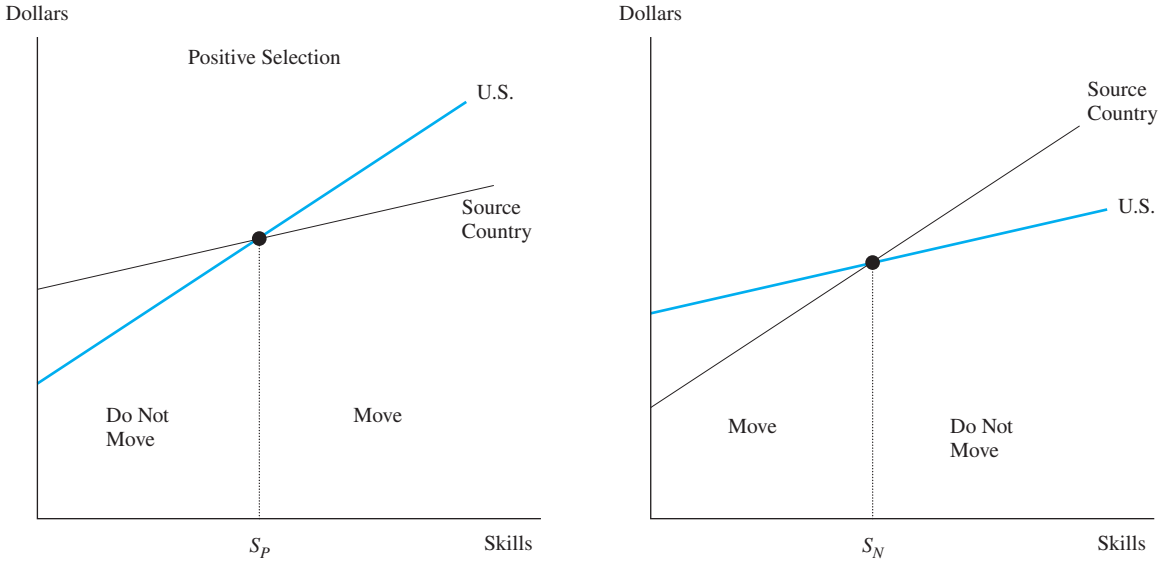
The influential **Roy model** describes how workers sort themselves among employment opportunities.[14] Workers now residing in a source country are trying to decide if they should migrate to the United States. Suppose that earnings in both countries depend only on the worker's skills, and that skills are general and completely transferable. Let s denote the number of efficiency units embodied in the worker. Figure 8-3 illustrates the frequency distribution of skills in the source country's population, and we wish to determine which subset of workers from that distribution chooses to emigrate.

Each worker makes the migration decision by comparing earnings in the two countries. To simplify, let's initially assume that workers do not incur any moving costs. The decision

[14] Andrew D. Roy, "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers* 3 (June 1951): 135–146. The model was first applied to the migration decision by George J. Borjas, "Self-Selection and the Earnings of Immigrants," *American Economic Review* 77 (September 1987): 531–553.

FIGURE 8-4 The Self-Selection of Immigrants

(*a*) If the return to skills is higher in the United States than in the source country (so that the wage–skills line is steeper in the United States), the immigrant flow is positively selected. Workers with more than s_P efficiency units find it profitable to move. (*b*) If the return to skills is lower in the United States, the immigrant flow is negatively selected. Workers with fewer than s_N efficiency units emigrate.



(*a*) Positive Selection

(*b*) Negative Selection

rule that determines migration is then very simple: A worker migrates to the United States whenever the U.S. earnings exceed earnings in the source country.[15]

Figure 8-4 illustrates the relation between wages and skills in each of the countries. The slope of the wage–skill lines gives the dollar payoff to an additional efficiency unit. Consider the sorting in Figure 8-4*a,* where the U.S. wage–skills line is steeper, implying that the payoff to an efficiency unit is higher in the United States than in the source country. Workers with fewer than S_P efficiency units earn more if they stay in the source country than if they emigrate. Workers with more than S_P efficiency units, however, earn more in the United States than in the source country.

The migration flow is then composed of workers in the upper tail of the skill distribution illustrated in Figure 8-3. This type of self-selection is called **positive selection.** Immigrants, on average, will be very skilled and do quite well in the United States.

In Figure 8-4*b,* the wage–skill line is steeper in the source country, so the payoff to skills is higher there. Workers with fewer than s_N efficiency units earn more in the United States and will want to move. But workers who have more than s_N efficiency units have higher earnings in the source country and will stay put. The migrant flow will be composed of the

[15] The discussion is also implicitly assuming that immigration policy does not restrict the entry of any immigrants who find it worthwhile to move to the United States.

least-skilled workers. This type of self-selection is called **negative selection.** Immigrants, on average, are unskilled and perform poorly in the United States.

The key insight of the Roy model is straightforward: *The relative payoff for skills across countries determines the skill composition of the migrant flow.* Workers "selling" their skills behave just like firms selling their products. Both workers and goods flow to those markets where they can get the highest price.

The model produces a rule of thumb that can help us understand the skill composition of migration flows. Consider workers residing in a country that offers a low rate of return to a worker's human capital so that the skilled do not earn much more than the unskilled. This is typical in countries such as Sweden or Denmark that have relatively egalitarian income distributions and almost confiscatory income tax systems. Relative to the United States, these countries' tax able workers and insure the unskilled against poor labor market outcomes. The situation generates incentives for the skilled to emigrate because they have the most to gain by moving. The United States would then be the recipient of a "brain drain."

But consider instead workers in source countries that offer a high rate of return to human capital. This is typical in countries with substantial income inequality, as in many developing countries. In this case, it is the United States that taxes the skilled and subsidizes the unskilled (relative to the source country). The United States becomes a magnet for workers with relatively low earnings capacities.

The available evidence suggests that there is indeed a negative correlation between measures of the source country's income inequality (which proxies for the rate of return to skills) and the earnings of immigrants in the United States.[16]

Recent research on the self-selection of migrants has moved beyond estimating these types of correlations and instead examines data sets that report the skills and earnings of workers at the origin before any migration taking place. The data allow a comparison of the *premigration* characteristics of the eventual movers with the characteristics of those who stayed.

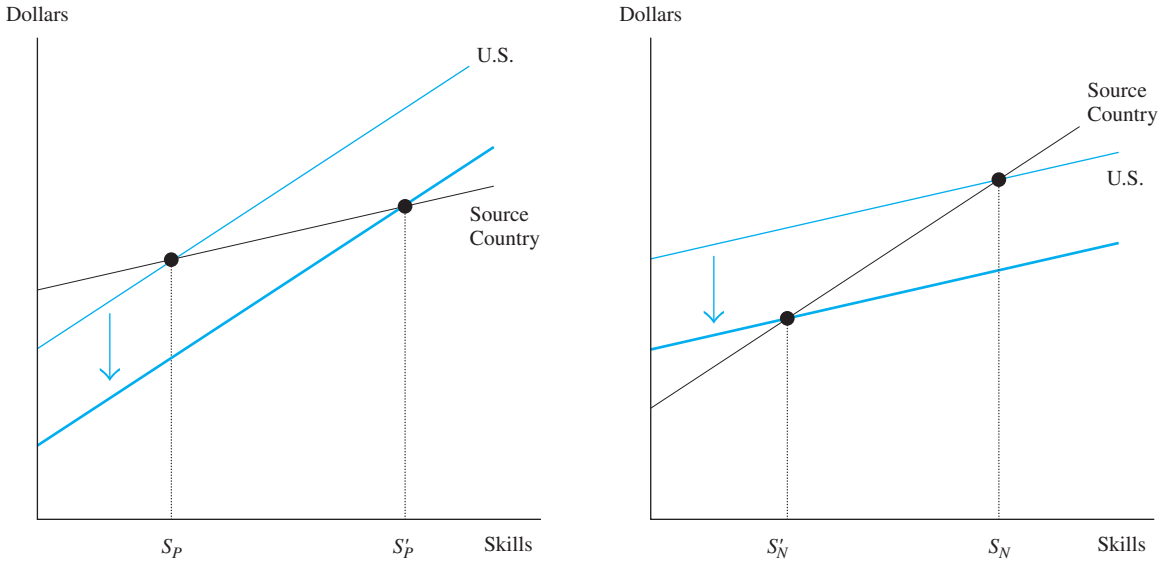
In the context of international migration, administrative data from Denmark—which contain information for the *entire* Danish population—shows that emigration from Denmark is marked by a very strong positive selection.[17] The Danish workers who eventually emigrated were highly skilled and had high earnings prior to their move. This finding is consistent with the implications of the Roy model because Denmark has a very low rate of return to skills and little income inequality. A related study of internal migration between Northern and Southern Italy also documents labor flows that are consistent with the Roy model.[18] The labor market in Northern Italy offers a

[16] A survey of the evidence is given by George J. Borjas, *Immigration Economics,* Cambridge, MA: Harvard University Press, 2014. See also Matthias Parey, Jens Ruhose, Fabian Waldinger, and Nicolai Netz, "The Selection of High-Skilled Emigrants," *Review of Economics and Statistics* 99 (December 2017): 776–792.

[17] George J. Borjas, Ilpo Kauppinen and Panu Poutvaara, "Self-Selection of Emigrants: Theory and Evidence on Stochastic Dominance in Observable and Unobservable Characteristics," *Economic Journal,* forthcoming 2018.

[18] Christian Bartolucci, Claudia Villosio, and Mathis Wagner, "Who Migrates and Why? Evidence from Italian Administrative Data," *Journal of Labor Economics,* forthcoming 2018.

FIGURE 8-5 The Impact of a Drop in U.S. Incomes
If U.S. incomes drop (or if there is an increase in moving costs), the wage–skills line for the United States shifts down and fewer workers migrate. The drop in U.S. incomes, however, does not change the type of selection that characterizes the immigrant flow.



(*a*) Positive Selection

(*b*) Negative Selection

relatively low rate of return to skills, and it attracts a negatively selected migrant flow from the South.[19]

Change in Income Levels and Moving Costs

One surprising implication of the Roy model is that the "baseline level" of income in the source country or in the United States (as measured by the height of the wage–skills lines in Figure 8-4) does not determine the type of selection that characterizes the immigrants. Income levels, however, do affect the number of immigrants.

Suppose that U.S. incomes drop because of a severe recession, pushing down the wage–skills line. If the payoff to skills in the United States exceeds the payoff in the source country, as in Figure 8-5*a,* the threshold S_P increases to S'_P. Fewer workers will move. The workers who do move are the ones above the new threshold S'_P, so that the immigrant flow is still positively selected.

[19] In the U.S. context, the empirical studies have mainly examined the selection characterizing Mexican immigrants, beginning with Daniel Chiquiar and Gordon Hanson, "International Migration, Self-Selection, and the Distribution of Wages: Evidence from Mexico and the United States," *Journal of Political Economy* 113 (April 2005): 239–281. Chiquiar and Hanson find that the probability of emigration to the United States is highest for Mexican workers in the middle of the Mexican skill distribution. More recent work, however, provides conflicting evidence; see Jesús Fernández-Huertas Moraga, "New Evidence on Emigrant Selection," *Review of Economics and Statistics* 93 (February 2011): 72–96; and Robert Kaestner and Ofer Malamud, "Self-Selection and International Migration: New Evidence from Mexico," *Review of Economics and Statistics* 96 (March 2014): 78–91.

Similarly, if the payoff to skills is higher in the source country, as in Figure 8-5*b,* the threshold falls to S'_N. Because only workers who have skill levels below S'_N will move, the drop in U.S. incomes again cuts the number of immigrants. But immigrants are still negatively selected.

Although we have assumed that workers do not incur any moving costs, it is easy to introduce moving costs into the framework. Suppose that it costs, say, $5,000 to move to the United States, *regardless* of the worker's skills. Moving costs then reduce the net income the worker can expect to receive in the United States and would shift down the U.S. wage–skills line. In other words, the introduction of moving costs is equivalent to the reduction in U.S. incomes illustrated in Figure 8-5. If migration costs are constant in the population, therefore, an increase in migration costs reduces the number of immigrants but does not alter the type of selection that characterizes the migrant flow.[20]

8-5 Immigrant Assimilation

How well do movers adapt to the different economic conditions in their new location? This question has received a lot of attention in the context of immigration, as the assimilation of immigrants plays a crucial role in the debate over immigration policy in every immigrant-receiving country. Immigrants who can adapt well and are relatively successful in their new environment can make a significant contribution to economic growth.

Cross-Section Age–Earnings Profiles

To assess the relation between earnings and assimilation, many early studies used *cross-section* data sets (that is, data sets that give a snapshot of the population at a point in time, such as a particular U.S. census).[21] A cross-section data set lets us compare the *current* (as of the time the snapshot is taken) earnings of newly arrived immigrants with the *current* earnings of immigrants who migrated years ago.

The typical analysis estimates variants of the Mincer earnings function in a cross-section. The regressions are given by

$$\text{Native regression: } \log w = \alpha_0 + \alpha_1 s + \alpha_2 t + \alpha_3 t^2 \qquad \textbf{(8-8)}$$

$$\text{Immigrant regression: } \log w = \beta_0 + \beta_1 s + \beta_2 t + \beta_3 t^2 + \beta_4 y + \beta_5 y^2 \qquad \textbf{(8-9)}$$

The native earnings regression is the typical Mincer equation, relating a worker's (log) wage w to years of schooling (s) and years of experience (t). The immigrant equation adds a key variable to the model, the number of years the immigrant has resided in the United States (denoted by y).

It is then easy to use the estimated regression coefficients to graph the implied parabolas. Figure 8-6 illustrates the age–earnings profiles produced by the data from the 1970 census cross-section.

[20] The predictions of the model can change if migration costs vary across workers; see Daniel Chiquiar and Gordon Hanson, "International Migration, Self-Selection, and the Distribution of Wages: Evidence from Mexico and the United States," *Journal of Political Economy* 113 (April 2005): 239–281.

[21] Barry R. Chiswick, "The Effect of Americanization on the Earnings of Foreign-Born Men," *Journal of Political Economy* 86 (October 1978): 897–921.

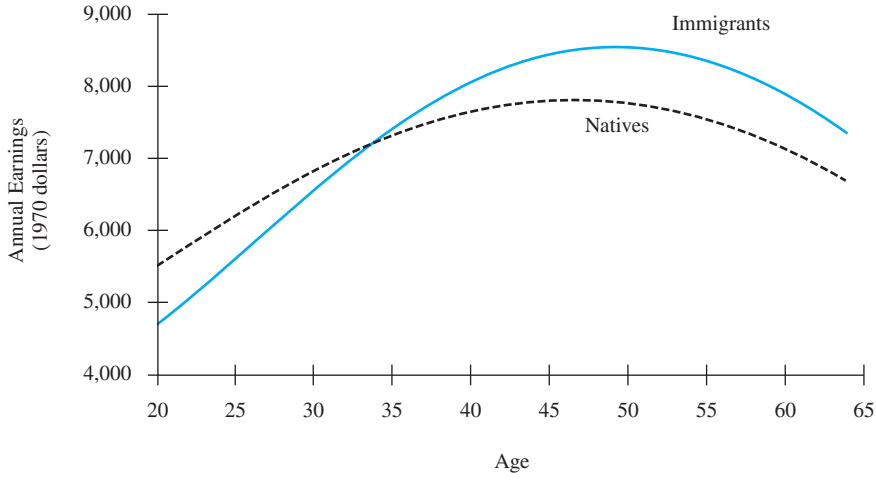


FIGURE 8-6
The Age–Earnings Profiles of Immigrant and Native Men in the Cross Section

Source: Barry R. Chiswick, "The Effect of Americanization on the Earnings of Foreign-Born Men," *Journal of Political Economy* 86 (October 1978): Table 2, Column 3. The figure assumes immigrants enter the country at age 20.

At the time of entry into the United States at age 20, the earnings of immigrant men are about 15 percent lower than the earnings of comparable native men. The age–earnings profile of immigrants, however, is steeper. After 14 years, the earnings of immigrants "overtake" the earnings of native-born workers. The typical immigrant who has been in the United States for 30 years earns about 10 percent more than comparable natives.

There are three notable results in Figure 8-6. First, immigrant earnings are initially below the earnings of natives. This finding is typically interpreted as follows: When immigrants first arrive, they lack skills that are valued by American employers. These "U.S.-specific" skills include language, educational credentials, and information on what the best-paying jobs are and where they are located.

The second is that the immigrant age–earnings profile is steeper. The human capital model implies that greater volumes of human capital investment steepen the age–earnings profile. As immigrants learn English and learn about the U.S. labor market, the immigrants' human capital stock grows relative to that of natives, and economic assimilation occurs in the sense that immigrant earnings begin to converge to the earnings of natives.

The human capital model thus provides a story for why immigrant earnings start out below and then grow faster than the earnings of natives. But it cannot explain the third finding in the figure: After 14 years, immigrants begin to earn more than natives. Why would immigrants end up accumulating more human capital than natives?

The explanation of the overtaking phenomenon typically resorts to a selection argument: Immigrants are not randomly selected from the population. Perhaps only those workers who have exceptional ability, or a lot of drive and motivation, pack up everything they own, leave family and friends behind, and move to a foreign country to start life anew. If immigrants are indeed selected from the population in this manner, it would not be surprising to find that immigrants earn more once they acquire the necessary U.S.-specific skills.

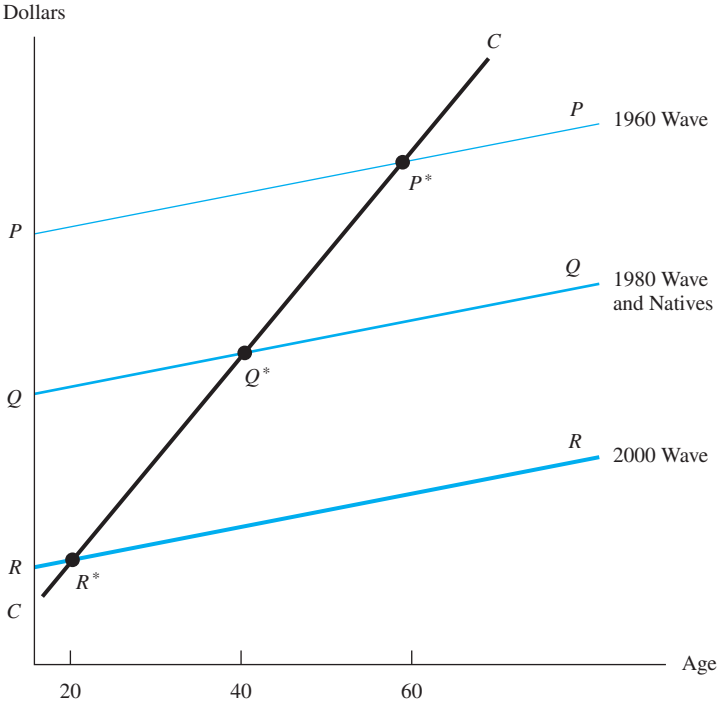
Cohort Effects

The "assimilationist" interpretation of the cross-section age–earnings profiles in Figure 8-6 states that immigrants who migrated many years ago have acquired U.S.-specific skills and thereby improved their economic status. There is one problem with this interpretation: We are drawing inferences about how immigrant earnings evolve over time from a snapshot taken at a single point in time. It might be the case, for example, that newly arrived immigrants are inherently different from those who migrated 20 years ago. It would then be invalid to use the economic experience of those who migrated 20 years ago to forecast the future labor market performance of the newly arrived. Figure 8-7 illustrates this alternative hypothesis.[22]

To simplify, consider a hypothetical situation where there are three separate immigrant waves, and those waves have distinct productivities. One wave arrived in 1960, the second in 1980, and the last in 2000. Suppose also that all immigrants enter the United States at age 20.

FIGURE 8-7 Cohort Effects and the Immigrant Age–Earnings Profile

The typical person migrating in 1960 is skilled and has age–earnings profile *PP;* the 2000 immigrant is unskilled and has age–earnings profile *RR;* the 1980 immigrant has the same skills as the typical native and has age–earnings profile *QQ*. Suppose all immigrants arrive at age 20. The 2000 census cross section reports the wage of immigrants who have just arrived (point R^*); the wage of immigrants who arrived in 1980 when they are 40 years old (point Q^*); and the wage of immigrants who arrived in 1960 when they are 60 years old (point P^*). The cross-sectional age–earnings profile *CC* erroneously suggests that immigrant earnings grow faster than those of natives.



[22] George J. Borjas, "Assimilation, Changes in Cohort Quality, and the Earnings of Immigrants," *Journal of Labor Economics* 3 (October 1985): 463–489.

Let's also assume that the earliest cohort has the highest productivity of any group in the population, including U.S.-born workers. If we could observe the earnings of those immigrants every year after they arrive, their age–earnings profile would be given by the line *PP*. Suppose also that the last cohort of immigrants (that is, the 2000 arrivals) is the least productive of any group in the population. If we could observe their earnings throughout their working lives, their age–earnings profile would be given by the line *RR*. Finally, suppose that the immigrants who arrived in 1980 have the same skills as natives. If we could observe their earnings at every age, the age–earnings profiles of this cohort and of natives would be given by the line *QQ*. Note that the age–earnings profiles of each of the immigrant cohorts is parallel to the age–earnings profile of the native population. There is *no* wage convergence between immigrants and natives in this hypothetical example.

We have access to data drawn from the 2000 decennial census, which reports how old the worker was in 2000 and how much he earned, as well as his country of birth and year of arrival in the United States. This means that the 2000 cross-section tells us the wage of immigrants who have just arrived as part of the 2000 cohort when they are 20 years old (see point R^* in the figure). It also tells us the wage of immigrants who arrived in 1980 when they are 40 years old (point Q^*), and it tells us the wage of immigrants who arrived in 1960 when they are 60 years old (point P^*). A cross-section, therefore, allows us to observe only one point on each of the immigrant age–earnings profiles.

If we connect points P^*, Q^*, and R^*, we trace out the immigrant age–earnings profile produced by cross-section data, or line *CC* in Figure 8-7. This cross-section line is steeper than the native age–earnings profile, making it seem as if there is wage convergence between immigrants and natives, when in fact there is none. Moreover, the cross-section line *CC* crosses the native line at age 40, giving the appearance that immigrant earnings overtake those of natives after 20 years. In fact, no immigrant group experienced such an overtaking.

Figure 8-5 shows that cross-section age–earnings profiles can yield an erroneous perception about assimilation if there are intrinsic differences in productivity across immigrant cohorts. The skill differences across cohorts are called **cohort effects.**[23]

Evidence on Cohort Effects

The data indeed suggest that there are skill differences across immigrant cohorts and that these cohort effects may be sizable.[24] The typical study "tracks" the earnings of a specific immigrant cohort across censuses. For instance, the 1980 census reports the average wage of persons who migrated in 1980 when they are 25 years old; the 1990 census reports the average wage of the same immigrants when they are 35 years old; and the 2000 census reports the average wage for the same persons when they are 45 years old. The tracking of specific immigrant cohorts across censuses then traces out the age–earnings profile for each of the cohorts.

[23] The discussion of Figure 8-7 assumed that the cohort effects arise because more recent immigrant waves are less skilled than earlier waves. Cohort effects may also arise because of nonrandom return migration by immigrants, so that earlier waves have been "filtered." See Darren Lubotsky, "Chutes or Ladders? A Longitudinal Analysis of Immigrant Earnings," *Journal of Political Economy* 115 (October 2007, no. 5): 820–867.

[24] The evidence is surveyed in George J. Borjas, *Immigration Economics,* Cambridge, MA: Harvard University Press, 2014.

FIGURE 8-8 Evolution of Wages for Specific Immigrant Cohorts over the Life Cycle (Relative to Wages of Comparably Aged Native Men)

Source: Francine D. Blau and Christopher Mackie, editors, *The Economic and Fiscal Consequences of Immigration,* Washington, DC: National Academies Press, 2016, p. 110.

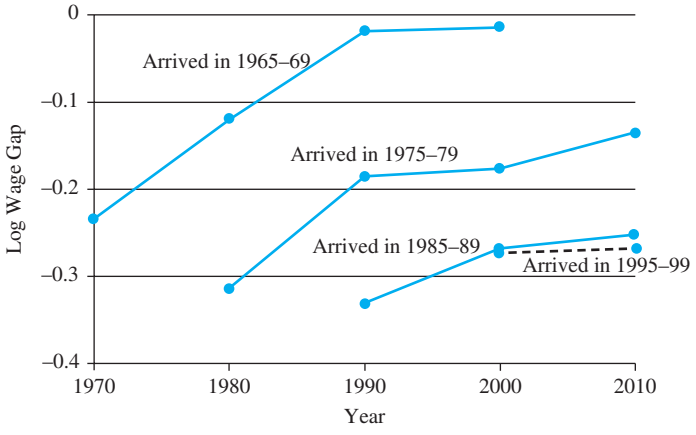


Figure 8-8 illustrates the age–earnings profiles produced by such tracking, as reported in a 2016 report by the National Academy of Sciences. Newly arrived immigrant men in 1970 earned about 24 percent less than natives at the time of entry. By 1990, the entry wage of the newest wave was 33 percent less than that of natives. In other words, there seems to have been a decline in immigrant skills across successive waves through 1990. The figure also shows that the earnings of immigrants who arrived in the late 1960s caught up with the earnings of native workers within two decades. The immigrant cohort that arrived in the late 1980s, however, started out at a greater disadvantage and experienced less wage growth.

8-6 The Job Match and Job Turnover

We now turn to the last type of labor mobility observed frequently in many labor markets: job turnover—the flow of workers from one job to another.

In the standard model of a competitive labor market, the interaction of workers looking for the best jobs and employers maximizing profits equalizes the value of marginal product of labor across firms. The equilibrium allocation of workers to firms maximizes the value of labor's contribution to national income. A worker's value of marginal product would not increase if he or she were to switch to another firm, so there is no reason for *any* type of job separation to occur.

Nevertheless, both quits and layoffs are observed frequently. These job separations occur partly because workers differ in their abilities and because firms offer different working conditions. Moreover, workers lack information about which firm provides the best opportunities, and firms lack information about the worker's true productivity.[25]

[25] Boyan Jovanovic, "Job Matching and the Theory of Turnover," *Journal of Political Economy* 87 (October 1979): 972–990; and Derek Neal, "The Complexity of Job Mobility among Young Men," *Journal of Labor Economics* 17 (April 1999): 237–261.

Suppose that different firms offer different work environments. At Microsoft, for example, the supervisor is always organized, plans the worker's schedule well in advance, and gives the worker ample time to complete an assigned task. At Joe's Start-Up, Joe waits until the last minute to inform the worker of an upcoming task (such as writing code for the latest tweak in a video game), and then imposes a tight deadline. If a particular worker does not perform well under stressful conditions, the value of the match between that worker and Microsoft will be higher than the value of the match at Joe's. Other workers, however, might find that their productive juices flow best when faced with tight deadlines, and the value of the match at Joe's would be higher.

The notion that each **job match** (that is, each particular pairing of a firm and a worker) has its own unique value implies that both workers and firms can improve their situations by shopping around. It matters if a particular programmer is employed at Microsoft or at Joe's. The joint search by both workers and firms for the best match increases the worker's wage and the firm's profits.

If workers and firms knew exactly which particular match had the highest value, workers would look for the best firm, firms would look for the best worker, and there would be no need for any turnover once the "marriage" was consummated. The sorting of workers and firms would be the optimal sorting, the one that maximizes the total value of labor's product.

But both firms and workers are ill-informed about the true value of the match when the job begins. Over time, both the worker and the firm may realize that they incorrectly predicted that value. Moreover, they both know that there are other workers and firms out there that would lead to a better match. Job turnover is the mechanism that labor markets use to correct matching errors and moves the market to a better allocation of resources. This type of turnover is called **efficient turnover,** for it increases the total value of labor's product in a competitive market.

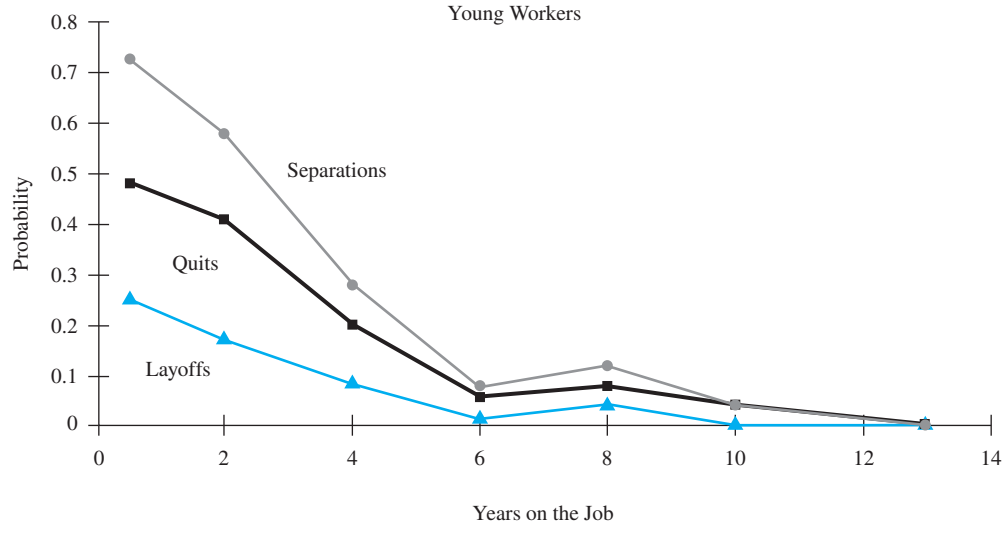
As Figure 8-9 shows, newly hired workers indeed have a very high probability of changing jobs. Nearly, 75 percent of newly hired young workers and about 40 percent of newly hired older workers change jobs within 2 years. The separation rate, however, declines sharply once the worker has been on the job even for 2 or 3 years. Note that the pattern is observed both among quits (or employee-initiated separations) and layoffs (or employer-initiated separations). Newly hired workers likely have the highest quit and layoff rates because both workers and firms are "testing the waters" and finding out the true value of the job match.

Interestingly, the figure suggests that separation rates continue to decline even after the worker has been on the job 2 or 3 years. As noted in the chapter on the earnings distribution, the steady decline in quit and layoff rates with job seniority is consistent with the presence of firm-specific human capital.[26] At the beginning of an employment relationship, the worker and firm have not yet invested in skills that are specific to that job, and there is no "bond" between the two parties. Once specific training is acquired, the worker's productivity in the firm that provided the training exceeds his wage (lowering the probability of layoff) and the worker's wage in that firm exceeds the wage he can get elsewhere (lowering the probability of a quit).

[26] Henry S. Farber, "Mobility and Stability: The Dynamics of Job Change in Labor Markets," in Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics,* vol. 3B, Amsterdam: Elsevier, 1999, pp. 2440–2483; and Lalith Munasinghe, "Specific Training Sometimes Cuts Wages and Always Cuts Turnover," *Journal of Labor Economics* 23 (April 2005): 213–233.

FIGURE 8-9 Probability of Job Turnover for Young and Older Workers

Source: Jacob Mincer and Boyan Jovanovic, "Labor Mobility and Wages," in Sherwin Rosen, editor, *Studies in Labor Markets,* Chicago: University of Chicago Press, 1981, p. 25. The sample of young workers refers to men aged 19–29, while the sample of older workers refers to men aged 50–64.



In the end, long jobs tend to become the norm rather than the exception. For example, 17 percent of workers aged 45–59 are in jobs that have lasted at least 20 years, and an additional 28 percent are in jobs that have lasted 10–19 years.[27]

[27] U.S. Department of Labor, Bureau of Labor Statistics, *Employee Tenure in 2016*, Washington, DC: September 2016. Table 3.

Theory at Work

HEALTH INSURANCE AND JOB LOCK

The optimal allocation of workers to firms requires that workers sort to those jobs where they are most productive. A number of factors, however, may block workers from moving to better jobs and prevent the economy from attaining an efficient allocation of labor.

A worker's employer-provided health insurance is generally not portable across jobs in the United States. Prior to the enactment of Obamacare, many health insurance programs refused to cover a new worker's preexisting medical conditions (sometimes for up to 2 years). As a result, workers who had a health problem would not want to move to a job where they were more productive because of the potential cost associated with losing health insurance coverage. In fact, 30 percent of the respondents in a CBS/*New York Times* Poll reported that they had stayed at a job they wanted to leave mainly because they did not want to lose their health insurance. The employer-based health insurance system, therefore, creates a form of "job-lock," where workers are locked into their jobs even though this allocation of workers to firms might not be efficient.

There is evidence suggesting that this type of job-lock was a significant problem in the U.S. labor market. In particular, families where a wife is pregnant (a form of preexisting medical condition) show increased mobility if the workers have no health insurance, but reduced mobility if workers have employer-provided health insurance. Job-lock may have reduced the quit rate of workers with employer-provided health insurance by as much as 25 percent per year.

Sources: Brigitte C. Madrian, "Employment-Based Health Insurance and Job Mobility: Is There Evidence of Job-Lock?" *Quarterly Journal of Economics* 109 (February 1994): 27–54; and Mark C. Berger, Dan A. Black, and Frank A. Scott, "Is There Job Lock? Evidence from the Pre-HIPAA Era," *Southern Economic Journal* 70 (April 2004): 953–976.

8-7 Job Turnover and the Age–Earnings Profile

Job turnover changes the shape of the worker's age–earnings profile. Most obviously, quitters usually move on to higher-paying jobs, while workers who are laid off move on to lower-paying jobs.[28] In fact, the adverse consequences of losing a job involuntarily can be substantial. A study of displaced workers in the United Kingdom found that the subsequent wage of workers who lost their jobs because of a mass layoff is about 15–25 percent lower than the prelayoff wage.[29]

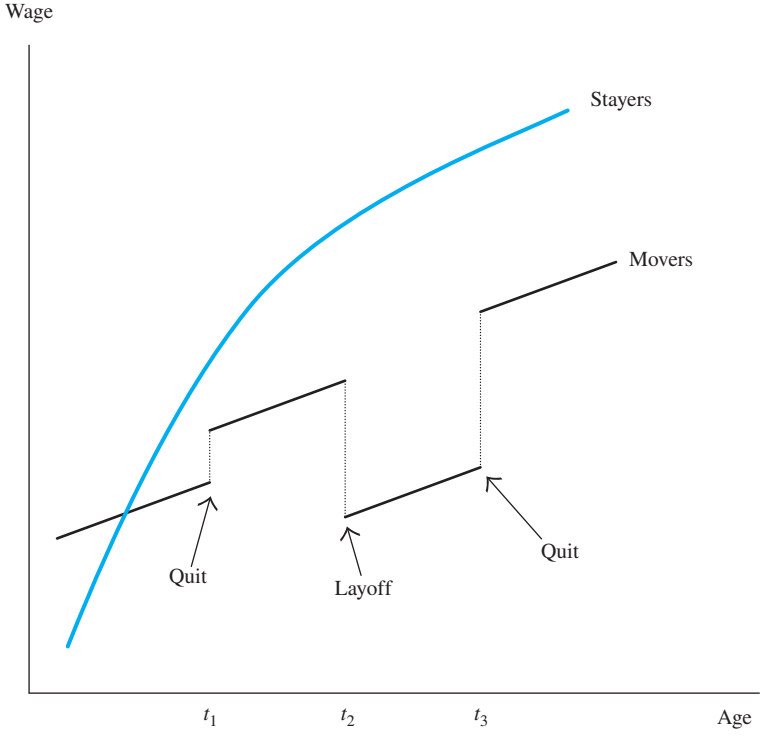
Job turnover, therefore, causes an immediate shift on the *level* of the mover's age–earnings profile, as illustrated in Figure 8-10. The wage level shifts up at ages t_1 and t_3, when the worker quits his job, and shifts down at age t_2 when he is laid off.

But labor turnover can also affect the *slope* of the age–earnings profile. Figure 8-10 also contrasts the age–earnings profile of two workers, a mover and a stayer. The stayer has a continuous profile that is steep, so that the rate of wage growth *within the job* is substantial. The mover gets either a raise or a wage cut with each job change (depending on whether it's a quit or a layoff). Within a given job, however, the mover's age–earnings profile is relatively flat.

[28] Jacob Mincer, "Wage Changes and Job Changes," *Research in Labor Economics* 8 (1986, Part A): 171–197.

[29] Alexander Hijzen, Richard Upward, and Peter W. Wright, "The Income Losses of Displaced Workers," *Journal of Human Resources* 45 (Winter 2010): 243–269.

FIGURE 8-10 Job Turnover and the Age–Earnings Profile
The age–earnings profile of movers is discontinuous, shifting up when they quit and shifting down when they are laid off. Long jobs encourage firms and workers to invest in specific training and steepen the age–earnings profile within the job.



Firm-specific training, in fact, suggests this relation between job turnover and the slope of the age–earnings profile within a job. Workers and firms engaged in a long-term relationship have incentives to invest in specific skills. Because workers pay part of the costs and collect part of the returns, wage growth is steeper in the jobs that have large specific capital investments—namely, longer jobs. A worker's earnings, therefore, depend not only on his years of labor market experience, as in the Mincer earnings function, but also on his job history and on job tenure, the number of years employed in the current job.

There is evidence that workers who have long job tenure earn more than newly hired workers, even after controlling for differences in the worker's age.[30] Although this evidence seems consistent with the specific training hypothesis, there is some skepticism about whether the positive correlation between job tenure and earnings truly reflects the influence of specific training.

The source of the problem is that the positive correlation between earnings and job tenure *across workers* can arise in a very different way. Suppose that some workers got lucky

[30] Michael R. Ransom, "Seniority and Monopsony in the Academic Labor Market," *American Economic Review* 83 (March 1993): 221–233.

and found high-paying jobs. These workers are in good matches and earn w_{HIGH} per year as long as they remain in their jobs. Suppose that the earnings of a well-matched worker do not grow over time.

Other workers are not as lucky. They are badly matched and have low earnings w_{LOW} per year as long as they remain in their bad jobs. Suppose also that the earnings of a poorly matched worker do not grow over time. In this hypothetical example, therefore, job tenure has no impact on earnings. Put differently, specific training plays no role in determining wages.

The lucky workers who earn w_{HIGH} are happy with their current situation and feel no need to look for alternative employment. Workers in good matches, therefore, will have low probabilities of job separation and high values of job tenure. But the workers who are not well matched are unhappy with their current situation, and will have high probabilities of job turnover and little seniority.

The correlation between earnings and job tenure across workers will be positive, suggesting that wages grow with job tenure when no such thing actually happens. For a given worker, wages do not grow with tenure. Across workers, however, longer job tenure is correlated with higher wages because workers with a lot of job seniority are likely to be in good matches, and workers with little seniority are in bad matches. It would be incorrect, therefore, to conclude that the cross-sectional correlation says anything about the importance of specific training.

To isolate the impact of job tenure on a given worker's wage, we would need to track a worker's earnings over time both as he gets older and as he accumulates firm-specific experience. However, the evidence on the relationship between wages and job tenure obtained by the tracking studies is mixed.[31] The most recent studies, however, suggest that first 10 years of job seniority may increase a worker's earnings by about 10 percent more than he could earn elsewhere. Put differently, each year of job tenure expands the worker's earnings opportunities by about 1 percent.

Summary

- The probability of moving across geographic regions depends on economic conditions in both the origin and destination regions, and on moving costs. A geographic move is more likely when incomes are low at the origin or when incomes are high at the destination. A move is also more likely when moving costs are low.
- If migration decisions are made jointly by all household members, the migrant flow will likely include a number of tied movers. Tied movers suffer a private loss from the move, but their loss is more than outweighed by the gains of other family members.

[31] See Katherine G. Abraham and Henry S. Farber, "Job Duration, Seniority, and Earnings," *American Economic Review* 77 (June 1987): 278–297; Joseph G. Altonji and Robert A. Shakotko, "Do Wages Rise with Job Seniority?" *Review of Economic Studies* 54 (July 1987): 437–459; Robert H. Topel, "Specific Capital, Mobility, and Wages: Wages Rise with Job Seniority," *Journal of Political Economy* 99 (February 1991): 145–176; Joseph G. Altonji and Nicolas Williams, "The Effects of Labor Market Experience, Job Seniority, and Mobility on Wage Growth," *Research in Labor Economics* 17 (1998): 233–276; and Margaret Stevens, "Earnings Functions, Specific Human Capital, and Job Matching: Tenure Bias Is Negative," *Journal of Labor Economics* 21 (October 2003): 783–806.

- If migration decisions are made jointly by all household members, some workers who should have moved will not, and become tied stayers. The tied stayer's private gain from moving is smaller than the family's loss.
- Migrants are not randomly chosen from the population. If the rate of return to skills in the destination exceeds the rate of return in the origin, the migrant flow is positively selected, and migrants have above-average skills. If the rate of return in the destination is lower than the rate of return in the origin, the immigrant flow is negatively selected, and immigrants have below-average skills.
- If there are cohort effects in the skill composition of the immigrant population, the observation that earlier immigrants earn more than newly arrived immigrants in a cross section need not indicate that immigrants experience significant assimilation. The correlation may instead reflect differences in productivity across immigrant waves.
- Efficient turnover improves the quality of the job match between worker and firm and increases labor's contribution to national income.
- Workers who have been on the job for a long time earn more than newly hired workers. This correlation arises partly because workers in good matches tend to stay on the job longer and because the accumulation of specific training increases the worker's productivity.

Key Concepts

cohort effects, *287*
efficient turnover, *289*
job match, *289*
labor mobility, *271*
negative selection, *282*
positive selection, *281*
repeat migration, *274*
return migration, *274*
Roy model, *280*
tied mover, *278*
tied stayer, *278*

Review Questions

1. Show how workers who wish to maximize the present value of lifetime earnings calculate the net gains to migration. And discuss how the net gain depends on incomes in the states of origin and destination and on moving costs.
2. Show how one can use the human capital framework to obtain an estimate of migration costs for the marginal person.
3. Why is there a difference between the private gains to migration and the family's gains to migration? Discuss how this difference generates tied stayers and tied movers. Can both the husband and the wife be tied movers?
4. Describe how the immigrant flow is chosen from the population of the country of origin. Why are some immigrant flows positively selected and other flows negatively selected?
5. Show how cohort effects in the immigrant population affect the interpretation of the cross-sectional age–earnings profiles of immigrants.
6. How do quits and layoffs help improve labor market efficiency?
7. How should one interpret the fact that—all other things equal—workers with a lot of seniority earn more than newly hired workers?

Problems

8-1. Suppose a worker with an annual discount rate of 10 percent currently resides in Pennsylvania and is deciding whether to remain there or to move to Illinois. There are three work periods left in the life cycle. If the worker remains in Pennsylvania, he will earn $20,000 in each of the three periods. If the worker moves to Illinois, he will earn $22,000 in each of the three periods. What is the highest cost of migration that a worker is willing to incur and still make the move?

8-2. Suppose high-wage workers are more likely than low-wage workers to move to a new state for a better job.

(a) Explain how this migration pattern can be due solely to differences in the distribution of wages.

(b) Explain how this migration pattern can take place even if the cost to moving is greater for high-wage workers.

8-3. Patrick and Rachel live in Seattle. Patrick's net present value of lifetime earnings in Seattle is $125,000, while Rachel's is $500,000. The cost of moving to Atlanta is $25,000 *per person.* In Atlanta, Patrick's net present value of lifetime earnings would be $155,000, while Rachel's would be $510,000. If Patrick and Rachel choose where to live based on their joint well-being, will they move to Atlanta? Is Patrick a tied-mover or a tied-stayer or neither? Is Rachel a tied-mover or a tied-stayer or neither?

8-4. Consider a household consisting of four college friends. The friends have made a commitment to live together for the next five years. Presently they live in Milwaukee where Abby will earn $200,000, Bonnie will earn $120,000, Cathy will earn $315,000, and Donna will earn $150,000 over the next five years. They have the option of moving to Miami. Moving to Miami would impose a one-time moving cost of $5,000 on each person. If they move to Miami, however, Abby will earn $180,000, Bonnie will earn $150,000, Cathy will earn $300,000, and Donna will earn $100,000 over the next five years. Moreover, each friend prefers to live in Miami over Milwaukee. In particular, Abby and Bonnie both value the quality of life in Miami versus Milwaukee over the next five years at $40,000 while Cathy and Donna place the value at $25,000 each. Should the household move to Miami or stay in Milwaukee? Is anyone a tied-mover or a tied stayer?

8-5. Suppose the United States enacts legislation granting all workers, including newly arrived immigrants, a minimum income floor of y dollars. (Assume there is positive selection of migrants from the home country to the U.S. before the policy change.)

(a) Generalize the Roy model to show how this type of welfare program influences the incentive to migrate to the United States. Ignore any issues regarding how the welfare program is funded.

(b) Does this welfare program change the selection of the immigrant flow? In particular, are immigrants more likely to be negatively selected than in the absence of a welfare program?

(c) Which types of workers, the highly skilled or the less skilled, are most likely to be attracted by the welfare program?

8-6. In the absence of any legal barriers on immigration from Neolandia to the United States, the economic conditions in the two countries generate an immigrant flow

that is negatively selected. In response, the United States enacts an immigration policy that restricts entry to Neolandians who are in the top 10 percent of Neolandia's skill distribution. What type of Neolandian would now migrate to the United States?

8-7. One trend in the U.S. labor market in the 2100s is telecommuting or working at home. More and more firms allow working from home, and many firms even allow employees to live and work in one city for most of the year, flying to the firm's headquarters for 3 or 4 days of work every quarter. How is this trend likely to affect job mobility (that is, workers switching jobs)? How is this trend likely to affect internal migration raters in the U.S. (that is, households moving cities)?

8-8. In addition to it being illegal to enter the U.S. without a visa or to over-stay one's visa, it is also illegal for U.S. employers to hire undocumented or "illegal" immigrants. Meanwhile, federal U.S. enforcement of immigration laws tends to concentrate resources on reducing illegal immigration rather than on prosecuting U.S. firms for employing undocumented workers. Using supply and demand analysis, show what would happen to the wage and employment level of undocumented workers if the government pursued more active enforcement of employers. According to your model, what would happen to the wage and employment level of documented workers?

8-9. Under 2001 tax legislation enacted in the United States, all income tax filers became eligible to deduct from their total income half of their expenses incurred when moving more than 50 miles to accept a new job. Prior to the change, only tax filers who itemized their deductions were allowed to deduct their moving expenses. (Typically, homeowners itemize their deductions and renters do not itemize.) How would this change in tax policy likely affect the mobility of homeowners and renters?

8-10. Suppose the immigrant flow from Lowland to Highland is positively selected. In order to mitigate the "brain drain" Lowland experiences as a result of this migration, public officials of Lowland successfully convince all Lowlanders who migrate to Highland to remit 10 percent of their wages to family members.

(a) What effect will this policy have on the immigrant flow?

(b) Provide a graph that details the extent to which this policy will limit the brain drain.

8-11. (a) According to standard migration theory, how will skill selection (positive versus negative) change on average as the distance between the source country and the destination country increases?

(b) The 1990 U.S. Census data can be used to estimate the average wage differential between immigrants to the U.S. by country of origin, and compare those to the average native wage of workers with similar characterizes such as education, age, occupation, etc. The data suggest that the average Canadian immigrant earns about 25 percent more than Americans while the average Mexican immigrant earns about 40 percent less. Similarly, Indian immigrants earn about 12 percent more than Americans while Vietnamese immigrants earn almost 20 percent less. Do these

empirical results support the idea that skill selection is a monotonic function of the distance between countries? If not, what might explain the differences?

8-12. (a) Explain how a universal healthcare system would likely cause a greater amount of efficient turnover.

(b) Defined-benefit retirement plans promise a fixed amount of retirement income to workers, but in order to receive benefits workers must be vested in the plan which usually requires working at the firm for 10 or 15 years. In contrast, a defined-contribution retirement plan specifies a fixed amount of money the firm contributes each pay period to a worker's retirement fund which the worker then largely controls and can access even if she changes jobs. Do defined-benefit or defined-contribution retirement plans allow for more efficient turnover?

(c) When federal workers in Washington D.C. move jobs from one federal agency to another, the worker keeps her same health insurance and retirement benefits. In order to quantify the degree to which ease of transfer of benefits affects job sorting, two groups of new economists Ph.Ds. who accept a job in Washington D.C. are observed. The first group is U.S. citizens. The second group is non-U.S. residents who eventually received permanent resident status after 3 years of work experience. By law, several government agencies cannot hire nonresidents. Among the group of U.S. citizens, 42 percent changed jobs within the first 3 years of work while 33 percent changed jobs during their fourth to sixth years of work. Among the group of non U.S. residents, 17 percent changed jobs in the 3 years before becoming a resident while 29 percent changed jobs in the 3 years after becoming a U.S. resident. Provide a difference-in-differences estimate of the effect of being a U.S. resident/citizen in Washington D.C. for Ph.D. economists on job sorting.

8-13. The Immigration Reform Act of 2006 provided fewer work visas than were available in previous years for college graduates to remain in the United States. The exception is that work visas remained plentiful for college graduates who majored in technical areas such as math, computer programming, and physics.

(a) How will this policy likely affect the skill distribution of immigrants to the United States and the age-earnings profile of immigrants in the United States?

(b) In the future a demographer uses the 2010 U.S. census to study immigrant wages and concludes that the U.S. policy actually had the unintended consequence of attracting immigrants with lower levels of productivity as shown by a flatter age-earnings profile. Using a graph similar to Figure 8-7, show why the demographer's conclusions are sensitive to cohort effects.

8-14. KAPC, a pharmaceutical company located in rural Kansas, is finding it difficult to retain its employees who frequently leave after just six months for jobs at pharmaceutical companies paying higher wages in Chicago. To address its problem with labor turnover, human resource officers at KAPC decide to run an experiment. Of their next 100 newly hired employees, 25 will randomly be selected to receive a housing voucher worth up to $4,000 per year to offset property taxes. To take advantage of this program, the employee must not only be randomly selected into the program but she must also purchase a home. Of the 25 employees selected

into the housing voucher program, 7 leave KAPC within 12 months of starting. Of the 75 employees not selected into the program, 37 leave KAPC within 12 months of starting.

(a) Provide an estimate of the effect the housing voucher program has on retention at KAPC.

(b) Suppose KAPC spends $10,000 in hiring costs each time a position is vacated. Would you endorse expanding the housing voucher program to all new employees? Justify your decision.

8-15. Consider the Roy model of potential immigrant flows as discussed in the chapter.

(a) Why is it that a source country can experience both an outflow of low-skill workers and an outflow of high-skill workers at the same time?

(b) Provide a graph of the returns to skills in the destination and source countries that would suggest both behaviors occur simultaneously.

(c) How do the social and economic (that is, tax) policies of the United States encourage both types of flows?

Selected Readings

Ran Abramitzky, Leah P. Boustan, and Katherine Eriksson, "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration," *American Economic Review* 102 (August 2012): 1832–1856.

George J. Borjas, "Assimilation, Changes in Cohort Quality, and the Earnings of Immigrants," *Journal of Labor Economics* 3 (October 1985): 463–489.

George J. Borjas, "Self-Selection and the Earnings of Immigrants," *American Economic Review* 77 (September 1987): 531–553.

Barry R. Chiswick, "The Effect of Americanization on the Earnings of Foreign-Born Men," *Journal of Political Economy* 86 (October 1978): 897–921.

Dora L. Costa and Matthew E. Kahn, "Power Couples: Changes in the Locational Choice of the College Educated, 1940–1990," *Quarterly Journal of Economics* 115 (November 2000): 1287–1314.

Brigitte C. Madrian, "Employment-Based Health Insurance and Job Mobility: Is There Evidence of Job-Lock?" *Quarterly Journal of Economics* 109 (February 1994): 27–54.

Robert H. Topel, "Specific Capital, Mobility, and Wages: Wages Rise with Job Seniority," *Journal of Political Economy* 99 (February 1991): 145–176.

Fabian Waldinger, "Quality Matters: The Expulsion of Professors and the Consequences for Ph.D. Student Outcomes in Nazi Germany," *Journal of Political Economy* 118 (August 2010): 787–831.

Chapter 9

Labor Market Discrimination

The way to stop discrimination on the basis of race is to stop discriminating on the basis of race.

—Chief Justice John Roberts

In other chapters, we have analyzed how differences in the work environment or in the skills of workers generate wage dispersion in competitive labor markets. We now demonstrate that wage differences may arise even among equally skilled workers employed in the same job simply because of the worker's race, gender, ethnicity, sexual orientation, or other seemingly irrelevant characteristics.

These differences are attributed to discrimination. Discrimination occurs when labor market participants take into account such factors as race and gender when making economic transactions. For instance, employers might care about the gender of their workers they hire; workers might care about the race of their coworkers; and customers might care about the race and gender of the seller. Although economists have little to say about the psychological roots of prejudice, we can easily reinterpret this type of behavior in the language of economics: The costs and benefits of an economic exchange depend on the race and gender of the persons involved in the exchange.

In fact, racial and gender differences in labor market outcomes might arise even if market participants are not prejudiced. We often "read" a person's socioeconomic background to learn more about that person's productivity and skills. For instance, we all know that teenagers are more likely to engage in reckless driving. Surely this information is useful to companies selling auto insurance. Similarly, employers, workers, and customers will use race, gender, and any other relevant traits to fill in information gaps about participants in the marketplace.

The chapter also shows how economists measure labor market discrimination and discusses the long-run trends in the black–white and male–female wage differentials. The study of these long-run trends provides important insights into the impact of government policies, such as affirmative action, designed to improve the relative economic status of minorities and women.

9-1 Race and Gender in the Labor Market

Table 9-1 reports various measures of human capital and labor market outcomes, by race and gender. Perhaps most striking are the large differences in annual earnings. Men earn more than women and whites typically earn more than nonwhites, although Asian men

TABLE 9-1 Gender and Racial Differences in Skills and Labor Market Outcomes, 2016–2017

Sources: The data on educational attainment refer to persons aged 25 and over and are drawn from U.S. Census Bureau, "Table A-2. Percent of People 25 Years and Oer Who Have Completed High School or College, by Race, Hispanic Origin, and Sex," and are available online at https://www.census.gov/data/tables/time-series/demo/educational-attainment/cps-historical-time-series.html. The data on labor force participation and unemployment rates refer to persons aged 20 and over and are drawn from U.S. Bureau of Labor Statistics, "Table A-2. Employment Status of the Civilian Population by Race, Sex, and Age," available online at www.bls.gov/cps/cpsatabs.htm. The data for Asians refer to persons aged 16 and over. The data on earnings refers to workers aged 25 and over and are drawn from U.S. Census Bureau, "Table PINC-03, Educational Attainment—People 25 Years Old and Over, by Total Money Earnings in 2009, Work Experience in 2009, Age, Race, Hispanic Origin and Sex," available online at https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-pinc/pinc-03.html.

	White		Black		Hispanic		Asian	
	Male	Female	Male	Female	Male	Female	Male	Female
Percent high school graduate or more	89.5	90.6	86.5	87.9	69.5	71.6	92.6	89.4
Percent bachelor's degree or more	34.0	35.0	22.1	25.4	15.8	18.6	56.6	53.2
Labor force participation rate	71.8	57.6	68.1	62.5	80.1	58.4	72.1	55.5
Unemployment rate	3.5	3.5	7.2	6.5	4.2	5.0	4.3	4.5
Annual earnings (in $1,000)	67.7	45.6	50.5	38.7	47.7	33.2	80.5	53.8
Annual earnings (among workers employed full-time, year-round) (in $1,000)	75.8	56.4	58.1	47.0	53.1	41.1	88.0	64.4

have the highest annual earnings of any group ($80,500). In contrast, white women earn $45,600, black men earn $50,500, and Hispanic women earn $33,200.

The data suggest that these earnings gaps arise partly because of differences in labor supply among the various groups. For example, the typical white man earns 48 percent more than the typical white woman ($67,700 versus $45,600). But the typical white man employed full-time earns "only" 34 percent more than a white woman employed full time (or $75,800 versus $56,400).

Part of the earnings gaps can also be attributed to differences in educational attainment. Only about 11 percent of white men do not have a high school diploma, as compared to 14 percent of black men and almost 30 percent of Hispanic men. Similarly, 34 percent of white men are college graduates, as compared to 22 percent of black men, and 16 percent of Hispanic men. If the rate of return to schooling is around 9 percent, these differences in educational attainment would clearly generate substantial wage differentials.

It is important to note that sizable race and gender wage differentials are not exclusive features of the U.S. labor market. In Malaysia, the Malay/Chinese wage ratio is 0.57 and the Indian/Chinese wage ratio is 0.81. Black men in Canada earn 18 percent less than Canadian whites; nonwhite immigrants in Britain earn 10–20 percent less than similarly skilled white immigrants; Jews of Oriental-Sephardic background in Israel earn less than Jews of Ashkenazic (that is, European) background; and there are substantial wage gaps among the various castes that make up Indian society.[1] Finally, there is a sizable wage gap

[1] William Darity Jr. and Jessica Gordon Nembhard, "Racial and Ethnic Economic Inequality: The International Record," *American Economic Review* 90 (May 2000): 308–311; Juliet Howland and Christos Sakellariou, "Wage Discrimination, Occupational Segregation and Visible Minorities in Canada," *Applied Economics* 25 (November 1993): 1413–1422; Biswajit Banerjee and J. B. Knight, "Caste Discrimination in the Indian Labour Market," *Journal of Development Economics* 17 (April 1985): 277–307.

between men and women in most countries. Men earn 27 percent more than women in Britain, 23 percent more in Germany, 18 percent more in Ireland, and 25 percent more in the Netherlands.[2]

9-2 The Discrimination Coefficient

The birth of the economic analysis of discrimination can be traced back to the 1957 publication of Nobel Laureate Gary Becker's doctoral dissertation entitled *The Economics of Discrimination.*[3] Much of the subsequent literature on discrimination is motivated and guided by the analytical framework set out in that influential study.

Becker's theory is based on the concept of **taste discrimination.** This concept translates the common-sense notion of racial prejudice into the language of economics.

Suppose there are two types of workers in the labor market: white workers and black workers. A competitive employer faces constant prices for these inputs; w_W is the wage rate for a white worker and w_B is the wage rate for a black worker. If the employer is prejudiced against blacks, the employer gets disutility from hiring black workers. In other words, even though it costs only w_B dollars to hire a black worker, the employer will *act as if* it costs $w_B(1 + d)$ dollars, where d is a positive number and is called the **discrimination coefficient.**

Racial prejudice blinds the employer to the actual monetary cost of the transaction. As a result, the employer's perceived cost of hiring a black worker exceeds the actual cost. Suppose that $w_B = \$10$ per hour and that $d = 0.5$. The employer will then act as if hiring a black worker costs \$15 per hour, a 50 percent increase in cost. The discrimination coefficient d, therefore, gives the percent "markup" in the cost of hiring a black worker attributable to the employer's prejudice. The greater the prejudice, the greater the disutility from hiring blacks, and the greater is the discrimination coefficient d.

Some employers might prefer to hire blacks. This type of behavior, called **nepotism,** implies that an employer's utility-adjusted cost of hiring a favored worker equals $w_B(1 - n)$ dollars, where the "nepotism coefficient" n is a positive number. For example, suppose black employers prefer to hire black workers. Black employers will then act as if hiring a black worker is cheaper than it actually is.

It is easy to apply Becker's definition of taste discrimination to other types of economic interactions. White workers, for instance, might dislike working alongside black workers. If a prejudiced white worker's wage equals w_W, she will act as if her wage

[2] Claudia Olivetti and Barbara Petrongolo, "Unequal Pay or Unequal Employment? A Cross-Country Analysis of Gender Gaps," *Journal of Labor Economics* 26 (October 2008): 621–654.

[3] Gary S. Becker, *The Economics of Discrimination,* 2nd ed., Chicago: University of Chicago Press, 1971 (1957). The voluminous literature inspired by Becker's framework is surveyed by Joseph G. Altonji and Rebecca M. Blank, "Race and Gender in the Labor Market," in Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics,* vol. 3C, Amsterdam: Elsevier, 1999, pp. 3143–1259. A recent attempt to empirically test aspects of Becker's theory is given by Kerwin Kofi Charles and Jonathan Guryan, "Prejudice and Wages: An Empirical Assessment of Becker's The Economics of Discrimination," *Journal of Political Economy* 116 (October 2008): 773–809.

Theory at Work

BEAUTY AND THE BEAST

There is a great deal of diversity in what we consider "beautiful" across cultures and over time. Ugangi men, for instance, are attracted to women with distended lower lips; European men in the eighteenth and nineteenth centuries fantasized about the plump women immortalized by Rubens; and today's Western men prefer lean women who would have been considered ill and undernourished 200 years ago.

Nevertheless, our attitudes about what defines a beautiful person at a particular point in time seem to have a strong impact on the labor market outcomes experienced by the beautiful and the ugly. As Jade Jagger (Mick's daughter) puts it: "God, what gorgeous staff I have. I just can't understand those who have ugly people working for them . . . Just call me a pathetic aesthetic."

There exist wage differentials not only on the basis of race and gender, but also on a worker's ranking in the beauty scale. American men who are perceived as having above-average looks earn 4 percent more than the average man, and beautiful women earn 8 percent more than the average woman. It seems as if the "look" of the workers enters the utility function of employers, so they are willing to pay a premium to be associated with "the beautiful people" and to penalize workers whose appearance they dislike.

In addition to the tabloid appeal of these results, the evidence may have substantial policy implications. The Americans with Disabilities Act of 1990 prohibits discrimination on the basis of physical disabilities. There already exist court precedents establishing that ugliness might be considered a physical disability. In 1992, the Vermont Supreme Court ruled that the lack of upper teeth is a disability protected by the state's Fair Employment Opportunities Act. Discrimination against ugly people, therefore, might already be a violation of the law. We still do not know, however, how many workers are willing to be certified ugly by a jury of their peers in order to get a raise.

Sources: Daniel S. Hamermesh and Jeff E. Biddle, "Beauty and the Labor Market," *American Economic Review* 84 (December 1994): 1174–1194; and Markus M. Mobius and Tanya S. Rosenblat, "Why Beauty Matters," *American Economic Review* 96 (March 2006): 222–235.

equals $w_W(1 - d)$ when she has to work alongside a black worker (where d is a positive number). The white worker then perceives her take-home pay to be less than it actually is. Similarly, white customers might dislike purchasing goods and services from black sellers. The white customer would act as if the price of the good is not p dollars, but instead equals $p(1 + d)$.

The discrimination coefficient, therefore, "monetizes" prejudice, regardless of whether the source of the prejudice is the employer (leading to **employer discrimination**), the employee (leading to **employee discrimination**), or the customer (leading to **customer discrimination**).

One can interpret Becker's theory in terms of the framework developed in the chapter on compensating differentials. The theory of compensating differentials is based on the idea that persons consider "the whole of the advantages and disadvantages" of an economic exchange. A prejudiced person includes the race, ethnicity, and gender of market participants in the long list of advantages and disadvantages that influence the value of the exchange. The labor market, therefore, will have to generate compensating differentials to compensate prejudiced persons for their utility loss or gain.

9-3 Employer Discrimination

There are two types of workers in the labor market: white and black workers.[4] Suppose that black and white workers are perfect substitutes. We can then write the firm's production function as

$$q = f(E_W + E_B) \tag{9-1}$$

where q is the firm's output, E_W gives the number of white workers, and E_B gives the number of black workers. For simplicity, we ignore the role of capital in the production process.

Note that the firm's output depends on the *total* number of workers hired, regardless of their race. The firm gets the same output if it hires 50 white workers and 50 black workers, or if it hires 100 black workers and no white workers, or if it hires 100 white workers and no black workers. As a result, the output produced by hiring one more worker, or the marginal product of labor MP_E, is the same regardless of whether that last worker is black or white. Because black and white workers are equally productive, any differences eventually observed in the economic status of the two groups cannot be attributed to skill differentials but must arise from the discriminatory behavior of market participants.

Consider a competitive firm that is deciding how many of each of the two types of workers to hire. Before introducing the employer's prejudice into the analysis, let's first review the decision of a firm that does not discriminate. This color-blind firm faces constant input prices of w_W and w_B dollars for white and black labor, respectively. Because both groups of workers have the same value of marginal product, a nondiscriminatory firm will hire whoever is cheaper. If the market wage for black workers were below the market wage for white workers, the firm would only hire blacks. The opposite would happen if the black wage exceeded the white wage.

Let's suppose that the market-determined wage of black labor is less than the market-determined wage of white labor, or $w_B < w_W$. A color-blind firm will then hire black workers up to the point where the black wage equals the value of their marginal product, or

$$w_B = VMP_E \tag{9-2}$$

Figure 9-1 illustrates this profit-maximizing condition. A color-blind firm, therefore, hires E_B^* black workers.

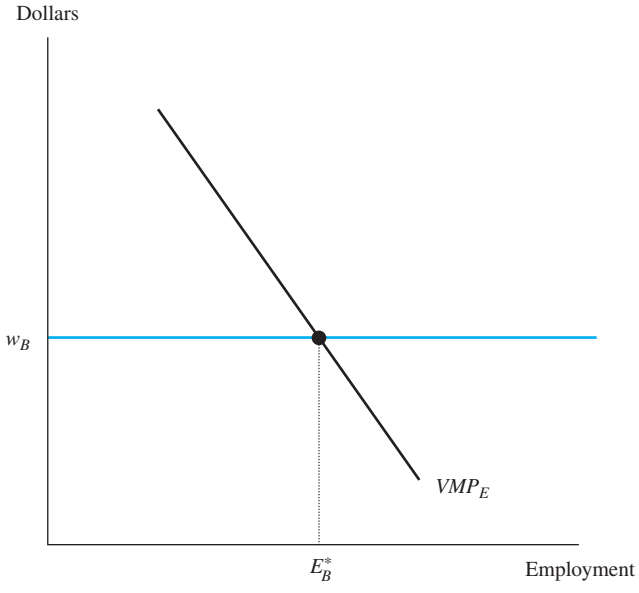
Employment in a Discriminatory Firm

Let's now describe the hiring decision of a firm that discriminates. The employer acts as if the black wage is not w_B, but instead equals $w_B(1 + d)$. The employer's hiring decision, therefore, is not based on a comparison of w_W and w_B, but on a comparison of w_W and $w_B(1 + d)$. The employer will hire the input that has a lower utility-adjusted price. The decision rule for a prejudiced employer is

$$\begin{aligned} &\text{Hire only blacks if } w_B(1 + d) < w_W \\ &\text{Hire only whites if } w_B(1 + d) > w_W \end{aligned} \tag{9-3}$$

[4] The presentation of Becker's theory of employer discrimination closely follows the exposition in Matthew S. Goldberg, "Discrimination, Nepotism, and Long-Run Wage Differentials," *Quarterly Journal of Economics* 97 (May 1982): 307–319.

FIGURE 9-1 The Hiring Decision of a Firm That Does Not Discriminate
If the market-determined black wage is less than the white wage, a nondiscriminatory firm hires only blacks. It hires black workers up to the point where the black wage equals the value of marginal product of labor, or E_B^*.



The decision rules in equation (9-3) highlight a key implication of the Becker model of employer discrimination: *Firms have a segregated workforce if black and white workers are perfect substitutes.* [5]

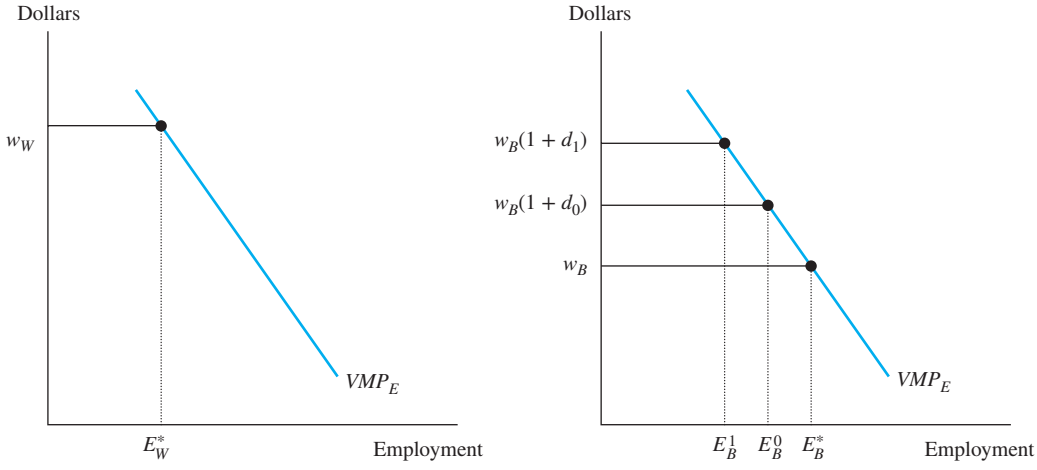
There are, therefore, two types of firms: those that hire an all-white workforce, which we will call "white firms"; and those that hire an all-black workforce, or "black firms." The race of the firm's workforce depends on the magnitude of the employer's discrimination coefficient. Employers with little prejudice, and small discrimination coefficients, will hire only blacks; employers who are very prejudiced, with large discrimination coefficients, will hire only whites. Figure 9-2*a* illustrates the employment decision of white firms, and Figure 9-2*b* illustrates the decision of black firms.

The white firm hires workers up to the point where the wage of white workers equals the value of marginal product, or $w_W = VMP_E$. We are assuming that the white wage exceeds the black wage. The white firm, therefore, pays an excessively high price for its workers and consequently hires few of them (or E_W^* in the figure).

Figure 9-2*b* shows that even black firms tend to hire too few workers. Recall that a color-blind firm hires E_B^* black workers, where the *actual* black wage equals the value of marginal product. A firm with discrimination coefficient d_0, however, acts as if the price of

[5] An empirical study of the racial composition of the firm's workforce is given by William J. Carrington and Kenneth R. Troske, "Interfirm Segregation and the Black/White Wage Gap," *Journal of Labor Economics* 16 (April 1998): 231–260; see also Kimberly Bayard, Judith K. Hellerstein, David Neumark, and Kenneth Troske, "Ethnicity, Language, and Workplace Segregation: Evidence from a New Matched Employer-Employee Data Set," *Journal of Labor Economics* 21 (October 2003): 877–922.

FIGURE 9-2 The Hiring Decision of a Prejudiced Firm
Firms that discriminate can be either white firms (if the discrimination coefficient is very high) or black firms (if the discrimination coefficient is relatively low). A white firm hires white workers up to the point where the white wage equals the value of marginal product. A black firm hires black workers up to the point where the utility-adjusted black wage equals the value of marginal product. Firms that discriminate hire fewer workers than firms that do not.



(*a*) White Firm (*b*) Black Firm

black labor is $w_B(1 + d_0)$. This discrimination coefficient is small enough that the firm will still want to hire only blacks. The firm hires black workers up to the point where the utility-adjusted price equals the value of marginal product, or $w_B(1 + d_0) = VMP_E$. As shown in Figure 9-3*b,* this firm hires E_B^0 workers. A firm with a larger discrimination coefficient d_1 hires even fewer workers (or E_B^1), and so on. The number of black workers hired, therefore, is smaller for firms with larger discrimination coefficients. Because employers do not like hiring black workers, they minimize their discomfort by hiring fewer of them.

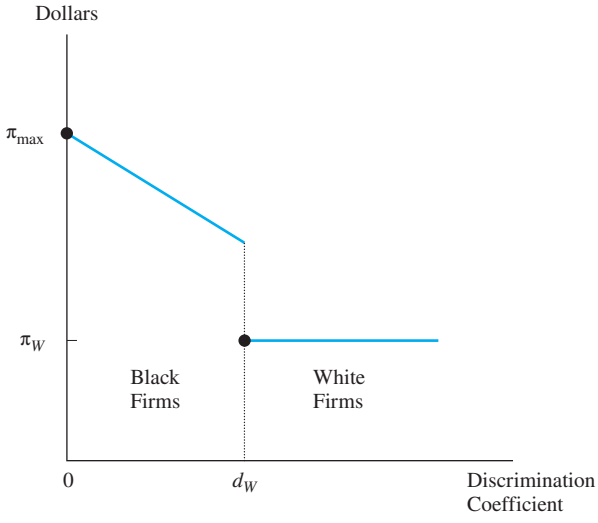
Discrimination and Profits

The analysis in Figure 9-2 leads to a fundamental insight: *Discrimination does not pay.*

To see why, consider first the profitability of white firms. Those firms hire E_W^* workers. This hiring decision is unprofitable in two distinct ways. First, the prejudiced employer could have hired the same number of black workers at a lower wage. In other words, because black and white workers are perfect substitutes, white firms could have produced the same output at a lower cost. In addition, white firms are hiring the *wrong* number of workers; a color-blind firm would hire many more workers, or E_B^*. By not hiring the right number of workers, white firms further reduce their profits. This argument also implies that even black firms that discriminate are giving up profits. Because discriminatory black firms are hiring too few workers (such as E_B^0 or E_B^1), they too are giving up profits as they minimize contact with black workers.

Figure 9-3 shows the relation between the firm's profits and the discrimination coefficient. The most profitable firm is the firm with a discrimination coefficient equal to zero. This color-blind firm hires an all-black workforce of E_B^* workers and has profits equal to π_{max} dollars.

FIGURE 9-3 Profits and the Discrimination Coefficient
Discrimination reduces profits in two ways. Even if the discriminatory firm hires only black workers, it hires too few workers. If the discriminatory firm hires only white workers, it hires too few workers at a very high wage.



Firms with slightly positive discrimination coefficients still have an all-black workforce but employ fewer black workers and earn lower profits. At some threshold level of prejudice, given by the discrimination coefficient d_W, the utility loss of hiring blacks is too large and the firm hires only whites. As a result, profits take a dramatic plunge (to π_W dollars) because the firm is paying a much higher wage than it needs to. Because all-white firms hire the same number of white workers (E_W^*) regardless of their discrimination coefficient, all white firms earn the same profits.

The Becker model of employer discrimination, therefore, predicts that discrimination is unprofitable. Discriminatory firms lose on two counts: They are hiring the "wrong color" of workers and/or they are hiring the "wrong number" of workers. Both of these hiring decisions move the firm away from the profit-maximizing level of employment, or E_B^* workers.

The implications of this prediction are far-reaching. If employers are indeed the source of racial prejudice in competitive markets, competition is a minority group's best friend. Free entry and exit of firms ensure that firms in the market do not earn excess profits. Discriminatory employers must then pay for the right to discriminate out of their own pocket. A color-blind firm, therefore, should eventually be able to buy out all the other firms in the industry, and employer discrimination would wither away.[6]

Labor Market Equilibrium

The comparison of the utility-adjusted price of black labor with the price of white labor in equation (9-3) tells us if a particular firm becomes a black firm or a white firm. Firms with small discrimination coefficients tend to become black firms and firms with large

[6] This conclusion assumes that all firms have the same production function. If discriminatory firms are more efficient and can produce output at lower costs, they can persist in their discriminatory behavior.

discrimination coefficients tend to become white firms. We can use this insight to derive the demand curve for black workers in the labor market.

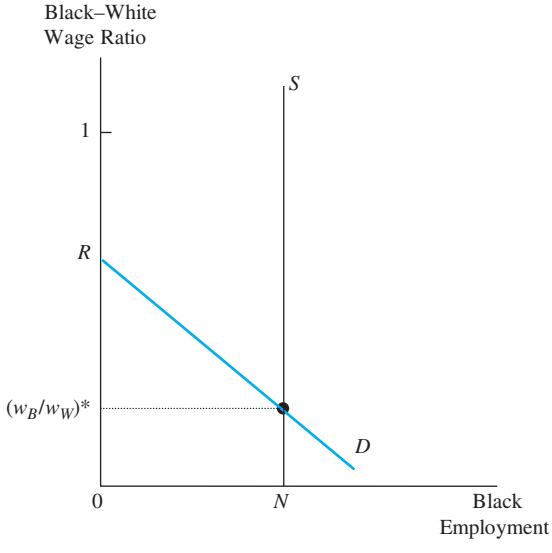
Assume initially that *all* employers discriminate against blacks; every firm has a positive discrimination coefficient. When the black wage exceeds the white wage, and the black–white wage ratio (w_B/w_W) is above 1, no employer, not even the employer who minds blacks the least and has the smallest discrimination coefficient, wants to hire blacks. As illustrated in Figure 9-4, there is zero demand for black workers. In fact, even if the black wage were slightly less than the white wage, the utility-adjusted black wage will probably exceed the white wage for all firms, and no employer will want to hire blacks.

But suppose now that the relative black wage decreases further. At some point, the firm with the least prejudice crosses a threshold (given by point *R* in the figure), and this firm becomes a black firm because blacks have become relatively cheaper than whites—even after accounting for the disutility from hiring them. As the relative black wage keeps on falling, more firms become black firms because the lower black wage compensates for the firms' prejudice. Moreover, those firms that were already hiring blacks take advantage of the lower black wage by hiring even more blacks. As the relative wage of blacks falls further and further, therefore, the quantity demanded of black workers increases. If the black wage is very low relative to the white wage, even firms with very large discrimination coefficients are "bought off" and will hire blacks. The market demand curve for black labor (or *D* in Figure 9-4) is downward sloping.

Of course, the equilibrium black–white wage ratio depends not just on the demand for black workers but also on their supply. For convenience, Figure 9-4 assumes that the supply curve of black workers is perfectly inelastic; there are *N* black workers regardless of the

FIGURE 9-4 Determination of Black–White Wage Ratio in the Labor Market

If the black–white wage ratio is very high, no firm in the labor market will want to hire blacks. As the black–white wage ratio falls, more and more firms are compensated for their disutility and the demand for black workers rises. The equilibrium black–white wage ratio is given by the intersection of supply and demand, and equals $(w_B/w_W)^*$. The assumption that all firms are prejudiced implies that the equilibrium black–white wage ratio is below 1.



relative black wage. The equilibrium black–white wage ratio, or $(w_B/w_W)^*$, is given by the intersection of the supply and demand curves. If the relative black wage is above equilibrium, too many black workers are looking for work relative to demand, and the relative black wage falls. Conversely, if the relative black wage is below equilibrium, there are too few black workers available relative to demand, and the relative black wage rises.

The Equilibrium Black–White Wage Differential

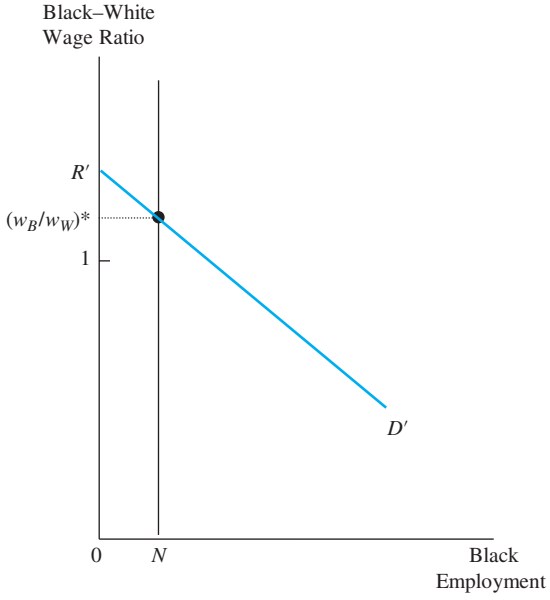
The equilibrium in Figure 9-4 has a number of interesting properties. Most important, the intersection of the supply and demand curves occurs below the point where the black–white wage ratio equals 1, so employer discrimination generates a racial wage gap. The employer cares about working conditions, particularly the race of the workforce. Because we assumed that all employers dislike hiring blacks, a compensating differential arises to compensate employers for hiring those workers. In effect, the market "compensates" employers so as to soften their resistance to hiring blacks.

Further, the allocation of black workers to firms is not random. Black workers are hired by the firms that chose to become black firms. And black firms are run by employers who have the smallest discrimination coefficients. Black workers, therefore, are matched with the least-prejudiced employers, while white workers are matched with the most-prejudiced employers.

We assumed that all firms discriminate against blacks. Some firms, however, might prefer to hire blacks. Because nepotistic firms get utility from hiring blacks, these firms would hire blacks even if the black wage were higher than the white wage. As a result, the demand curve for black labor starts at the intercept R' in Figure 9-5, where the relative black wage is above 1.

FIGURE 9-5 Nepotism and the Black–White Wage Ratio

If some firms prefer to hire blacks, they would be willing to hire blacks even if the black–white wage ratio exceeds 1, shifting the demand curve up to D'. If the supply of blacks is sufficiently small, it is then possible for the black–white wage ratio to exceed 1.



If the labor market has relatively few black workers, the equilibrium black–white wage ratio could be above 1, *even if most firms in the labor market dislike hiring blacks*. Because the labor market matches black workers with employers who prefer to hire blacks, blacks are then able to sell their services to those firms that are willing to pay for the right to hire them.

The observed black–white wage gap, therefore, should not be interpreted as a summary measure of how much prejudice there is in the labor market, such as the mean or median discrimination coefficient. The equilibrium relative black wage measures the discrimination coefficient of the *marginal* firm, the last firm that made the switch and hired blacks. The market black–white wage gap only measures what it took to compensate that marginal firm.

9-4 Employee Discrimination

The source of prejudice in the labor market need not be the employer. Instead, it might be fellow workers.

Suppose that all whites dislike working alongside blacks and that blacks are indifferent about the race of their coworkers. White workers who receive a wage of w_W dollars will then act as if their wage rate is only $w_W(1 - d)$, where d is the discrimination coefficient. Because black workers do not care about the race of their coworkers, both their actual and utility-adjusted wage rates are given by w_B. We continue to assume that black and white workers are perfect substitutes in production.

A white worker has two job offers. Both employers offer the same wage of $20 per hour, but working conditions vary in the two firms. In particular, one firm has a completely white workforce, and the other firm has an integrated workforce, consisting of black and white workers. Because the white worker dislikes blacks, the two firms are not offering equivalent utility-adjusted wages. From the worker's perspective, the integrated firm offers a lower wage. As a result, integrated firms will have to offer more than $20 per hour if they wish to attract prejudiced white workers.

But a color-blind profit-maximizing employer would never choose to have an integrated workplace. The employer would not want to hire both black and white workers because white workers have to be paid a compensating wage differential, yet they have the same value of marginal product as blacks. The firm will then hire only whites if the white wage is below the black wage and will hire only blacks if the black wage is below the white wage. Because it does not pay to "mix," black and white workers are employed by different firms. Employee discrimination, like employer discrimination, implies a completely segregated workforce.

Unlike employer discrimination, however, employee discrimination does *not* generate a wage gap between equally skilled black and white workers. Color-blind firms hire whichever labor is cheaper. If blacks are cheaper, employers hire blacks. If whites are cheaper, employers will hire whites. The competition for the cheapest workers will eventually equalize the wage of the two groups, as employers increase their demand for whichever group costs less. If blacks and whites were perfect substitutes, therefore, a model of employee discrimination could not explain why equally skilled blacks might earn less than equally skilled whites.

Finally, employee discrimination does not affect the profitability of firms. Because all firms pay the same price for an hour of labor, and because black and white workers are perfect substitutes, there is no advantage to being either a black or a white firm. There are no market forces, therefore, that will tend to diminish the importance of employee discrimination over time.[7]

9-5 Customer Discrimination

If customers have a taste for discrimination, their buying decisions are not based on the actual price of the good, p, but on the utility-adjusted price, or $p(1 + d)$, where d is the discrimination coefficient. If whites dislike buying from black sellers, customer discrimination will then reduce the demand for goods and services sold by minorities.

As long as a firm can allocate a particular worker to one of many different positions within the firm, customer discrimination may not matter much. The firm can place its black workers in jobs that require little customer contact (such as an assembly line), and place some of its white workers in the service division (where they regularly interact with customers). In effect, the employer segregates the workforce internally so that white workers mainly fill "contact" positions and black workers remain hidden from outside view.[8]

Customer discrimination could have an adverse impact on black earnings when the firm cannot easily hide black workers from public view. A firm employing a black worker in a sales job, for example, will have to lower the price of the product to compensate white buyers for their disutility. The black wage might then fall because black workers have to compensate the firm for the decline in profits. In short, the impact of customer discrimination on the black wage will depend on the relative demand for contact and noncontact jobs and the relative supply of black workers.

A large survey of employers in Atlanta, Boston, Detroit, and Los Angeles in the early 1990s allows us to do a simple difference-in-differences exercise that shows how the customers' race and the degree of contact between workers and customers influences hiring decisions. Suppose we classify the firms in the survey into two types: contact firms, where the workers talk face-to-face with the customers, and noncontact firms. Table 9-2 shows that 58 percent of newly hired workers in contact firms that have a mostly black customer base are black. This contrasts with the hiring of firms that have a mostly white customer base; only 9 percent of newly hired workers are black. The difference seems to suggest that customer discrimination reduces the fraction of black hires by 49.0 percentage points.

But contact firms with a mainly black customer base are probably located in black areas of the city. These firms would naturally attract many black job applicants, and the racial composition of the applicant pool would affect the racial composition of the firm's workforce.

[7] Barry R. Chiswick, "Racial Discrimination in the Labor Market: A Test of Alternative Hypotheses," *Journal of Political Economy* 81 (November 1973): 1330–1352.

[8] Lawrence M. Kahn, "Customer Discrimination and Affirmative Action," *Economic Inquiry* 24 (July 1991): 555–571; and George J. Borjas and Stephen G. Bronars, "Consumer Discrimination and Self-Selection into Self-Employment," *Journal of Political Economy* 97 (June 1989): 581–605.

TABLE 9-2 Customer Discrimination and Percentage of Newly Hired Workers Who Are Black

Source: Harry J. Holzer and Keith R. Ihlanfeldt, "Customer Discrimination and Employment Outcomes for Minority Workers," *Quarterly Journal of Economics* 113 (August 1998): 846.

Type of Firm	More Than Half of Firm's Customers Are Black (%)	More Than 75% of Firm's Customers Are White (%)	Difference (%)
Contact between customers and workers	58.0	9.0	49.0
No contact between customers and workers	46.6	12.2	34.4
Difference-in-differences	—	—	**14.6**

To isolate the impact of customer discrimination, therefore, we need a control group. The firms in the survey where workers do not have any contact with customers provide one possible control group. As Table 9-2 shows, the fraction of newly hired workers who are black falls from 46.6 to 12.2 percent as the customer base shifts from being mainly black to mainly white in the no-contact firms, a reduction of 34.4 percentage points. The 34.4-point gap estimates what one might expect to happen to black employment—*even in the absence of customer discrimination*—when a firm caters mainly to black customers and likely opens up shop in black neighborhoods and attracts many black job applicants.

The difference-in-differences estimate of the impact of customer discrimination is 14.6 percent. In other words, face-to-face contact between black workers and white customers lowers the probability that the firm hires black workers by about 15 percentage points.

Some of the most interesting evidence of customer discrimination has been uncovered in the market for baseball memorabilia. Collecting baseball cards is not a children's pastime. A 1909 Honus Wagner baseball card sold for $2.8 million in 2011. It turns out that the market price of baseball cards in the collectibles market depends not only on the most obvious factors—such as the number of career home runs and at-bats for a hitter and the number of wins and strikeouts for a pitcher—but also on the race of the player. Even after controlling for the position played and for the player's career "stats", the cards of white players cost about 10–13 percent more than the cards of black players.[9]

9-6 Statistical Discrimination

The theory of taste discrimination helps us understand how differences between equally skilled blacks and whites (or men and women) can arise in the labor market. But racial and gender differences may arise *even in the absence of prejudice* when membership in a particular group (for example, being a black woman) carries information about a person's skills and productivity.[10]

[9] Clark Nardinelli and Curtis Simon, "Customer Racial Discrimination in the Market for Memorabilia: The Case of Baseball," *Quarterly Journal of Economics* 105 (August 1990): 575–596.

[10] Edmund S. Phelps, "The Statistical Theory of Racism and Sexism," *American Economic Review* 62 (September 1972): 659–661; Dennis J. Aigner and Glen G. Cain, "Statistical Theories of Discrimination in Labor Markets," *Industrial and Labor Relations Review* 30 (January 1977): 175–187; and Shelly J. Lundberg and Richard Startz, "Private Discrimination and Social Intervention in Competitive Labor Markets," *American Economic Review* 73 (June 1983): 340–347.

The economic incentives that produce statistical discrimination are easy to describe. Suppose that a color-blind, gender-blind, and profit-maximizing employer has a job opening. The employer wants to add a worker to a finely tuned team that will develop a revolutionary fully immersive AI video game in the next few years. The employer is looking for a worker who, in addition to the usual requisites of intelligence and ambition, can be counted on being a team member over the long haul.

Two people apply for the job. Their résumés are identical. Both just graduated from the same college, majored in the same field, enrolled in the same courses, and had similar class rankings. Moreover, both applicants passed the interview with flying colors. They were bright, motivated, knowledgeable, and articulate. It just happens, however, that one of the applicants was a man and the other was a woman.

During the interview, the employer specifically asked them if they viewed the prospective job as one where they could grow and develop over the next few years. Both replied that they saw the job as a terrific opportunity and that it was hard to foresee how any other employment or nonmarket opportunities could conceivably compete. Based on the "paper trail" (that is, the résumé, the interview, and any other screening tests), the employer will find it difficult to choose between the two applicants. The employer knows, however, that because both applicants need a job, the assertion that they intend to stay at the firm for the next few years may not be sincere.

To make an informed decision (rather than just toss a coin), the employer will evaluate the employment histories of similarly situated men and women that this firm—or other firms—hired in the past. Suppose this review of the statistical record reveals that many women cut back on hours worked or leave the firm when they reach their late twenties.[11] The employer has no way of knowing if the female job applicant under consideration intends to follow this path. Nevertheless, the employer infers from the statistical data that the woman has a higher probability of quitting her job prior to the completion of the software program. Because a quit would disrupt the team's work and increase costs, the profit-maximizing employer offers the job to the man.

The decision to favor one group over another arises because the information gathered from the résumé and the interview is an imperfect predictor of the applicant's true productivity. The uncertainty encourages the employer to use statistics about the average performance of the group (hence the name **statistical discrimination**) to predict a particular applicant's productivity. As a result, applicants from high-productivity groups benefit from their membership in those groups, while applicants from low-productivity groups do not.[12]

[11] Despite the premise made in this illustrative example, some of the evidence does not suggest that women have higher quit rates than men; see Francine D. Blau and Lawrence M. Kahn, "Race and Sex Differences in Quits by Young Workers," *Industrial and Labor Relations Review* 34 (July 1981): 563–577; and W. Kip Viscusi, "Sex Differences in Worker Quitting," *Review of Economics and Statistics* 62 (August 1980): 388–398; and Nachum Sicherman, "Gender Differences in Departures from a Large Firm," *Industrial and Labor Relations Review* 49 (April 1996): 484–505.

[12] There is one crucial difference between statistical discrimination and the signaling model presented in the human capital chapter. In the signaling model, workers invest in education to separate themselves from the pack. In the statistical discrimination model, the traits that employers use to predict productivity—such as race, gender, or national origin—are immutable.

Firms use information on group membership in many contexts. For example, women tend to live longer than men. Suppose that a man and a woman who were born on the same day and have the same overall physical condition apply to buy life insurance. The insurance company has no way of knowing who will live longer, but its prior experience indicates the woman will likely outlive the man. This fact will surely matter when setting insurance premiums. Similarly, teenagers tend to have more car accidents than older drivers. If a teenager and a 40-year-old apply to buy auto insurance, the insurance company will typically charge more to the teenager—even though both drivers may have clean driving records. In short, competitive firms frequently use statistical discrimination to fill in information gaps when the firm cannot perfectly predict the risks or rewards of particular economic transactions.

Statistical Discrimination and Wages

Let's gather all the information in the applicant's résumé, the interview, and any other screening tests and give it a score, say, *T.* Suppose this score is perfectly correlated with productivity. A score of 15 indicates that the applicant's true value of marginal product is \$15; a score of 30 indicates a true value of marginal product of \$30; and so on. The job applicant would then be offered a wage that equals the test score.

Of course, the assumption that the test score perfectly predicts productivity is very unrealistic. Some low-scoring applicants will turn out to be quite productive, and some high-scoring applicants will be spectacular failures. Therefore, employers may want to link the applicant's wage offer not only to the applicant's own score *T,* but also to $\overline{T}$, the average test score of the group the applicant belongs to.

Under some conditions, the applicant's expected productivity will be a weighted average of the applicant's test score and of the group's average score. We can then write the firm's wage offer to that applicant as[13]

$$w = (1 - \alpha)\,\overline{T} + \alpha T \qquad \textbf{(9-4)}$$

If the parameter α is equal to one, the applicant's wage depends only on the applicant's test score. Because the employer ignores the group average, this is the extreme case where the screening test predicts the applicant's productivity perfectly. At the other extreme, the parameter α is equal to zero. Equation (9-4) then indicates that the worker's own test score is meaningless and plays no role in the wage-setting process. The employer relies entirely on the group average to set the worker's wage. The parameter α, therefore, reflects the correlation between the test score and true productivity. The higher the predictive power of the test, the higher the value of α.

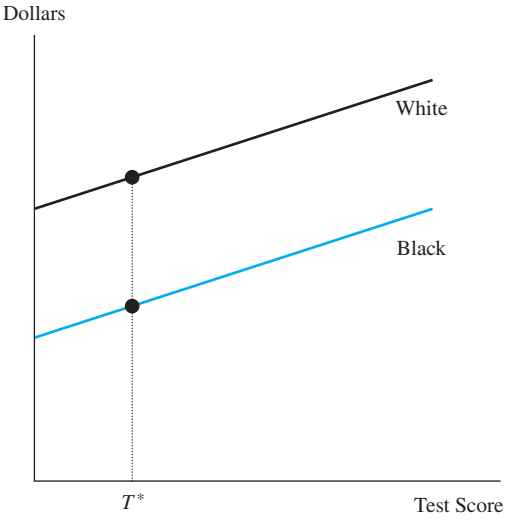
Equation (9-4) suggests that statistical discrimination can influence the wage of minorities and women in two distinct ways. Statistical discrimination affects both the intercept and the slope of the line relating the wage and the applicant's test score.

Consider first the situation illustrated in Figure 9-6*a.* The average black score in the test, $\overline{T}_B$, is lower than the average white score, $\overline{T}_W$, but the correlation between test scores and productivity (α) is assumed to be the same for the two groups. Equation (9-4) then

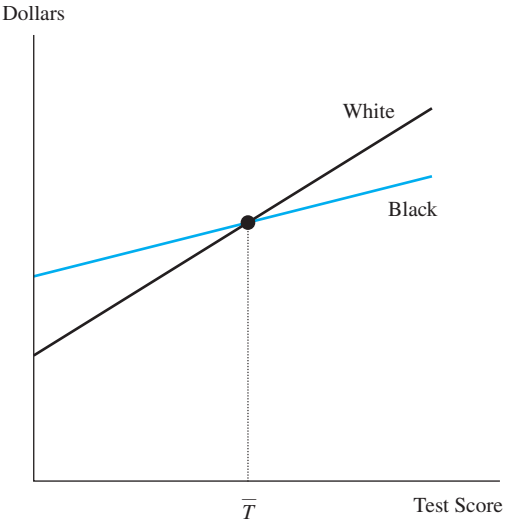
[13] Lundberg and Startz, "Private Discrimination and Social Intervention in Competitive Labor Markets." The key assumption used to derive equation (9-4) is that the frequency distribution of the unobserved component of an applicant's productivity follows a normal distribution.

FIGURE 9-6 Statistical Discrimination and Wages

The worker's wage depends on both his own test score and the mean score of workers in his racial group. (*a*) If black workers have a lower average score, a white worker who gets T^* points earns more than a black worker with the same score. (*b*) If the test is a better predictor of productivity for whites, high-scoring whites earn more than high-scoring blacks, but low-scoring whites earn less than low-scoring blacks.



(*a*) Whites Have a Higher Average Score



(*b*) Test Is Better Predictor for White Workers

implies that the white line lies above the black line because whites, on average, do better on the test, but that both lines have the same slope. If a black and a white worker get the same test score (say T^*), the white worker is offered a higher wage because employers expect the white applicant to be more productive than the black applicant.

It is also possible that the two groups have the same mean score $\overline{T}$, but the test may be more informative for one group. It has been argued that some tests predict the true productivity of blacks and other groups imprecisely because of "cultural bias." Standardized tests written by white male academics reflect a set of upper-middle-class values and experiences that may be unfamiliar to persons raised in different environments. As a result, the value of the parameter α may differ between blacks and whites as well as between men and women. If the test is a poor predictor of productivity for black workers, then $\alpha_B < \alpha_W$.

Figure 9-6*b* shows the impact of this type of cultural bias on the wage.[14] If the test were a poor predictor of productivity for black workers, the line relating the wage and the test score would be relatively flat for blacks. Because the test is such an imperfect predictor of productivity, employers would view most black workers as having roughly similar productivities and would offer them roughly similar wages. A black worker's wage would then be mostly set on the basis of the group average, while the white worker's wage would be

[14] The assumption that both groups have the same mean score but that $\alpha_B < \alpha_W$ implies that the intercept of the black wage line ($(1 - \alpha_B)\overline{T}$) exceeds the intercept of the white wage line ($(1 - \alpha_W)\overline{T}$).

mostly determined by the white worker's own qualifications. Low-scoring blacks benefit relative to high-scoring blacks because the employer does not trust the worker's test score. Statistical discrimination implies that low-scoring blacks earn more than low-scoring whites, but that the opposite will be true for high-scoring workers.

Interestingly, statistical discrimination does not necessarily predict that the average black will be paid less than the average white.[15] But it does raise the possibility that discrimination benefits some black workers, while harming others.

9-7 Experimental Evidence

It is very difficult to measure a particular employer's discrimination coefficient, or to determine if a particular employer is engaging in statistical discrimination. After all, it is illegal to discriminate on the basis of race and gender, so employers will not willingly reveal their prejudicial behavior.

A number of studies have attempted to bypass this measurement problem by conducting labor market experiments. In these experiments, researchers contact a number of employers at random. The experiments are cleverly designed to induce employers to reveal their hiring preferences.

One particularly well-known experiment tried to determine if employers were more willing to hire workers named Emily or Greg than workers named Lakisha or Jamal.[16] Researchers sent out about 5,000 fake résumés in response to over 1,000 job ads that actually appeared in Boston and Chicago newspapers. The résumé did not specify the applicant's race. But the researchers gave employers a *hint* of the applicant's race by giving the fake applicant a name that was either "white-sounding" or "black-sounding." Among the white-sounding names were Emily Walsh and Greg Baker, while the black-sounding names included Lakisha Washington and Jamal Jones.[17] The résumés also varied slightly in describing the applicant's marketable skills. Some résumés stated that the applicant had many years of labor market experience, or that the applicant had attained some type of certification, or that the applicant knew a foreign language.

After mailing out the fake résumés, the researchers sat back and waited for employers to call back the fake applicants for interviews. Holding the skills in the applicant's résumé constant, the applicants with white-sounding names got about one callback for every 10 résumés sent. In contrast, the applicants with black-sounding names got one callback

[15] Empirical tests of the statistical discrimination hypothesis include Joseph G. Altonji and Charles R. Pierret, "Employer Learning and Statistical Discrimination," *Quarterly Journal of Economics* 116 (February 2001): 313– 350; and David H. Autor and David Scarborough, "Does Job Testing Harm Minority Workers? Evidence from Retail Establishments," *Quarterly Journal of Economics* 123 (February 2008): 219–277.

[16] Marianne Bertrand and Sendhil Mullanaithan, "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review* 94 (September 2004): 991–1013.

[17] The naming conventions used by black and white parents have diverged in recent years. A study of the names given to every single child born in California between 1961 and 2000 discovered that 40 percent of the black girls born in California in that period were given a name that not a single white girl born in those years was given; see Roland Fryer and Steven Levitt, "The Causes and Consequences of Distinctive Black Names," *Quarterly Journal of Economics* 119 (August 2004): 767–805.

for every 15 résumés sent. A black applicant would need eight more years of work experience to even out the gap!

The experimental approach has been extended beyond the simple act of mailing out fake résumés. Some researchers have actually sent out "experimental" human beings in actual job interviews to see how employers react to the characteristics of live applicants. In these "hiring audits," two matched job applicants are similar in all respects, except that they differ in their race or gender. The audit is conducted at a number of firms and the data are then examined to determine if the outcome of the job application differed between black and white applicants, or between men and women.

In the summer of 1989, for example, a hiring audit was conducted of employers in the Chicago and San Diego areas.[18] The employers, who were trying to fill entry-level jobs requiring few skills, were chosen at random from the classified ads in the Sunday edition of the *Chicago Tribune* and the *San Diego Union.* The average job applicant participating in the audit was a neatly dressed 22-year-old man who had a high school diploma, did not have a criminal record, had some college credits, and some work experience as a stockperson or waiter. The only notable difference between the matched pair of applicants sent to a particular firm was that one was Hispanic, with dark hair and light brown skin, and a slight accent. The other was a non-Hispanic white with brown, blonde, or red hair, and who did not have an accent. The white job applicant was 33 percent more likely to be interviewed and 52 percent more likely to receive a job offer.

A similar hiring audit of low-priced and high-priced restaurants also suggests that employers view male and female workers differently.[19] Young men and women carrying identical (and fictitious) résumés were sent out to apply for jobs at Philadelphia restaurants. A waiter can typically do much better, in terms of wages and tips, at a fancy restaurant. Even though the applicants looked alike on paper, 8 of the 10 job offers made by low-priced restaurants went to women, but 11 of the 13 job offers made by high-priced restaurants went to men.

Audit studies have been criticized because even though the researcher might think that the experimental job applicants are identical and have been trained to respond in a similar fashion in job interviews, employers might not get the same impression in real-world interactions. Moreover, the job applicants participating in the audit know the purpose of their visits to the various employers and that information might (subconsciously or not) influence their behavior during the application process. Finally, the theory of taste discrimination predicts that it is the discrimination coefficient of the *marginal* firm that determines the wage gap between two groups in the labor market. The information provided by the audit studies, which tend to measure the preferences of the average firm, may not be all that relevant in explaining observed wage differences.[20]

[18] Harry Cross, *Employer Hiring Practices: Differential Treatment of Hispanic and Anglo Job Seekers,* Urban Institute Report 90-4, Washington, D.C.: The Urban Institute Press, 1990.

[19] David Neumark, Roy J. Bank, and Kyle D. Van Nort, "Sex Discrimination in Restaurant Hiring: An Audit Study," *Quarterly Journal of Economics* 111 (August 1996): 915–941.

[20] It is worth noting that one careful appraisal of the audit literature concludes that carefully controlling for the skills of black and white job applicants suggests little evidence of differential treatment between equally qualified black and white applicants; see James J. Heckman, "Detecting Discrimination," *Journal of Economic Perspectives* 12 (Spring 1998): 101–116.

Theory at Work

ORCHESTRATING IMPARTIALITY

For decades, the musicians who played in the major symphony orchestras of the United States were handpicked by the music director of the orchestra. The director would typically audition the students of a selected group of teachers and would single-handedly choose the winner. This hiring process typically led to a symphony orchestra composed of mostly male musicians. The typical symphony orchestra has around 100 musicians, and fewer than 10 were women.

As part of an effort to make the hiring process fairer and to increase diversity, the major orchestras adopted a process of "blind" auditions in the 1980s and 1990s. Job applicants would play a musical piece hidden behind a screen, typically a large piece of heavy cloth hanging from the ceiling. The music director and other persons involved in the hiring decision could hear the applicant play but could not see who the applicant was.

The introduction of blind auditions greatly increased the representation of women in the major symphony orchestras. In particular, the use of the screen increased the probability that a female musician advanced out of the preliminary rounds by 50 percent. By the 1990s, more than 20 percent of the players in the major orchestras were women, and about half of that increase be directly traced to the adoption of the blind screening process.

Source: Claudia Goldin and Cecilia Rouse, "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians," *American Economic Review* 90 (September 2000): 715–741.

9-8 Measuring Discrimination

Suppose that we have two groups of workers, male and female. The average male wage is $\overline{w}_M$ and the average female wage is $\overline{w}_F$. One possible definition of discrimination is given by the difference in mean wages, or

$$\Delta \overline{w} = \overline{w}_M - \overline{w}_F \qquad \textbf{(9-5)}$$

This definition is unappealing because it is comparing apples and oranges. Many factors, other than discrimination, generate wage differences between men and women. Men, for example, may be more likely to have advanced degrees in high-paying computer fields. We would not want to claim that employers discriminate against women if men earn more partly because men are more likely to acquire those valuable degrees. A more appropriate definition of labor market discrimination compares the wage of equally skilled workers.

Therefore, we would like to adjust the "raw" wage gap given by $\Delta \overline{w}$ for differences in skills between men and women. This adjustment is typically done by estimating regressions that relate earnings to a wide array of socioeconomic and skill characteristics. To simplify, suppose that only one variable, schooling (s), affects earnings. The earnings functions for the two groups can then be written as

$$\text{Male earnings function:} \quad w_M = \alpha_M + \beta_M s_M$$

$$\text{Female earnings function:} \quad w_F = \alpha_F + \beta_F s_F \qquad \textbf{(9-6)}$$

The coefficient β_M tells us by how much a man's wage increases if he gets one more year of schooling, while the coefficient β_F gives the same statistic for a woman. If employers value the education acquired by women as much as they value the education acquired by men, these two coefficients would be equal ($\beta_M = \beta_F$). Similarly, the intercepts α_M and

α_F give the intercepts of the earnings function. If employers valued the skills of men and women who have zero years of schooling equally, the two intercepts would be the same ($\alpha_M = \alpha_F$).

The regression model implies that the raw wage differential can be written as

$$\Delta \overline{w} = \overline{w}_M - \overline{w}_M = \alpha_M + \beta_M \overline{s}_M - \alpha_F - \beta_F \overline{s}_F \qquad \textbf{(9-7)}$$

where $\overline{s}_M$ and $\overline{s}_F$ give the mean schooling of men and women, respectively.

The Oaxaca–Blinder Decomposition

We can decompose the raw wage differential $\Delta \overline{w}$ into a portion that arises because men and women, on average, have different skills and a portion attributable to labor market discrimination. To conduct this decomposition, which has come to be known as the **Oaxaca–Blinder decomposition,** let's play a harmless algebraic trick.[21] In particular, add and subtract the term ($\beta_M \times \overline{s}_F$) to the right-hand side of equation (9-7). Rearranging the terms in the equation then leads to

$$\Delta \overline{w} = \underbrace{[(\alpha_M - \alpha_F) + (\beta_M - \beta_F)\overline{s}_F]}_{\text{Due to discrimination}} + \underbrace{\beta_M(\overline{s}_M - \overline{s}_F)}_{\text{Due to skills}} \qquad \textbf{(9-8)}$$

Equation (9-8) shows that the raw wage differential consists of two parts. It is useful to begin by discussing the last term in the equation. That term is zero if men and women have the same average schooling (or $\overline{s}_M = \overline{s}_F$). Part of the raw wage gap between men and women, therefore, arises because the two groups may differ in their skills.

The bracketed first term in equation (9-8) will be positive if employers either value a man's schooling by more than a woman's schooling ($\beta_M > \beta_F$) or just pay men more than women for any level of schooling so that the male intercept is larger ($\alpha_M > \alpha_F$). The wage gap that arises because of this differential treatment of men and women is typically defined as discrimination.

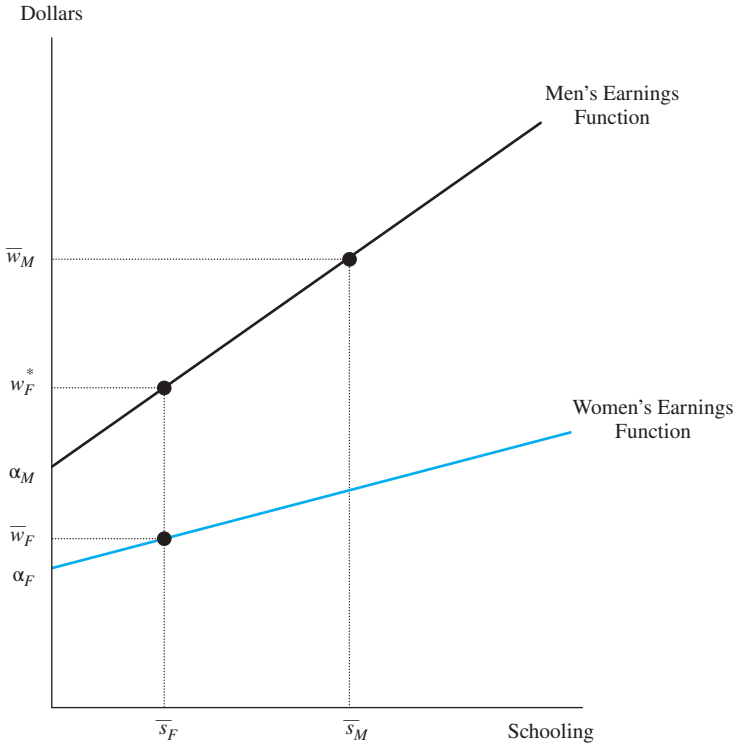
Figure 9-7 illustrates the intuition behind the decomposition. As drawn, the regression line relating earnings and schooling has a higher intercept and a steeper slope for men than for women. In other words, men start off with an advantage (they get paid more than women even if the two groups have zero years of schooling), and then get a bigger payoff from each additional year of schooling.

Suppose now that men, on average, have more education. The raw wage gap between men and women is given by the vertical difference $\overline{w}_M - \overline{w}_F$. The average woman with $\overline{s}_F$ years of schooling would earn w_F^* if she were "treated like a man." Therefore, the difference $w_F^* - \overline{w}_F$ can be attributed to discrimination. Part of the raw differential, however, arises because men have more schooling. The difference $\overline{w}_M - w_F^*$ is the part attributable to the average skill differential between men and women.

To simplify the exposition, we derived the Oaxaca–Blinder decomposition in a model where schooling is the only explanatory variable in the earnings function. The exercise can

[21] Ronald L. Oaxaca, "Male–Female Wage Differentials in Urban Labor Markets," *International Economic Review* 14 (October 1973): 693–709; and Alan S. Blinder, "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources* 8 (Autumn 1973): 436–455.

FIGURE 9-7 Measuring the Impact of Discrimination on the Wage
The average woman has $\overline{s}_F$ years of schooling and earns $\overline{w}_F$. The average man has $\overline{s}_M$ years of schooling and earns $\overline{w}_M$. Part of the wage differential arises because men have more schooling than women. If the average woman was paid as if she were a man, she would earn w_F^* dollars. A measure of discrimination is $w_F^* - \overline{w}_F$.



be easily extended if there are many other variables that affect earnings, including labor market experience, marital status, and region of residence. The basic insight is the same: The raw wage differential can be decomposed into a portion due to differences in characteristics between the two groups and a portion that remains unexplained and that we call discrimination.

What Does the Oaxaca–Blinder Decomposition Really Measure?

The claim that the Oaxaca–Blinder decomposition isolates the impact of discrimination on wages depends largely on whether we have controlled for *all* the ways in which the skills of the two groups differ. If there are some variables that affect earnings but are left out of the regression model, we will have an incorrect measure of discrimination.

In fact, we seldom observe all the variables that make up a worker's human capital stock. Most surveys, for example, provide little information on the quality of a worker's education (as opposed to the number of years of schooling). If men and women or blacks and whites systematically attend institutions that vary in quality, the Oaxaca–Blinder decomposition produces a biased measure of discrimination. For example, suppose that blacks attend

lower-quality schools. There will be a wage gap between black and white workers who have the same level of schooling. It would be incorrect to label wage differences between black and white workers with the same schooling as discrimination because, in fact, the two groups are not equally skilled.

Anyone who doubts that discrimination plays an important role in the labor market can always point out that a variable was left out of the model used to calculate the Oaxaca–Blinder decomposition. Even if we try to include in the model every single measure of skills that we can think of *and* that we can observe, someone can still assert that we have omitted such variables as ability, effort, motivation, and drive, and that these variables differ between the groups.

On the other hand, one could argue that defining discrimination as the wage gap between observationally equivalent men and women or blacks and whites underestimates the importance of discrimination. It is no coincidence that blacks have less schooling and attend lower-quality schools than whites or that women become grammar school teachers but not plumbers and electricians.

Cultural discrimination as well as differential funding of black and white schools influences the human capital accumulation of the various groups prior to their entry into the labor market. Even though employers are not responsible for these "preexisting" skill differences, other institutions may be. A more complete accounting of discrimination, therefore, might not want to net out the differences in skills among groups and would focus more on the raw wage gap.

9-9 Policy Application: The Black–White Wage Gap

In 1995, black workers earned about 21 percent less than white workers. Table 9-3 reports the results obtained from two alternative Oaxaca–Blinder decompositions of this wage gap. The first adjusts for differences in educational attainment, age, sex, and region of residence between the two groups. The second controls for all of these variables as well as for differences in the occupation and industry of employment.

The extent of measured discrimination clearly depends on the controls used. In the first decomposition, racial differences in educational attainment, age, and region of residence produced an 8.2 percent wage differential between the two groups so that labor market discrimination accounts for the residual, or a 13.4 percent wage gap. But if the

TABLE 9-3 The Oaxaca Decomposition of the Black–White Wage Differential, 1995

Source: Joseph G. Altonji and Rebecca M. Blank, "Race and Gender in the Labor Market," in Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics,* vol. 3C, Amsterdam: Elsevier, 1999, Table 5. The log wage differential between any two groups can be interpreted as being approximately equal to the percentage wage differential between the groups.

	Controls for Differences in Education, Age, Sex, and Region of Residence	Controls for Differences in Education, Age, Sex, Region of Residence, *and* Occupation and Industry
Raw log wage differential	−0.211	−0.211
Due to differences in skills	−0.082	−0.114
Due to discrimination	−0.134	−0.098

analysis also adjusts for differences in occupation and industry of employment, there is an 11.4 percent wage gap attributable to skills and "only" a 9.8 percent wage gap attributable to discrimination.

This exercise raises the conceptual question alluded to earlier: Which is the right set of controls? Should one calculate the wage differences among similarly skilled blacks and whites employed in the same occupation and industry before we decide if there is discrimination? Or is it possible that part of the racial differences in occupation and industry is due to barriers that prevent blacks from moving into certain types of jobs? One lesson from Table 9-3 is that one should look carefully at the "fine print" behind any Oaxaca–Blinder decomposition before one concludes that discrimination either plays a small or a substantial role in the labor market.

The Trend in the Black–White Wage Ratio

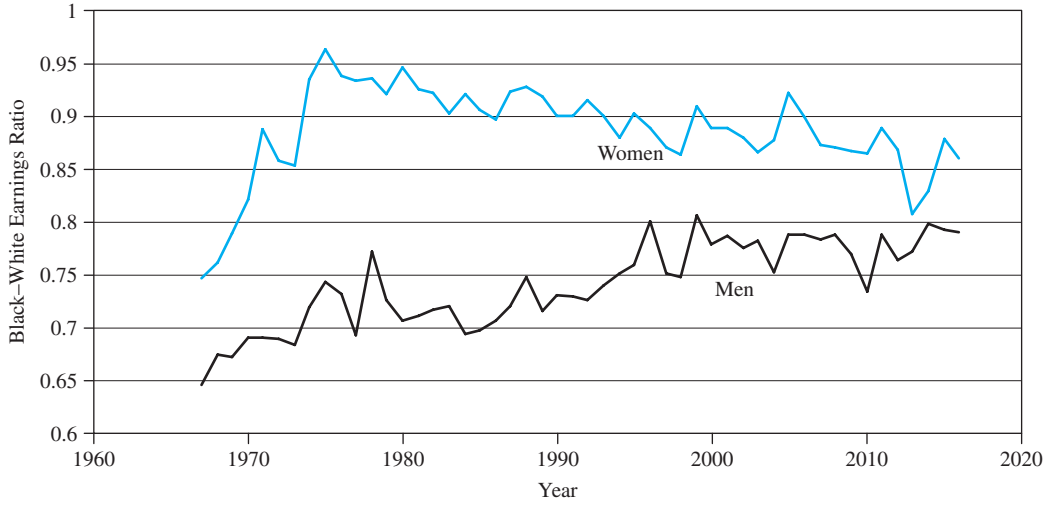
As Figure 9-8 illustrates, the wage ratio between black and white men rose dramatically in the past 50 years. In 1967, the ratio stood at about 0.65; by 1980, it had risen to 0.71; and by 2016, it stood at 0.79. This improvement in the relative economic status of black men is a continuation of long-run trends; the ratio was about 0.4 around 1940.

Figure 9-8 also shows that the wage ratio between black and white women rose rapidly before the mid-1970s, but there has been a slow downward drift since the 1980s. Between 1967 and 1975, this ratio rose from 0.75 to 0.96; it now stands at around 0.86. Despite this downward drift, the relative wage of both black men and women is substantially higher today than it was in the late 1960s.

Several hypotheses have been proposed to explain the improving economic status of African American men. The first is that the increasing human capital of black workers, particularly in terms of the quantity and quality of education, contributed greatly to the

FIGURE 9-8 Trend in Black–White Earnings Ratio, 1967–2016

Source: U.S. Bureau of the Census, "Historical Income Tables—People," Table P-38. "Full-Time, Year-Round Workers by Median Earnings and Sex," https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-people.html. The earnings refer to the median earnings of full-time, full-year workers aged 15 or above.



increase in the black relative wage.[22] In 1940, the typical 30-year-old white man had 9.9 years of schooling, as compared to 6.0 years for a comparable black. By 1980, the typical 30-year-old white man had 13.6 years of schooling and the comparable black had 12.2 years, a difference of only 1.4 years.

The disparity in school quality also narrowed dramatically. In the 1920s, pupil–teacher ratios in southern states were about 50 percent higher in black schools than in white schools. By the late 1950s, this differential had disappeared. As a result, the racial gap in the rate of return to school also vanished. The rate of return to school for white workers who entered the labor market around 1940 was 9.8 percent, but for black workers it was only 4.7 percent. For the cohorts that entered the labor force in the late 1970s, blacks actually had a higher rate of return to school (9.6 percent versus 8.5 percent for whites).

Affirmative Action

Part of the increase in the black relative wage can be attributed to the impact of government programs, particularly the enactment of the 1964 Civil Rights Act.[23] This landmark legislation prohibits employment discrimination on the basis of race and gender. Title VII of the act established the Equal Employment Opportunity Commission (EEOC) to monitor compliance. It is under this provision that costly class action suits can be initiated to force the discontinuation of discriminatory hiring practices by the targeted employers, and to compensate the affected workers for past discrimination.

The federal civil rights program was further strengthened in the 1960s by Executive Order No. 11246 and No. 11375, which prohibited discrimination by race and gender among government contractors. Federal contractors must agree "not to discriminate against any employee or applicant for employment because of race, color, religion, gender, national origin, and to take affirmative action to ensure that applicants and employees are treated during employment without regard to their race, color, gender, or national origin."

The enforcement effort to ensure compliance has been substantial. Federal contractors who have at least $50,000 worth of contracts and 50 employees must fill out an annual form where they report their total employment by occupation, race, and gender. These data can trigger "compliance reviews" that audit the contractor's employment practices and may lead to costly negotiations or litigation to change the employer's hiring behavior. It is not surprising that affirmative action programs had an almost immediate impact on hiring decisions.

[22] James P. Smith and Finis R. Welch, "Black Economic Progress after Myrdal," *Journal of Economic Literature* 27 (June 1989): 519–564; and David Card and Alan B. Krueger, "School Quality and Black–White Relative Earnings: A Direct Assessment," *Quarterly Journal of Economics* 107 (February 1992): 151–200.

[23] John J. Donohue and James J. Heckman, "Continuous versus Episodic Change: The Impact of Civil Rights Policy on the Economic Status of Blacks," *Journal of Economic Literature* 29 (December 1991): 1603–1643; Kenneth Y. Chay, "The Impact of Federal Civil Rights Policy on Black Economic Progress: Evidence from the Equal Employment Opportunity Act of 1972," *Industrial and Labor Relations Review* 51 (July 1998): 608–632; and Peter Hinrichs, "The Effects of Affirmative Action Bans on Educational Attainment, College Enrollment, and the Demographic Composition of Universities," Review of Economics and Statistics 94 (August 2012): 712–722. The literature is surveyed in Harry Holzer and David Neumark, "Assessing Affirmative Action," *Journal of Economic Literature* 38 (September 2000): 483–568.

Some of the strongest evidence is given by employment trends among manufacturing firms in South Carolina.[24] There was little change in the share of black employment in the textile industry (the main manufacturing employer in that state) between 1910 and 1964. The fraction of black employment in that industry stood at roughly 4–5 percent throughout those decades. The South Carolina textile industry, however, sold 5 percent of its output to the U.S. government, so it was clearly targeted by the executive orders. By 1970, nearly 20 percent of the workers in the industry were black.

It is tempting to interpret the rising wage of blacks after 1964 as the result of affirmative action programs, but the black wage was rising even prior to the 1960s.[25] But there is, in fact, a straightforward "back-door" way through which affirmative action increased black wages. The executive orders affect mainly large firms, and large firms tend to pay higher wages. The number of blacks employed by large firms increased substantially in the 1970s. The increasing representation of blacks in the workforce of large firms may account for about 15 percent of the increase in the black–white wage ratio over the period.[26]

The Decline in Black Labor Force Participation

Despite the increase in the wage of black men in recent decades, the labor force participation rate of this group fell precipitously. Figure 9-9 illustrates this important trend. In the mid-1950s, 85 percent of both black and white men were in the labor force. By 2015, the gap between the black and white participation rates was 6 percentage points.

Suppose that black workers who drop out of the labor market are relatively low-skill. This would imply that the average wage of *working* blacks would rise over time, simply because blacks at the lower tail of the wage distribution are no longer included in the calculations.[27] In other words, the observed increase in the black wage need not indicate an improvement in black opportunities. It might instead be indicating that the least-skilled blacks are no longer working.

Figure 9-10 illustrates a wage distribution for blacks and shows how a decline in labor force participation can lead to an increase in the mean black wage. As we saw in the labor supply chapter, persons decide whether to work or not by comparing the reservation wage with the market wage. The reservation wage of black workers is initially given by w_1^*, so that all blacks who command a wage greater than w_1^* will work. The mean wage observed in the sample of workers is then given by $\overline{w}_1$.

Suppose that for some reason, such as the introduction of large-scale public assistance programs in the 1960s, the reservation wage of blacks increased to w_2^*. This increase in

[24] James J. Heckman and Brook S. Payner, "Determining the Impact of Federal Antidiscrimination Policy on the Economic Status of Blacks: A Study of South Carolina," *American Economic Review* 79 (March 1989): 138–177.

[25] Richard B. Freeman, "Changes in the Labor Market for Black Americans," *Brookings Papers on Economic Activity* 20 (1973): 67–120; and Harry Holzer and David Neumark, "Are Affirmative Action Hires Less Qualified? Evidence from Employer–Employee Data on New Hires," *Journal of Labor Economics* 17 (July 1999): 534–569.

[26] William J. Cvarrington, Kristin McCue, and Brooks Pierce, "Using Establishment Size to Measure the Impact of Title VII and Affirmative Action," *Journal of Human Resources* 35 (Summer 2000): 503–523.

[27] Richard J. Butler and James J. Heckman, "The Government's Impact on the Labor Market Status of Black Americans: A Critical Review," in Leonard J. Hausman, editor, *Equal Rights and Industrial Relations,* Madison, WI: Industrial Relations Research Association, 1977.

FIGURE 9-9 Male Labor Force Participation Rates, by Race, 1955–2015

Sources: U.S. Bureau of the Census, *Historical Statistics of the United States, Colonial Times to 1970,* Washington, D.C.: Government Printing Office, 1975; Bureau of Labor Statistics, Labor Force Characteristics by Race and Ethnicity, 2015, Table 4; www.bls.gov/opub/reports/race-and-ethnicity/2015/home.htm.

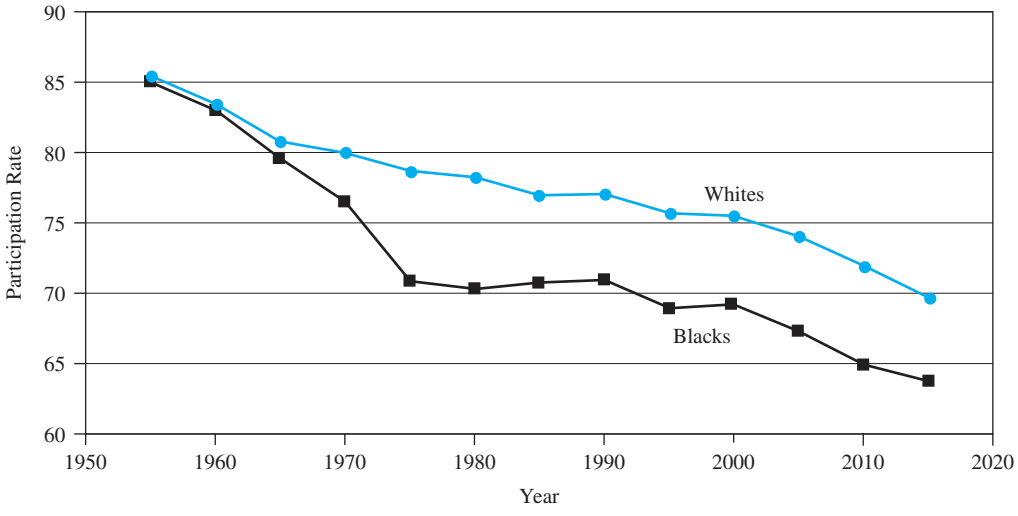
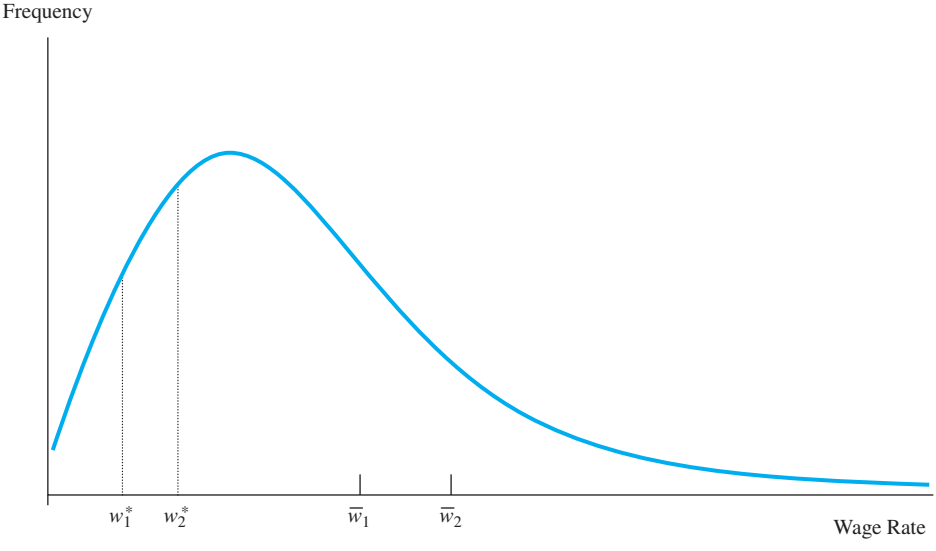


FIGURE 9-10 Decline in Black Labor Force Participation of Blacks and the Average Black Wage

If the reservation wage of blacks is w_1^*, the mean wage observed among black workers is $\overline{w}_1$. Suppose the reservation wage rises to w_2^*. The black labor force participation rate falls, and the mean wage observed among blacks still working rises to $\overline{w}_2$.



the reservation wage reduces the number of black workers but increases the average wage in the subset of blacks who do work to $\overline{w}_2$. The upward drift in the relative wage of black men observed since the 1960s might then be an illusion produced by sample selection bias.

Some studies conclude that about a third of the improvement in the relative black wage between 1969 and 1989 can be attributed to the declining labor force participation of the black population.[28]

Unobserved Skill Differences

The empirical measure of discrimination based on the Oaxaca–Blinder decomposition effectively measures the wage gap between black and white workers who are "observationally equivalent" in the sense that they have the same number of years of schooling, the same amount of labor market experience, live in the same region, work in the same industry and occupation, and so on. But there may well be other skill differences between the groups that are not observed, and they may account for part of the wage differential that the Oaxaca–Blinder decomposition labels "discrimination."

Some studies investigate if such unobserved skill differences exist. These studies often use a particular measure of skills: the test score in the Armed Forces Qualification Test (AFQT). As the name implies, this standard test is given to all recruits in the U.S. military. The test was also administered to a randomly chosen sample of young persons in the 1980s (regardless of whether they planned to be in the military).

There are sizable racial differences in the AFQT score; blacks tend to have lower scores than whites. An influential study documented that the racial difference in the AFQT score accounts for practically the entire wage gap between young black and white workers.[29] Even though the *actual* black–white wage ratio is about 0.8 for these young workers, the *adjusted* black–white wage ratio jumps to about 0.95 once we control for differences in AFQT scores between the groups. Put differently, although the typical black worker earns 20 percent less than a white worker, the typical black worker earns only 5 percent less than a white worker with the same AFQT score. In short, much of the wage gap between young black and white workers disappears once the wage data are adjusted for the difference in AFQT scores.

Although there is little doubt about the validity of the evidence, the interpretation is not clear. What exactly is the AFQT score measuring? The AFQT score is not a straightforward measure of innate ability. Persons who have more schooling or go to better schools have higher AFQT scores. The score in this particular test, therefore, partly measures skills that were acquired prior to the time of labor market entry. Because much of the wage gap between young blacks and whites can be attributed to skill differences, the evidence seems to suggest that the role played by labor market discrimination may have diminished substantially in recent decades.

[28] Chinhui Juhn, "Labor Market Dropouts and Trends in Black and White Wages," *Industrial and Labor Relations Review* 56 (July 2003): 643–662; Amitabh Chandra, "Labor Market Dropouts and the Racial Wage Gap: 1940–1990," *American Economic Review* 90 (May 2000): 333–338; and Derek Neal, "The Measured Black–White Wage Gap among Women Is Too Small," *Journal of Political Economy* 112 (February 2004): S1–S28.

[29] Derek A. Neal and William R. Johnson, "The Role of Premarket Factors in Black–White Wage Differences," *Journal of Political Economy* 104 (October 1996): 869–895; see also Kevin Lang and Michael Manove, "Education and Labor Market Discrimination," *American Economic Revie*w 101 (June 2011): 1467–1496.

9-10 The Relative Wage of Hispanic and Asians

The resurgence of large-scale immigration in the past few decades greatly altered the racial and ethnic mix of the U.S. population and sparked interest in documenting the wage determination process for other racial and ethnic groups.

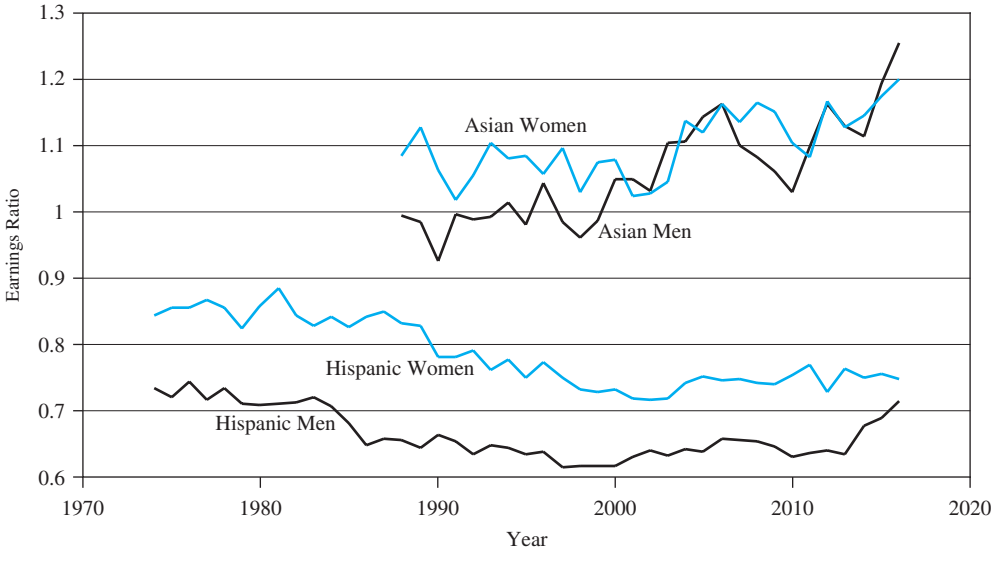
The growth of the Hispanic population is astounding. In 1980, Hispanics made up only 6.4 percent of the population, while blacks made up 11.7 percent. By 2016, Hispanics had become the largest minority group, comprising 17.3 percent of the population, and the proportion of blacks had risen slightly to 13.3 percent.

Figure 9-11 illustrates the trend in the Hispanic-white earnings ratio. This ratio declined between 1980 and 2010 for both Hispanic men and Hispanic women (although the male wage ratio increased sharply since 2010). Because the number of Hispanic immigrants grew substantially in the past few decades, the long-term trends in the relative wage of Hispanics may reflect the changing composition of the Hispanic population, rather than a growing disadvantage to a fixed group of workers.

A careful study of the sizable wage gap between men of Mexican origin and non-Hispanic whites concludes that over three-quarters of the gap can be attributed to differences in observable skill measures, particularly educational attainment.[30] In other words, the largest group of Hispanic Americans earns less not because of their "Hispanic-ness," but mainly because they are less skilled.

FIGURE 9-11 Trend in Relative Earnings of Hispanics and Asians, 1974–2016

Source: U.S. Bureau of the Census, "Historical Income Tables—People," Table P-38. "Full-Time Year-Round Workers by Median Earnings and Sex," https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-people.html. The earnings refer to the median earnings of full-time, full-year workers aged 15 or above. The denominator in the ratios gives the earnings of white men or women, respectively.



[30] Stephen J. Trejo, "Why Do Mexican Americans Earn Low Wages?" *Journal of Political Economy* 105 (December 1997): 1235–1268.

Theory at Work

SHADES OF BLACK

There is a lot of variation in skin tone among African-Americans. There are also substantial differences in economic outcomes by skin tone within the black population. Typically, African-Americans with a lighter skin tone have more education and earn more than African-Americans with darker tones.

In one survey, for example, the average white man earned $15.94 an hour. Black respondents were asked to classify their skin tone in one of three categories: light black, medium black, or dark black. Those who indicated they had a light skin tone earned $14.42, those indicating a medium tone earned $13.23, and those indicating a dark tone earned $11.72. Moreover, these wage gaps remained even after controlling for observed differences in socioeconomic characteristics, including education and age. The typical light-skin black earned roughly the same as a comparably skilled white man, while dark-skin blacks earned about 10 percent less.

Interestingly, skin color has equally strong effects on earnings in the immigrant population, even if we compare immigrants originating in the same country and belonging to the same race. Light-skinned immigrants earn 17 percent more than dark-skinned immigrants.

There are many potential explanations for the link between skin tone and socioeconomic outcomes. For instance, it may well be that persons with lighter skin tones are perceived to be more attractive, and we know that "beauty" leads to better labor market outcomes. Or perhaps many employers find it difficult to pigeon-hole light-skinned African-Americans or immigrants into a specific race group, particularly in an increasingly multiracial society.

And it is not only the visual aspects of "blackness" that matter. Speech that can be distinctly identified as belonging to a black speaker is penalized in the labor market. Blacks whose speech cannot be differentiated from "white speech" earn essentially the same as comparably skilled whites, while blacks whose speech can be distinctly identified earn about 12 percent less.

Sources: Arthur H. Goldsmith, Darrick Hamilton, and William Darity, Jr., "From Dark to Light: Skin Color and Wages Among African-Americans," *Journal of Human Resources* 62 (Fall 2007): 701–738; Joni Hersch, "Profiling the New Immigrant Worker: The Effects of Skin Color and Height," *Journal of Labor Economics* 26 (April 2008): 345–386; and Jeffrey Grogger, "Speech Patterns and Racial Wage Inequality," *Journal of Human Resources* 46 (Winter 2011): 1–25.

The Asian population also grew rapidly in recent decades. In 1980, only 1.5 percent of the population was of Asian ancestry. By 2016, the Asian share had almost quadrupled, to 5.7 percent.

Figure 9-11 also shows the trend in the Asian–White earnings ratio. These ratios typically hover between 1.0 and 1.2 for both men and women. In other words, the typical Asian-American has a wage advantage over whites. Much of this advantage can be attributed to the fact that many Asian workers have relatively high levels of educational attainment.

It is important to emphasize that the observed economic performance of particular groups can be contaminated by the fact that workers *choose* which group to "bond with" when they respond in census-type surveys. For example, the frequency of intermarriage in the Mexican-American population gives the offspring of these couplings a choice: Do they self-identify as Mexican or even as Hispanic? It turns out that the self-selected group of third- and higher-order generation workers who choose to report their Mexican ancestry is not a random sample of the relevant population. Instead, it consists of the persons who have Mexican ancestry *and* who have relatively low levels of economic performance.[31] This self-selection greatly complicates the interpretation of trends in the socioeconomic status of some ethnic or racial groups.

[31] Brian Duncan and Stephen J. Trejo, "Intermarriage and the Intergenerational Transmission of Ethnic Identity and Human Capital for Mexican Americans," *Journal of Labor Economics* 29 (April 2011): 195–227.

9-11 Policy Application: The Male–Female Wage Gap

The oldest documented wage differential between men and women dates back to the days of the Old Testament:

> The Lord spoke to Moses and said, Speak to the Israelites in these words. When a man makes a special vow to the Lord which requires your valuation of living persons, a male between twenty and fifty years old shall be valued at fifty silver shekels, that is shekels by the sacred standard. If it is a female, she shall be valued at thirty shekels. (Leviticus 27:1–4)

By 1999, the Biblical female–male wage ratio of 0.6 had increased to 0.78 in the Netherlands, 0.76 in the United Kingdom, and 0.72 in the United States.[32] The study of male–female wage differentials focuses on a simple question: What explains the existence and persistence of this wage gap?

The Gender Wage Gap and Labor Market Experience

There is an ongoing debate over how much of the sizable wage gap between men and women remains after we control for differences in socioeconomic characteristics between the two groups.

As Table 9-4 shows, women earned about 28.6 percent less than men in 1995. The Oaxaca–Blinder decomposition reveals that differences in education, age, and region of residence account for only a trivial part of that wage gap. In fact, even after adjusting for occupation and industry, the wage gap attributable to discrimination is about 21 percent.

It is not surprising that differences in educational attainment, region of residence, and age fail to explain the bulk of the gender wage gap. On average, men and women have roughly the same level of schooling, are about the same age, and live in the same towns. Discrimination—in the Oaxaca–Blinder sense—must then accounts for the bulk of the wage difference.

The exercise reported in Table 9-4, however, ignores a key determinant of female earnings: Men and women with the same age might have very different labor market histories.

It is not uncommon for some women to drop out of the labor market (or limit their work activities) during the child-raising years. A woman's career path, for example, might consist of a spell of employment after she completes school, followed by a spell in the household sector or with curtailed effort in the labor market, and then a full-time return to employment after the child-raising years.[33]

This discontinuity in labor market attachment would generate a gender wage gap.[34] The argument can be easily stated. Human capital is more profitable the longer the payoff period. Consider the payoffs to investments made by new labor market entrants. Because the vast majority of men expect to work throughout their entire lives, the human capital acquired by men

[32] Claudia Olivetti and Barbara Petrongolo, "Unequal Pay or Unequal Employment? A Cross-Country Analysis of Gender Gaps," *Journal of Labor Economics* 26 (October 2008): 621–654.

[33] Francine D. Blau and Lawrence F. Kahn, "Swimming Upstream: Trends in the Gender Wage Differential in the 1980s," *Journal of Labor Economics* 15 (January 1997): 1–42; June O'Neill and Solomon Polachek, "Why the Gender Gap in Wages Narrowed in the 1980s," *Journal of Labor Economics* 11 (January 1993, Part 1): 205–228; Anne M. Hill and June E. O'Neill, "Intercohort Change in Women's Labor Market Status," *Research in Labor Economics* 13 (1992): 215–286.

[34] Jacob Mincer and Solomon W. Polachek, "Family Investments in Human Capital: Earnings of Women," *Journal of Political Economy* 82 (March 1974 Supplement): S76–S108.

TABLE 9-4 The Oaxaca Decomposition of the Female–Male Wage Differential, 1995

Source: Joseph G. Altonji and Rebecca M. Blank, "Race and Gender in the Labor Market," in Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics,* vol. 3C, Amsterdam: Elsevier, 1999, Table 5. The log wage differential between any two groups can be interpreted as being approximately equal to the percentage wage differential between the groups.

	Controls for Differences in Education, Age, Sex, and Region of Residence	Controls for Differences in Education, Age, Sex, Region of Residence, *and* Occupation and Industry
Raw log wage differential	–0.286	–0.286
Due to differences in skills	–0.008	–0.076
Due to discrimination	–0.279	–0.211

has a long payoff period. Some women expect to spend time to the household sector, shortening the payoff period and reducing the returns on the investment. This difference in work histories would imply that women, on average, will acquire less postschool human capital.

Moreover, those skills might depreciate somewhat during the years when a woman is more actively engaged in household production. Skills that are not used or kept up-to-date can be forgotten or become obsolete. The value of a woman's human capital stock, therefore, would be reduced by her intermittent labor market attachment.

The discontinuity in female labor supply over the life cycle then generates a gender wage gap for two distinct reasons. First, it creates a wage gap because men tend to acquire more human capital. Second, the household production years further increase the wage gap because a woman's skills may depreciate during that period.

Some of the evidence is consistent with the human capital hypothesis, although there is disagreement over how much of the gender wage gap can be explained by the difference in labor market histories.[35] A clear illustration of the role of labor market experience is provided by an analysis of graduates from the University of Michigan law school.[36] Fifteen years after graduation, male attorneys earned $141,000 annually as compared to $86,000 for female attorneys. But about two-thirds of the gender wage gap was attributable to differences in the work histories of male and female attorneys. If a female attorney decided to work part-time just for 3 years to care for her children, as many women did, her earnings were *permanently* reduced by 17 percent. This wage reduction might occur because a full-time attachment to the profession enlarges the attorney's client base and increases opportunities for future career advancement.

A related study of the earnings of MBA graduates from the Booth School of Business at the University of Chicago reports similar findings.[37] Although there is no gender wage gap when these graduates begin their careers, the earnings of men and women begin to diverge

[35] Steven H. Sandell and David Shapiro, "The Theory of Human Capital and the Earnings of Women: A Reexamination of the Evidence," *Journal of Human Resources* 13 (Winter 1978): 103–117; Donald Cox, "Panel Estimates of the Effects of Career Interruptions on the Earnings of Women," *Economic Inquiry* 22 (July 1984): 386–403; and Per-Anders Edin and Magnus Gustavsson, "Time Out of Work and Skill Depreciation," *Industrial and Labor Relations Review* 61 (January 2008): 163–180.

[36] Robert G. Wood, Mary E. Corcoran, and Paul N. Courant, "Pay Differences among the Highly Paid: The Male–Female Earnings Gap in Lawyers' Salaries," *Journal of Labor Economics* 11 (July 1993): 417–441.

[37] Marianne Bertrand, Claudia Goldin, and Lawrence F. Katz, "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors," *American Economic Journal: Applied Economics* 2 (July 2010): 228–255.

quickly. Men earn 30 percent more after 5 years and 60 percent more after 15 years. The gender difference in accumulating labor market experience again explains much of the wage gap. Female MBA graduates are much more likely to have interrupted work careers: Over 40 percent of women had a career interruption (defined as a 6-month period out of the workforce) within 15 years after graduation, as compared to only 10 percent of men. In addition, women are far more likely to work part-time. Almost a quarter of female MBA graduates work part-time after 15 years, as compared to only 4 percent of male MBA graduates.

In short, although there is disagreement over the extent to which the human capital hypothesis can account for the gender wage gap, differences in human capital accumulation between men and women do matter.

Despite its influence, the human capital story faces an important conceptual obstacle. The hypothesis asserts that because women might work fewer years during the life cycle, they will invest less in on-the-job training and other forms of human capital, and they will earn lower wages as a result. But a low female wage, perhaps due to discrimination, reduces incentives to work. Did a woman's weaker work attachment lead to lower wage rates through reduced human capital investments? Or did the lower wage rate lead to a weaker work attachment?[38]

Occupational Crowding

There is a lot of occupational segregation by gender. Fewer than 6 percent of aircraft mechanics are women, but nearly 98 percent of kindergarten teachers and receptionists are women.[39] A discrimination-based explanation of this difference, known as the **occupational crowding** hypothesis, argues that women are intentionally segregated into particular occupations.[40]

This crowding need not be the result of employer discrimination. It may instead reflect the social climate where young women are taught that some occupations "are not for girls" and get channeled into "appropriate" jobs. The crowding of women into a relatively small number of occupations reduces the wage of so-called female jobs and generates a gender wage gap.

A woman working in an occupation where at least 75 percent of the coworkers are women earns about 14 percent less than an equally skilled woman working in an occupation where more than 75 percent of the coworkers are men. Moreover, a *man* working in an occupation that is predominantly female also earns 14 percent less than a man working in an occupation that is predominantly male. In short, it is the "femaleness" of the job that leads to lower wages, regardless of whether the worker employed in that job is a man or a woman.[41]

A blatant example of occupational crowding is given by the so-called marriage bars that restricted the employment of married women in some sectors of the U.S. labor market before 1950.[42] The marriage bars prohibited married women from working, primarily in

[38] Reuben Gronau, "Sex-Related Wage Differentials and Women's Interrupted Labor Careers—The Chicken or the Egg," *Journal of Labor Economics* 6 (July 1988): 277–301; and David Neumark, "Sex Discrimination and Women's Labor Market Outcomes," *Journal of Human Resources* 30 (Fall 1995): 713–740.

[39] Bureau of Labor Statistics, "Employed Persons by Detailed Occupation, Sex, Race, and Hispanic or Latino Ethnicity"; see www.bls.gov/cps/cpsaat11.htm.

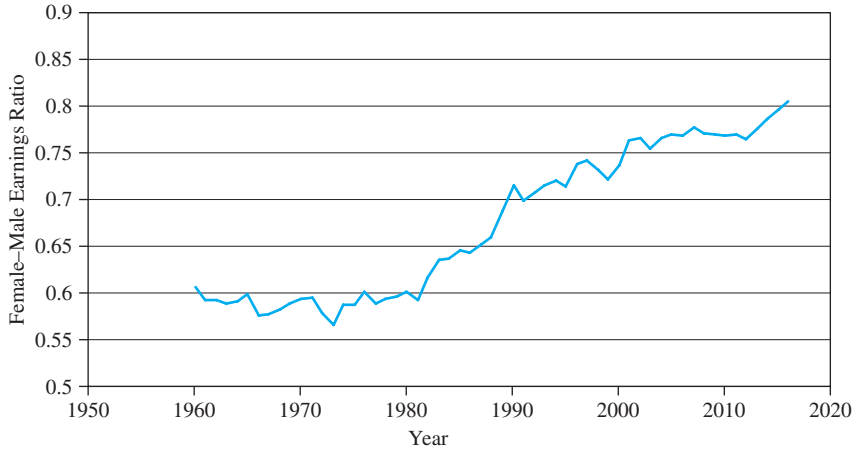
[40] Barbara F. Bergmann, "The Effect on White Incomes of Discrimination in Employment," *Journal of Political Economy* 79 (March/April 1971): 294–313.

[41] David A. Macpherson and Barry T. Hirsch, "Wages and Gender Composition: Why Do Women's Jobs Pay Less?" *Journal of Labor Economics* 13 (July 1995): 426–471, Table 4.

[42] Claudia Goldin, *Understanding the Gender Gap: An Economic History of American Women,* New York: Oxford University Press, 1990, pp. 159–179.

FIGURE 9-12 Trend in Female–Male Earnings Ratio, 1960–2016

Source: U.S. Bureau of the Census, "Historical Income Tables—People," Table P-40, "Women's Earnings as a Percentage of Men's Earnings by Race and Hispanic Origin." The earnings refer to the median earnings of full-time, full-year workers aged 15 or above.



teaching and clerical jobs. Married women looking for work in these occupations would not be hired, and single women working in these jobs were often fired once they married. Marriage bars, however, did not affect the hiring of women in manufacturing, or employed as waitresses and domestic servants. The marriage bars, therefore, served as a device driving well-educated women out of the labor market or crowding them into lower paying jobs.

The human capital model provides an alternative, "supply-side" explanation of why women *choose* certain occupations and avoid others. Some occupations (for example, child care workers) require skills that do not have to be updated frequently, whereas other occupations (such as physicists) require skills that must be updated constantly. Women who wish to maximize the present value of lifetime earnings will not enter occupations where their skills might depreciate during the years spent in the household sector. Although there is some evidence indicating that women tend to choose occupations where skills are less likely to depreciate, it is also the case that women's career decisions are influenced by many other factors, including the gender of their mentors.[43]

The Trend in the Female–Male Wage Ratio

Figure 9-12 illustrates the historical trend in the female–male earnings ratio in the U.S. labor market. Among persons employed full-time year-round, the ratio hovered around 0.6 between 1960 and 1980. Beginning in the early 1980s, however, the female–male earnings ratio began to increase rapidly and stood at 0.81 in 2016.

As we have seen, wage inequality increased since the 1980s. This increase in wage inequality might have been expected to further widen the wage gap between men and women. As

[43] Solomon W. Polachek, "Occupational Self-Selection: A Human Capital Approach to Sex Differences in Occupational Structure," *Review of Economics and Statistics* 63 (February 1981): 60–69; Florian Hoffmann and Philip Oreopoulos, "A Professor Like Me: The Influence of Instructor Gender on College Achievement," *Journal of Human Resources* 44 (Spring 2009): 479–494; and Scott E. Carrell, Marianne E. Page, and James E. West, "Sex and Science: How Professor Gender Perpetuates the Gender Gap," *Quarterly Journal of Economics* 125 (August 2010): 1101–1144.

Figure 9-12 reveals, however, women's economic status improved rapidly in the past few decades. There is evidence suggesting that part of the rise in the female wage can be attributed to an increase in the labor market experience of women, as the intermittent labor market attachment of earlier generations is slowly replaced by a more permanent attachment to the workforce.[44]

The data also suggest that affirmative action had little impact on the employment prospects of white women, but a sizable impact on black women. For example, federal contractors employed 28 percent of all working white women in 1970, but only 30 percent in 1980. In contrast, federal contractors employed 35 percent of black women in 1970, but almost half of all black women by 1980. Affirmative action thus induced a huge increase in the demand for black women by these firms.[45]

It is worth noting that the interpretation of trends in the female–male earnings ratio is complicated by the rapid rise in the female labor force participation rate. This implies that the average female wage is being calculated in very different samples of working women in different decades.

Suppose, for instance, that the new labor market entrants in the 1960s and 1970s had lower earnings potential than women already working. Adding the low-wage persons to the sample of female workers would tend to flatten the trend in the earnings ratio in those decades, potentially masking any improvement in female earnings. Conversely, perhaps it is the women with the highest earnings potential who are more likely to enter the labor force since 1990. This would lead to an increase in the relative female wage, but that increase could be due entirely to changes in sample composition. There is, in fact, evidence that adjusting for the increasing self-selection of high-skill women into the labor force flattens the steep post-1980 improvement documented in Figure 9-12.[46]

Uber and the Gender Wage Gap

The past decade witnessed the explosive growth of what is being called the "gig economy," where firms offer short-term employment arrangements to independent contractors. In a typical setting, a consumer wants to buy a service. A digital platform connects the consumer to someone who is willing to provide that service for a fee. Because all the information involving any particular transaction is instantaneously recorded, the gig economy generates immense amounts of "big data" that can be analyzed in many new ways, and that may provide important insights into how labor markets work.

An analysis of the gender wage differential at the ridesharing service Uber shows the promise of this new type of research.[47] The Uber software platform links consumers who want to

[44] Francine D. Blau, and Lawrence M. Kahn, "Swimming Upstream: Trends in the Gender Wage Differential in the 1980s," *Journal of Labor Economics,* 15 (1997), 1–42; June O'Neill and Solomon Polachek, "Why the Gender Gap in Wages Narrowed in the 1980s," *Journal of Labor Economics* 11 (January 1993, Part 1): 205–228.

[45] Smith and Ward, "Women in the Labor Market and in the Family," *Journal of Economic Perspectives* 3 (Winter 1989): p. 15.

[46] Casey B. Mulligan and Yona Rubinstein, "Selection, Investment, and Women's Relative Wages over Time," *Quarterly Journal of Economics* 123 (August 2008): 1061–1110.

[47] Cody Cook, Rebecca Diamond, Jonathan Hall, John A. List, and Paul Oyer, "The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers," National Bureau of Economic Research Working Paper No. 24372, June 2018.

get somewhere with a driver willing to drive them. The rider requests a ride through a smartphone app. That request is sent to an Uber driver who happens to be nearby. The driver can either accept or reject the request after seeing the rider's pickup location. If the driver declines to provide the ride, the request is sent to another Uber driver who is also nearby. The computer algorithm that sends the rider's request to a particular driver does not know the gender of the driver; it only knows that the driver happens to be near the rider's pickup location.

The appeal of this work arrangement to Uber drivers is obvious: drivers have full discretion about where and when to work. They choose which locations to service and they set their own hours.

Uber drivers are paid according to a formula, which includes a base fare for the ride, plus a premium that depends on how long the ride takes and on the distance driven. Three is also a "surge" multiplier that depends on the demand for rides at that particular time and in that particular location. Notably, the formula does not take into account any driver characteristics, such as years of experience at Uber or the number of hours that the driver works.

Gender plays no role in the allocation of riders to drivers, and it does not enter the formula that determines the driver's compensation for a particular ride. Nevertheless, an analysis of nearly 2 million Uber drivers, involving nearly 750 million rides, reveals that the hourly wage rate of the typical male driver is 7 percent higher than that of the typical female driver.

It turns out that three variables explain the entire 7 percent wage gap, and these variables reflect the importance of supply factors and driver choices in generating gender wage differentials.

First, there are certain times of the day, as well as certain locations, where it is more profitable to drive. Men often choose their driving hours and locations to take advantage of these differences, maximizing their earnings potential. Second, work experience matters. Male drivers are more likely to have been with Uber for at least 2 years and work more hours per week. The accumulation of driving experience provides valuable "on-the-job training" that informs drivers about which rides are more profitable. For instance, an experienced driver is more likely to predict when surge premiums will be going into effect. Finally, men drive 4 percent faster. Although the pay formula depends both on the number of minutes per ride and on the distance driven, driving faster typically lets the driver get to the next ride quicker, accumulating more miles and further increasing their earnings.

Summary

- Taste discrimination affects the employer's hiring decision because prejudice blinds the employer to the actual monetary cost of hiring a particular worker. An employer who discriminates will act as if the cost of hiring a black or female worker exceeds the actual cost.
- If black and white workers are perfect substitutes in the production process, employer discrimination leads to the segregation of black and white workers in the firm and to unequal pay for equal work. The firm's discriminatory behavior also reduces profits.

- Employee discrimination leads to segregation of black and white workers but does not create a wage differential between the two groups. Customer discrimination might create a wage differential between black and white workers if employers cannot place blacks in positions where they have little contact with customers.
- Wage differentials by race, ethnicity, and gender can arise even if employers are not prejudiced. When firms do not have complete information on a particular worker's productivity, they might use aggregate characteristics of the group as an indicator of the worker's productivity. Statistical discrimination leads to differential treatment of equally skilled workers belonging to different groups.
- The Oaxaca–Blinder decomposition defines discrimination as the wage gap between workers who belong to different groups (such as race or gender), but are observationally equivalent (that is, they have the same education, experience, and other socioeconomic characteristics). If the empirical exercise does not control for all the dimensions in which skills might differ across workers, the measure of discrimination does not isolate the impact of prejudice or statistical discrimination on the wage of minorities and women.
- The wage ratio between black and white workers in the United States has risen significantly in the past few decades. In 1995, whites earned about 24 percent more than blacks, and about half of this wage gap could be attributed to differences in observable skills.
- The wage ratio between female and male workers in the United States has risen significantly since 1980. Part of the gender wage gap can be attributed to the fact that women, on average, have less labor market experience than men.

Key Concepts

customer discrimination, *302*
discrimination coefficient, *301*
employee discrimination, *302*
employer discrimination, *302*
nepotism, *301*
Oaxaca–Blinder decomposition, *318*
occupational crowding, *330*
statistical discrimination, *312*
taste discrimination, *301*

Review Questions

1. What is the discrimination coefficient?
2. Discuss the implications of employer discrimination for the hiring decisions of the firm, for the profitability of the firm, and for the black–white wage ratio in the labor market.
3. Can the wage of blacks exceed that of whites if most firms in the labor market discriminate against blacks?
4. Derive the implications of employee discrimination for the employment decisions of firms and for the black–white wage differential.

5. Discuss the implications of customer discrimination for the employment decisions of firms and for the black–white wage differential.
6. What is statistical discrimination? Why do employers use group membership as an indicator of a worker's productivity? What is the impact of statistical discrimination on the wage of the affected workers? Must statistical discrimination reduce the average wage of blacks or women?
7. Derive the Oaxaca measure of discrimination. Does this statistic truly measure the impact of discrimination on the relative wage of the affected groups?
8. Discuss the factors that might explain why the black–white wage ratio rose significantly in the past few decades.
9. Discuss why part of the female–male wage differential might be attributable to "supply-side" factors, such as a woman's decision to work and acquire human capital.

Problems

9-1. Feeling that local firms follow discriminatory hiring practices, a nonprofit firm conducts the following experiment. It has 200 white individuals and 200 black individuals, all of whom are similar in age, experience, and education, apply for local retail jobs. Each individual applies to two jobs, one in a predominantly black part of town and one in a predominantly white part of town. Of the white applicants, 120 are offered jobs in the white part of town while only 80 are offered jobs in the black part of town. Meanwhile, 90 of the black applicants are offered jobs in the black part of town while only 50 are offered jobs in the white part of town. Using a difference-in-differences estimator, do you find evidence of discriminatory hiring practices? If there is evidence of discrimination, is it appropriate to conclude that all employers in the white part of town are discriminatory?

9-2. Suppose black and white workers are complements in that the marginal product of whites increases when more blacks are hired. Suppose also that white workers do not like working alongside black workers. Under what conditions will this employee discrimination lead to a segregated workforce? Under what conditions will it not?

9-3. Suppose a restaurant hires only women to wait on tables, and only men to cook the food and clean the dishes. Is this most likely to be indicative of employer, employee, consumer, or statistical discrimination?

9-4. A firm's production function is given by $Q = 40\ ln(E_W + E_B + 1)$ where E_W and E_B are the number of whites and blacks employed by the firm, respectively. From this it can be shown that the marginal product of labor is

$$MP_E = \frac{40}{E_W + E_B + 1}.$$

Suppose the market wage for blacks is \$50, the market wage for whites is \$100, and the price of each unit of output is \$20.

(a) How many workers of each race would a nondiscriminating firm hire? How much profit is earned if there are no other costs?

(b) How many workers of each race would a firm with a discrimination coefficient of 0.6 against blacks hire? How much profit is earned if there are no other costs?

(c) How many workers of each race would a firm with a discrimination coefficient of 1.2 against blacks hire? How much profit is earned if there are no other costs?

9-5. Suppose years of schooling, *s,* is the only variable that affects earnings. The equations for the weekly salaries of male and female workers are given by:

$$w_m = 500 + 100s$$

and

$$w_f = 300 + 75s.$$

On average, men have 14 years of schooling and women have 12 years of schooling.

(a) What is the male-female wage differential in the labor market?

(b) Using the Oaxaca–Blinder decomposition, calculate how much of this wage differential could be due to discrimination?

9-6. Suppose the firm's production function is given by

$$\boldsymbol{q = 10\sqrt{E_w + E_b}},$$

where E_w and E_b are the number of whites and blacks employed by the firm, respectively. It can be shown that the marginal product of labor is then

$$MP_E = \frac{5}{\sqrt{\boldsymbol{E_w + E_b}}}.$$

Suppose the market wage for black workers is \$10, the market wage for whites is \$20, and the price of each unit of output is \$100.

(a) How many workers would a firm hire if it does not discriminate? How much profit does this nondiscriminatory firm earn if there are no other costs?

(b) Consider a firm that discriminates against blacks with a discrimination coefficient of 0.25. How many workers does this firm hire? How much profit does it earn?

(c) Finally, consider a firm that has a discrimination coefficient equal to 1.25. How many workers does this firm hire? How much profit does it earn?

9-7. Cindy, a tenured, full professor of French literature at a large university, is paid \$60,000. The university reports median salaries by gender and rank as a new initiative on faculty compensation. From reading the report, Cindy learns that she is paid \$20,000 below the median for male, tenured, full professors. She is also paid \$12,000 below the median for female, tenured, full professors. What factors might explain Cindy's position in the wage distribution? Why might or might not the university be engaged in gender discrimination?

9-8. Consider the following log-wage regression results for women (W) and men (M) where wages are predicted by schooling (S) and age (A).

$$w_W = 2.19 + 0.075\,S_W + 0.023\,A_W \text{ and } w_M = 2.42 + 0.072\,S_M + 0.017\,A_M.$$

Sample means for the variables by gender are: women average a logged wage of 3.83, 13.5 years of schooling, and 41.2 years-old; men average a logged wage of

3.92, 13.2 years of schooling, and 44.3 years old. Decompose the raw difference in average logged wages using the Oaxaca–Blinder decomposition. Specifically, decompose the raw difference into the portion due to differences in schooling, differences in age, and the portion left unexplained, possibly due to gender discrimination.

9-9. Each employer faces competitive weekly wages of \$2,000 for whites and \$1,400 for blacks. Suppose employers under-value the efforts/skills of blacks in the production process. In particular, every firm is associated with a discrimination coefficient, d where $0 \leq d \leq 1$. In particular, although a firm's actual production function is $Q = 10(E_W + E_B)$, the firm manager acts as if its production function is $Q = 10E_W + 10(1 - d)E_B$. Every firm sells its output at a constant price of \$240 per unit up to a weekly total of 150 units of output. No firm can sell more than 150 units of output without reducing its price to \$0.

(a) What is the value of the marginal product of each white worker?

(b) What is the value of the marginal product of each black worker?

(c) Describe the employment decision made by firms for which $d = 0.2$ and $d = 0.8$, respectively.

(d) For what value(s) of d is a firm willing to hire blacks and whites?

9-10. After controlling for age and education, it is found that the average woman earns \$0.80 for every \$1.00 earned by the average man. After controlling for occupation to control for compensating differentials (that is, maybe men accept riskier or more stressful jobs than women, and therefore are paid more), the average woman earns \$0.92 for every \$1.00 earned by the average man. The conclusion is made that occupational choice reduces the wage gap 12 cents and discrimination is left to explain the remaining 8 cents.

(a) Explain why discrimination may explain more than 8 cents of the 20-cent differential (and occupational choice may explain less than 12 cents of the differential).

(b) Explain why discrimination may explain less than 8 cents of the 20-cent differential.

9-11. Consider a town with 10 percent blacks (and the remainder is white). Because blacks are more likely to work the night shifts, 20 percent of all cars driven at night are driven by blacks. One out of every twenty people driving at night is drunk, regardless of race. Persons who are not drunk never swerve their cars, but 10 percent of all drunk drivers, regardless of race, swerve their cars. On a typical night, 5,000 cars are observed by the police force.

(a) What percent of blacks driving at night are driving drunk? What percent of whites driving at night are driving drunk?

(b) Of the 5,000 cars observed, how many are driven by blacks? How many of these cars are driven by a drunk? Of the 5,000 cars observed at night, how many are driven by whites? How many of these cars are driven by a drunk? What percent of nighttime drunk drivers are black?

(c) The police chief believes the drunk-driving problem is mainly due to black drunk drivers. He orders his policemen to pull over all swerving cars *and*

one in every two nonswerving cars that is driven by a black person. The driver of a nonswerving car is then given a breathalyzer test that is 100 percent accurate in diagnosing drunk driving. Under this enforcement scheme, what percent of people arrested for drunk driving will be black?

9-12. Suppose 100 men and 100 women graduate from high school. After high school, each can work in a low-skill job and earn $200,000 over his or her lifetime, or each can pay $50,000 and go to college. College graduates are given a test. If someone passes the test, he or she is hired for a high-skill job paying lifetime earnings of $300,000. Any college graduate who fails the test, however, is relegated to a low-skill job. Academic performance in high school gives each person some idea of how he or she will do on the test if they go to college. In particular, each person's GPA, call it x, is an "ability score" ranging from 0.01 to 1.00. With probability x, the person will pass the test if he or she attends college. Upon graduating high school, there is one man with $x = 0.01$, one with $x = 0.02$, and so on up to $x = 1.00$. Likewise, there is one woman with $x = 0.01$, one with $x = 0.02$, and so on up to $x = 1.00$.

(a) Persons attend college only if the expected lifetime payoff from attending college is higher than that of not attending college. Which men and which women will attend college? What is the expected pass rate of men who take the test? What is the expected pass rate of women who take the test?

(b) Suppose policymakers feel not enough women are attending college, so they take actions that reduce the cost of college for women to $10,000. Which women will now attend college? What is the expected pass rate of women who take the test?

9-13. Suppose the discrimination coefficient increases as the firm employs more black workers. In particular, the discrimination coefficient is $d = 0.01E_B$ where E_B is the number of blacks hired by the firm so that each employer facing competitive wages of w_W for whites and w_B for blacks acts as if she faces competitive wages of w_W for whites and $w_B(1 + d)$ for blacks. Lastly, assume that the firm must employ 200 workers. Define the wage ratio to be w_W / w_B. Solve for the number of blacks hired as a function of the wage ratio. Graph the number of blacks hired (x-axis) against the wage ratio (y-axis).

9-14. Consider a data set with the following descriptive statistics.

TABLE 1 Descriptive Statistics

	Men			Women		
	Mean	Min	Max	Mean	Min	Max
Ln(wages)	3.562	1.389	5.013	3.198	1.213	4.875
Black	0.231	0	1	0.191	0	1
Age	42.2	19	68	39.2	19	63
Work experience	18.1	0	42	16.1	0	35
Schooling	13.9	9	21	14.1	9	21
% female occupation	18.2	2.3	95.4	62.3	6.7	98.5

Wage is the worker's hourly wage; Black takes on a value of 1 if the worker is Black and a value of 0 otherwise; work experience is actual years of work experience, schooling is measured in years; and % female occupation is the percent of all employees in the worker's occupation who are female. The following table reports the regression results from a log-wage regression.

TABLE 2 Regression Results

	Men	Women
Constant	2.314	2.556
Black	–0.198	–0.154
Age	0.054	0.037
Years of work experience	0.042	0.059
Years of schooling	0.085	0.083
Percent female in occupation	–0.0012	0.0024

Decompose the raw difference in average wages using the Oaxaca–Blinder decomposition. Specifically, decompose the raw difference into the portion due to differences in personal characteristics (schooling, race, age, and experience), the portion due to occupation, and the portion left unexplained possibly due to gender discrimination.

9-15. In 2006, Evo Morales assumed the presidency in Bolivia, a South American country in which official commerce is done in Spanish. Morales was the first Bolivian president of indigenous decent. As president, he quickly instituted reforms that were designed to reduce discrimination against indigenous populations with the aim of eventually reducing inequality. Suppose discrimination before Morales took two forms–discrimination in education by not providing state funds to educate all children (and particularly not educating indigenous children in Spanish), and discrimination in the job market by firms not willing to hire indigenous workers.

(a) In terms of education, which policy would be better at combating discrimination and inequality: (1) providing state funds to educate all people in their native languages, or (2) providing state funds for a public education system that requires all people to learn Spanish and a second, indigenous language? Why?

(b) In terms of the job market, which policy would be best at combating discrimination and inequality: (1) increasing the minimum wage, (2) requiring all firms with at least 50 workers to hire some indigenous workers, or (3) improving the legal system to protect economic rights and activities? Why?

Selected Readings

Marianne Bertrand, Claudia Goldin, and Lawrence F. Katz, "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors," *American Economic Journal: Applied Economics* 2 (July 2010): 228–255.

Marianne Bertrand and Sendhil Mullanaithan, "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review* 94 (September 2004): 991–1013.

Francine D. Blau, "Trends in the Well-Being of American Women, 1970–1995," *Journal of Economic Literature* 36 (March 1998): 112–165.

Cody Cook, Rebecca Diamond, Jonathan Hall, John A. List, and Paul Oyer, "The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers," National Bureau of Economic Research Working Paper No. 24732, June 2018.

Kerwin K. Charles and Jonathan Guryan, "Prejudice and Wages: An Empirical Assessment of Becker's The Economics of Discrimination," *Journal of Political Economy* 116 (October 2008): 773–809.

Matthew S. Goldberg, "Discrimination, Nepotism, and Long-Run Wage Differentials," *Quarterly Journal of Economics* 97 (May 1982): 307–319.

Jeffrey Grogger, "Speech Patterns and Racial Wage Inequality," *Journal of Human Resources* 46 (Winter 2011): 1–25.

Daniel S. Hamermesh and Jeff E. Biddle, "Beauty and the Labor Market," *American Economic Review* 84 (December 1994): 1174–1194.

James J. Heckman and Brook S. Payner, "Determining the Impact of Federal Antidiscrimination Policy on the Economic Status of Blacks: A Study of South Carolina," *American Economic Review* 79 (March 1989): 138–177.

Jacob Mincer and Solomon W. Polachek, "Family Investments in Human Capital: Earnings of Women," *Journal of Political Economy* 82 (March 1974 Supplement): S76–S108.

Chapter 10

Labor Unions

Union gives strength.

—*Aesop*

Up to this point, we have ignored the institution of labor unions. The omission may seem surprising. After all, supporters of the union movement often argue that labor unions, as the key institution representing workers' interests in the labor market, are mainly responsible for the improvement in working conditions in many industrialized countries. And even though union membership in the United States declined rapidly in recent decades, unions *still* represent 11 percent of workers.

This chapter argues that unions, like workers maximizing utility and firms maximizing profits, choose among various options to maximize the well-being of their members. As a result, the labor market impact of unions depends not only on the political and institutional environment that regulates the employer–union relationship, but also on the factors that motivate unions to pursue certain strategies (such as making wage demands that may lead to a strike) and to ignore others.

It has long been recognized that unions can arise and prosper only under certain conditions. Because the free entry and exit of firms into the marketplace reduce profits to a normal return on investment (that is, zero excess profits), unions can flourish only when firms earn above normal profits, or what economists call "rents." In a sense, unions provide an institutional mechanism through which employers share the rents with workers.

This chapter investigates how unions change the terms of the employment relationship between workers and firms. We will find that unions influence practically every aspect of the employment contract, including hours of work, wages, and productivity, and also have important effects on the firm's profits.[1]

[1] Detailed surveys of the evidence include Barry T. Hirsch and John T. Addison, *The Economic Analysis of Unions: New Approaches and Evidence,* Boston: Allen & Unwin, 1986; John H. Pencavel, *Labor Markets under Trade Unionism: Employment, Wages, and Hours,* Cambridge, MA: Basil Blackwell, 1991.

10-1 A Brief History of American Unions

Prior to the Great Depression, social attitudes and the political climate toward unions were quite unfavorable.[2] Various legal restrictions and employer practices kept union membership in check. For instance, in the 1908 *Loewe v. Lawlor* decision, the Supreme Court upheld a judgment against the Hatters' Union because the union had organized a consumer boycott against a nonunion producer. The Supreme Court concluded that the union's actions reduced the flow of goods in interstate commerce and was a "restraint of trade" prohibited by the Sherman Antitrust Act. In subsequent decisions, the Court used the antitrust analogy to outlaw strikes that affected interstate commerce.

Employers also made use of **yellow-dog contracts.** These contracts stipulated that as a condition of employment, the worker would not join a union. When unions attempted to organize workers who had signed these contracts, the unions were found guilty of inducing a breach of contract.

Beginning with the New Deal legislative program, the legal environment facing unions changed substantially. Four major pieces of federal legislation form the new ground rules:

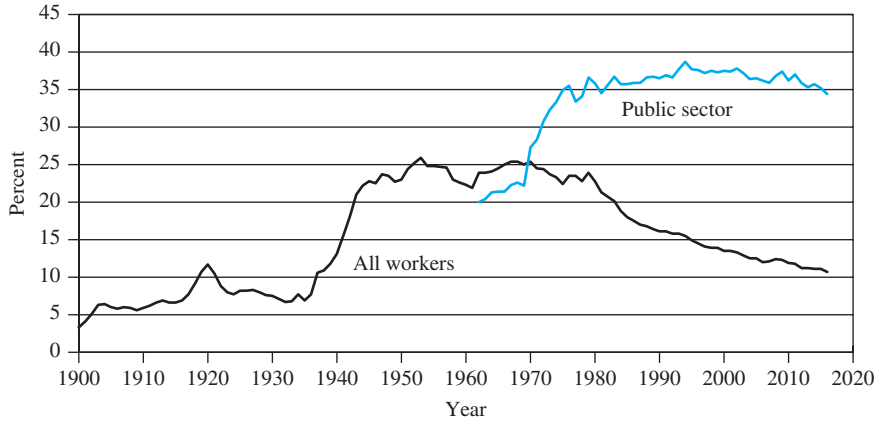
- *The Norris–LaGuardia Act of 1932.* This legislation tried to "even out" the game by restricting the employer's use of court orders and injunctions to hamper union organizing drives, and by making yellow-dog contracts unenforceable in federal courts.
- *The National Labor Relations Act of 1935* (also known as the *Wagner Act*). This legislation further increased the power of unions by defining a set of **unfair labor practices** by employers. Among the specific unfair labor practices were the firing of workers involved in union activities and discrimination against workers who support the union. The Wagner Act also established the National Labor Relations Board (NLRB) to enforce the provisions of the legislation. The NLRB can investigate unfair labor practices and can order that such practices be stopped. The NLRB also runs the elections where workers decide if a particular union is to represent them in collective bargaining. These elections are called **certification elections.**
- *The Labor–Management Relations Act of 1947* (also known as the *Taft–Hartley Act*). This legislation curbed union power by permitting states to pass **right-to-work laws.** These laws prohibit unions from requiring that workers become union members as a condition of employment in unionized firms. By 2018, 28 states had enacted right-to-work laws. The Taft–Hartley Act also permits workers to hold elections that would decertify a union from representing them in collective bargaining (or **decertification elections**).
- *The Labor–Management Reporting and Disclosure Act of 1959* (also known as the *Landrum–Griffin Act*). This legislation, passed in reaction to increasing evidence of corruption among union leaders, requires the complete disclosure of union finances.

Figure 10-1 illustrates the trend in union membership. Fewer than 10 percent of civilian workers in 1930 were union members. As a result of the major legislative changes during the New Deal, union membership began to rise rapidly. By the early 1950s, over a quarter of the civilian workforce was unionized. Unionization rates remained roughly at that level until the mid-1960s, when a steady decline in union membership began, with the decline

[2] A history of the union movement is given by Albert Rees, *The Economics of Trade Unions,* 2nd ed., Chicago: University of Chicago Press, 1977, Chapter 1.

FIGURE 10-1 Union Membership, 1900–2016 (Percent of Workers Unionized)

Sources: Barry T. Hirsch and John T. Addison, *The Economic Analysis of Unions: New Approaches and Evidence,* Boston, MA: Allen & Unwin, 1986, pp. 46–47; Barry T. Hirsch and David A. Macpherson, *Union Membership and Earnings Data Book: Compilations from the Current Population Survey (2017 Edition),* Washington, DC: Bureau of National Affairs, 2017.



accelerating in the 1980s. By 2017, only 10.7 percent of civilian workers were unionized. The phenomenon of the vanishing union is even more evident if we look at the fraction of unionized workers in the private sector: Only 6.5 percent of workers in the private sector are now unionized.

Prior to the 1960s, public-sector workers were specifically prohibited from forming unions. In 1962, President John Kennedy, through Executive Order No. 10988, gave federal workers the right to organize. The Civil Service Reform Act of 1978, which superseded President Kennedy's executive order, now regulates unions in the federal sector. This legislation prohibits strikes and protects the right of federal workers to either join or not join unions. A number of state laws also extended the right to organize to state and local workers. Not surprisingly, there has been a remarkable rise in public-sector unionization rates at the same time that union membership in the private sector was collapsing. Figure 10-1 also shows that although only about 20 percent of public-sector workers were unionized in the 1960s, this fraction had jumped to 34.4 percent by 2017.

It is useful to think of the union movement in the United States today as a pyramid. At the top of the pyramid is the AFL-CIO (which stands for American Federation of Labor and Congress of Industrial Organizations). The AFL-CIO is a federation of unions. The diverse set of unions affiliated with the AFL-CIO, which includes the American Federation of Teachers, the United Mine Workers, and the Actors' Equity Association, account for about 80 percent of all union members in the United States. Most of the unions affiliated with the AFL-CIO are national unions, representing workers throughout the country (and sometimes even representing workers outside the United States). In turn, these national unions are composed of "locals," or unions established at the city level or even the plant level. These locals are at the bottom of the pyramid. The main objective of the AFL-CIO is to provide a single, national voice for the diverse unions under its umbrella, to engage in political lobbying, and to support political candidates who are sympathetic to labor's social and economic agenda.

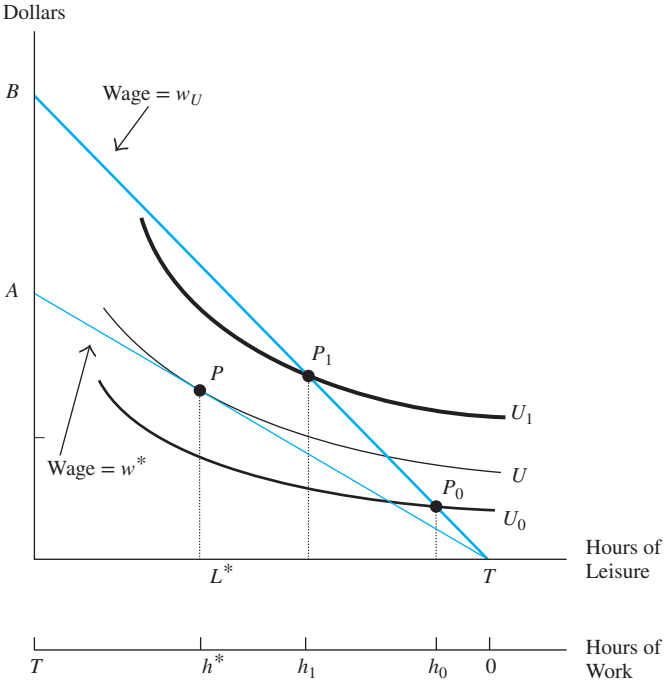
10-2 Determinants of Union Membership

Workers choose whether to join a union. He joins if the union offers him a wage–employment package that makes him better off than the package offered by a nonunion employer.[3]

Figure 10-2 uses the familiar model of the labor–leisure choice to illustrate the worker's decision. The worker is initially at a nonunion firm offering the competitive wage w^*. At this wage rate, the worker's budget line is given by AT. A worker maximizes utility by choosing the consumption–leisure bundle where the indifference curve U is tangent to the budget line (or point P). The nonunion worker consumes L^* hours of leisure and works h^* ($h^* = T - L^*$) hours.

FIGURE 10-2 The Decision to Join a Union

The budget line is given by AT. The worker maximizes utility at point P and works h^* hours. The union wage increase (from w^* to w_U) shifts the budget line to BT. If the employer cuts hours of work to h_0, the worker is worse off (utility falls from U to U_0). If the employer cuts hours to h_1, the worker is better off.



[3] Although the worker's utility obviously depends on many aspects of the job (including fringe benefits and working conditions), we focus on the simpler model where the characteristics of the job include only wages and hours of work. See Henry S. Farber and Daniel H. Saks, "Why Workers Want Unions: The Role of Relative Wages and Job Characteristics," *Journal of Political Economy* 88 (April 1980): 349–369; and Henry S. Farber, "The Determination of the Union Status of Workers," *Econometrica* 51 (September 1983): 1417–1437.

The firm is targeted by union organizers, and these organizers promise a new and improved employment contract. Specifically, the union promises a wage increase to w_U dollars. The worker's budget line, therefore, shifts to *BT*.

The worker knows that there is no free lunch. The wage increase comes at a cost, and the cost may be a cutback in employment. Suppose that the firm's demand curve for labor is downward sloping and elastic. If the firm responds to the union wage increase by moving up the labor demand curve, the union-mandated wage increase reduces the worker's workweek to, say, h_0 hours, placing him at point P_0 on the *BT* budget line. If the union organizes the firm's workforce, therefore, the worker would be worse off (he moves to the lower indifference curve U_0). This worker, therefore, opposes the union in the certification election.

If the firm's demand curve for labor is inelastic, the employment reduction is small and the union offers the wage–employment combination at point P_1 (where the workweek lasts h_1 hours). The union shifts the worker to a higher indifference curve (given by U_1) and the worker supports the union in the certification election.

Who Joins a Union?

Workers are more likely to support unionization when the union organizer can promise a high wage and a small decline in hours of work. Moreover, because there are additional costs to joining a union (such as union dues), the worker will be more likely to support unions when these costs are small. These factors generate the worker's "demand" for union jobs.

The ability of union organizers to deliver union jobs depends on the costs of organizing the workforce, on the legal environment that permits certain types of union activities and prohibits others, on the resistance of management to the introduction of collective bargaining, and on whether the firm is making excess rents that can be captured by the union membership. These forces, in effect, determine the potential "supply" of union jobs.

The extent of unionization observed in the labor market is determined by the interaction of these two forces. As a result, the unionization rate will be higher the more workers have to gain by becoming unionized and will be lower the harder it is to convert jobs from nonunion to union status. This "cost–benefit" approach helps us understand differences in unionization rates across demographic groups, across industries, and over time. Table 10-1 summarizes some of the key differences in the U.S. labor market.

Men have slightly higher unionization rates than women. In 2016, 11 percent of working men and 10 percent of women were unionized. The gender differential in unionization rates arises partly because women are more likely to be employed in part-time jobs or in jobs that offer flexible work schedules. Those types of jobs tend not to be unionized.

Blacks have higher unionization rates than whites. The unionization rate of black workers was 13.4 percent, as compared to about 10 percent for whites and 9 percent for Hispanics and Asians. It is not surprising that blacks are more likely to support unions because, as we will see below, unions compress wages within the firm, helping to reduce the impact of labor market discrimination on the black wage. The somewhat lower unionization rate of Hispanics might be due to the predominance of immigrant workers in the Hispanic population; many of these workers might be on the "fringes" of the labor market, and it is unlikely that those types of jobs are unionized.

TABLE 10-1 Union Membership by Selected Characteristics, 2016 (Percent of Workers Who Are Union Members)

Source: Barry T. Hirsch and David A. Macpherson, *Union Membership and Earnings Data Book: Compilations from the Current Population Survey (2017 Edition),* Washington, DC: Bureau of National Affairs, 2017, Tables 3a and 7a.

Gender:		Industry:	
Men	11.2	Private workforce	6.4
Women	10.2	Agriculture	1.8
		Mining	5.6
Race:		Construction	14.7
White	10.1	Manufacturing	8.8
Black	13.4	Transportation	25.2
Hispanic	8.8	Wholesale trade	3.7
Asian	9.0	Retail trade	4.4
		Finance	1.9
		Government	34.4

There are also sizable differences in unionization rates across private-sector industries, with workers in construction, manufacturing, and transportation being the ones most likely to be unionized, and workers in agriculture and finance being the least likely. The available evidence, in fact, suggests that workers employed in concentrated industries, where most of the output is produced by a few firms, are more likely to be unionized.[4] This result is consistent with our cost–benefit approach to understanding differences in unionization rates. Firms in concentrated industries earn excess profits because of their monopoly power, so unions have a better chance of extracting some of the rents for the workers.

Finally, the legal environment regulating the employer–union relationship has a large impact on the success of union organizing drives. States with right-to-work laws have much lower unionization rates than other states. In 2016, for instance, the five states with the lowest unionization rates (South Carolina, North Carolina, Georgia, Arkansas, and Texas) were also states with right-to-work laws. In these states, the unionization rate ranged from 1.6 to 4.0 percent.[5]

Are American Unions Obsolete?

The most noticeable feature of the American union movement today is the steady decline in unionization rates since 1970.[6] There were major changes in the structure

[4] Barry T. Hirsch and Mark C. Berger, "Union Membership Determination and Industry Characteristics," *Southern Economic Journal* 50 (January 1984): 665–679.

[5] Barry T. Hirsch and David A. Macpherson, *Union Membership and Earnings Data Book: Compilations from the Current Population Survey (2017 Edition),* Washington, DC: Bureau of National Affairs, 2017, Tables 4a. U.S. Bureau of the Census, *Statistical Abstract of the United States, 2008,* Washington, DC: Government Printing Office, 2008, Table 644.

[6] Henry S. Farber, "The Decline of Unionization in the United States: What Can Be Learned from Recent Experience," *Journal of Labor Economics* 8 (January 1990): 75–105; Henry S. Farber and Bruce Western, "Accounting for the Decline of Unions in the Private Sector, 1973–1998," *Journal of Labor Research* 22 (Summer 2001): 459–486; and William T. Dickens and Jonathan S. Leonard, "Accounting for the Decline in Union Membership: 1950–1980," *Industrial and Labor Relations Review* 38 (April 1985): 323–334.

of the economy during this period. In 1960, 31 percent of workers were employed in manufacturing, where union organizing drives have typically been successful. By 2001, the proportion of workers in manufacturing had fallen to 14 percent. The location of jobs also shifted. In the 1950s, only 42 percent of the jobs were located in southern and western states (which tend to have less favorable environments for union organizing, such as right-to-work laws). By 2001, 57 percent of the jobs were located in these states. There is, in fact, evidence suggesting that manufacturing activity is substantially higher in right-to-work states.[7]

In addition to the structural shifts in the economy, workers' demand for unionization may have declined. There were notable changes in how workers voted in union certification elections. The NLRB holds an election to certify a union as a collective bargaining agent after 30 percent of the workers petition for such an election. The union can represent the workers if they get a simple majority of the workers who will make up the bargaining unit. In 1955, unions won over two-thirds of certification elections. By 1990, unions won fewer than half of the elections.[8]

The worsening performance of unions in certification elections during this period has been attributed to aggressive antiunion tactics by management.[9] Management can reduce the success of union organizing drives in many ways, including filing petitions to delay the certification election, firing workers for union activities, and hiring consultants to handle the management campaign. The aggressive management response may be the result of the rise in foreign competition and the deregulation of many unionized industries (such as trucking, airlines, and railroads).[10] The tide of foreign goods into the United States captured part of the excess rents that were previously shared between firms and workers in the affected industries. Similarly, the deregulation of unionized industries introduced competitive forces into the marketplace, again dissipating excess rents. Not surprisingly, firms became much more resistant to union wage demands and to the introduction of union work rules.

[7] Thomas J. Holmes, "The Effect of State Policies on the Location of Manufacturing: Evidence from State Borders," *Journal of Political Economy* 106 (August 1998): 667–705.

[8] The *Annual Reports of the National Labor Relations Board* contain detailed statistics on the election results. The union "win rate" has increased since 2000, but there are now dramatically fewer certification elections, suggesting that unions are choosing their targets more carefully. See Henry S. Farber, "Union Organizing Decisions in a Deteriorating Environment: The Composition of Representation Elections and the Decline in Turnout," *Industrial and Labor Relations Review* 68 (October 2015): 1126–1156.

[9] Richard B. Freeman and Morris Kleiner, "Employer Behavior in the Face of Union Organizing Drives," *Industrial and Labor Relations Review* 43 (April 1990): 351–365; William T. Dickens, "The Effect of Company Campaigns on Certification Elections: *Law and Reality* Once Again," 36 (July 1983): 560–575.

[10] John M. Abowd and Thomas Lemieux, "The Effects of International Competition on Collective Bargaining Outcomes: A Comparison of the United States and Canada," in John M. Abowd and Richard B. Freeman, editors, *Immigration, Trade, and the Labor Market,* Chicago: University of Chicago Press, 1991.

Theory at Work

THE RISE AND FALL OF PATCO

The Professional Air Traffic Controllers Organization (PATCO) represented air controllers in collective bargaining negotiations with their employer, the Federal Aviation Administration (FAA). The union's brief (and militant) 13-year history ended when they called a strike in 1981. Because controllers are federal civil servants, their salaries were set by Congress and their right to strike was specifically prohibited by law. Nevertheless, much of PATCO's history was marked by the union's demands that they should be able to bargain directly over wages and that they had a right to strike.

PATCO began as an organization of New York City controllers in January 1968. By July 1968, under the leadership of attorney F. Lee Bailey—a future member of the "dream team" that defended O. J. Simpson at his murder trial—PATCO had already sponsored a work slowdown that seriously disrupted commercial air travel.

In 1980, air controllers earned high wages and had extraordinarily generous retirement and disability programs. They were among the highest-paid government employees and could retire at age 50 after 20 years of service. In contrast, most other federal workers needed 30 years of service if they wished to retire at age 55.

Despite the high salary and generous benefits, the PATCO leadership decided that 1981 would be a crucial year for the union and prepared to aggressively demand even higher earnings and better benefits. Most important, the leadership decided that the way to persuade Congress to agree was through a strike.

PATCO made unreasonable demands in the initial rounds of the negotiation: An immediate $20,000 salary increase, a 32-hour workweek, and more generous pension and disability benefits. The Reagan administration countered with an immediate pay raise of $4,000, overtime pay after 36 hours per week (rather than 40), and various other benefits. If PATCO had accepted the administration's offer (and Congress had consented), controllers would have gotten pay increases exceeding 11 percent, more than twice what other federal employees got.

But PATCO wanted much more, and the rest is history. PATCO's strike began at 7 a.m. on August 3, 1981. The FAA was prepared and moved quickly to staff the control towers at airports with military personnel, retirees, supervisors, and controllers who refused to strike.

Four hours after the strike began, President Reagan personally announced that the law would be enforced and that any striker not on the job within 48 hours would be fired and would not be reemployed by any other federal agency. About one-fourth of the 16,395 controllers did not go on strike and another 875 returned to work before the deadline. The 48 hours passed and 11,301 controllers were fired. It soon became obvious that the system was overstaffed. It eventually reached full capacity again with about 20 percent fewer controllers.

The militancy of the PATCO leadership, combined with President Reagan's resolve to enforce the law, created a political and cultural environment that likely influences labor relations to this day.

Source: Herbert R. Northrup, "The Rise and Demise of PATCO," *Industrial and Labor Relations Review* 37 (January 1984): 167–184.

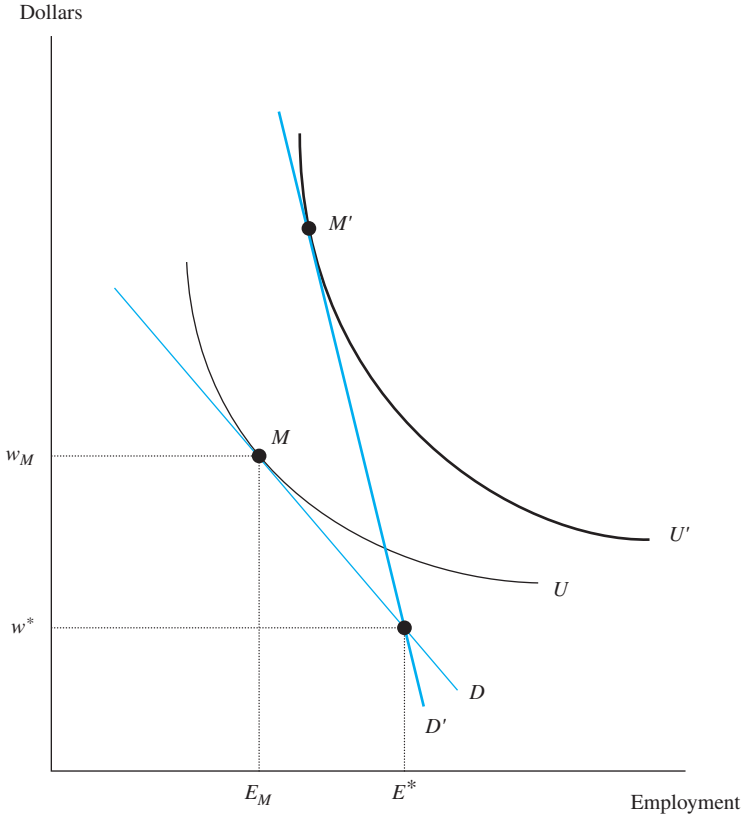
10-3 Monopoly Unions

Samuel Gompers, founder of the American Federation of Labor, was once asked what unions wanted. His reply was simple and memorable: "More." Economists keep this response in mind when building models of union behavior.[11] We typically assume that the union's utility depends on the wage w and employment E, and that unions want more of both. The union's utility function is then given by $U(w, L)$ and the union's indifference curves have the usual shape, as shown in Figure 10-3 (see curves U and U').

[11] Henry S. Farber, "The Analysis of Union Behavior," in Orley Ashenfelter and Richard Layard, editors, *Handbook of Labor Economics,* vol. 2, Amsterdam: Elsevier, 1986, pp. 1039–1089.

FIGURE 10-3 Monopoly Unions

A monopoly union maximizes utility by choosing the point on the labor demand curve D that is tangent to the union's indifference curve. The union demands wage w_M and the employer cuts employment to E_M (from the competitive level E^*). If the demand curve were inelastic (as in D'), the union could demand a higher wage and get more utility.



The union wants to maximize utility.[12] Suppose that the union is dealing with a profit-maximizing competitive firm. This firm has a downward-sloping labor demand curve that specifies how many workers it is willing to hire at any wage. In a sense, the firm's labor demand curve D in Figure 10-3 is a constraint on union behavior. Unions would then maximize utility by choosing a point like M, where the labor demand curve D is tangent to the union's indifference curve U.

[12] There is a conceptual problem with the utility-maximizing approach to modeling the behavior of unions: Where exactly does the union's utility function come from? The union is not a person but is composed of many workers. If all workers had the same preferences over wages and employment, and if the leadership were elected democratically, the union's preferences would be identical to that of the typical worker. But it is doubtful all workers have the same preferences; see Henry S. Farber, "Individual Preferences and Union Wage Determination," *Journal of Political Economy* 86 (October 1978): 923–942.

Suppose that the competitive wage is w^*. In the union's absence, the firm would hire E^* workers. The union, however, demands a wage of w_M and the firm responds by cutting employment to E_M.

This solution has some interesting properties. Most important, note that the union chooses the wage and the firm then moves along the demand curve to set the profit-maximizing level of employment. The model of union behavior summarized in Figure 10-3 is called a model of **monopoly unions.** The union has an effective monopoly on the sale of labor to the firm. The union sets the price of its product (that is, it sets the wage) and firms then look at the labor demand curve to determine how many workers to hire.

The model of monopoly unions implies that some workers lose their jobs as a result of the union's wage demand. It is not surprising, therefore, that unions get more utility when the labor demand curve is inelastic. If the demand curve were given by D' in Figure 10-3 (which is more inelastic than D), the union would want an even higher wage (at point M') and jump to a higher indifference curve because employment does not fall very much.

As we noted in our discussion of Marshall's rules of derived demand, unions will want to manipulate the labor demand elasticity by making it difficult for firms to substitute between union and nonunion labor and for consumers to substitute between goods produced by union and nonunion firms. Because workers choose to join unions, union organizing drives will be more successful in firms that have inelastic labor demand curves. The evidence indeed suggests that the elasticity of labor demand is smaller in union firms than in nonunion firms.[13]

10-4 Policy Application: The Efficiency Cost of Unions

The wage–employment solution implied by the model of monopoly unions is inefficient because unions reduce the total value of labor's contribution to national income. If employers move along the demand curve as a result of union-mandated wage increases, unions reduce employment in union firms and increase employment in nonunion firms (as long as the displaced workers move to nonunion jobs). Because the wage (and the value of marginal product of labor) differs between the two sectors, unionism introduces an inefficiency into the economy. The last worker hired by nonunion firms would have a greater productivity if he or she had been hired in the union sector, and the value of labor's contribution to national income would increase if some workers were reallocated across sectors.[14]

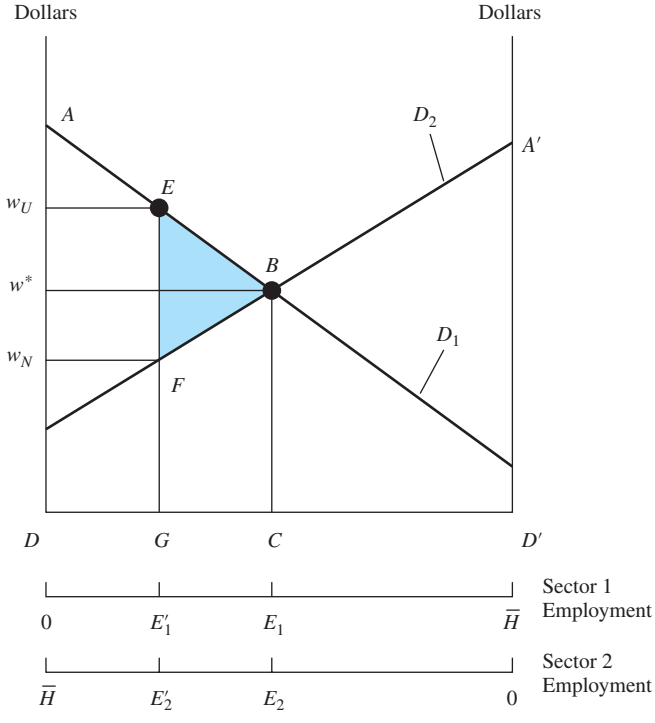
Figure 10-4 illustrates the efficiency loss. There are two sectors in the economy: sector 1 and sector 2. Sector 1's labor demand curve is given by D_1 and sector 2's demand curve is given by D_2. For expositional convenience, both demand curves are superimposed in the same graph. The demand curve for sector 1 is drawn in the typical fashion, whereas the demand curve for sector 2 goes from right to left. Finally, we assume labor supply is inelastic, so that a total of $\overline{H}$ workers must be employed in one of the two sectors.

[13] Richard B. Freeman and James L. Medoff, "Substitution between Production Labor and Other Inputs in Unionized and Nonunionized Manufacturing," *Review of Economics and Statistics* 64 (May 1982): 220–233.

[14] Albert Rees, "The Effects of Unions on Resource Allocation," *Journal of Law and Economics* 6 (October 1963): 69–78; Robert DeFina, "Unions, Relative Wages, and Economic Efficiency," *Journal of Labor Economics* 1 (October 1983): 408–429.

FIGURE 10-4 Unions and Labor Market Efficiency

In the absence of unions, the competitive wage is w^* and national income is given by the sum of the areas *ABCD* and *A′BCD′*. Unions increase the wage in sector 1 to w_U. The displaced workers move to sector 2, lowering the nonunion wage to w_N. National income is now given by the sum of areas *AEGD* and *A′FGD′*. The misallocation of labor reduces national income by the area of the triangle *EBF*.



In the absence of unions, the competitive wage equals w^*. At this wage, all workers are employed. Sector 1 employs E_1 workers and sector 2 employs E_2 workers (or $\overline{H} - E_1$). Because the labor demand curve gives the value of marginal product of labor, the area under the demand curve measures the value of total product. In the initial equilibrium, therefore, the value of output in sector 1 equals the area of the trapezoid *ABCD* and the value of output in sector 2 equals the area of the trapezoid *A′BCD′*. The sum of these two areas equals national income.

Suppose that a monopoly union represents workers in sector 1, raising that sector's wage to w_U. Because employment in the union sector falls to E_1', employment in the nonunion sector must increase to E_2' and the nonunion wage falls to w_N.

The value of output in the union sector is now given by the area of the trapezoid *AEGD* and the value of output in the nonunion sector increases to the area of the trapezoid *A′FGD′*. Note that the sum of these two areas is smaller than national income in the absence of a union, the shortfall being the area of the shaded triangle *EBF*. This triangle is the deadweight loss that arises because the union sector is hiring too few workers and the nonunion sector is hiring too many.

The analysis in Figure 10-4 suggests a simple way for calculating the efficiency loss resulting from unionization. The area of the shaded triangle *EBF* is

$$\text{Efficiency loss} = \frac{1}{2} \times (w_U - w_N) \times (E_1 - E_1') \qquad \textbf{(10-1)}$$

After rearranging terms in this equation, the efficiency loss as a fraction of national income can be written as[15]

$$\begin{aligned}\frac{\text{Efficiency loss}}{GDP} = \frac{1}{2} &\times (\text{Percent union–nonunion wage gap}) \\ &\times (\text{Percent decline in employment in union sector}) \\ &\times (\text{Fraction of labor force, that is, unionized}) \\ &\times (\text{Labor's share of national income})\end{aligned} \qquad \textbf{(10-2)}$$

Suppose that unions increase wages by 15 percent. Further, let's assume that the demand curve for union workers is unit elastic so that employment in the union sector also falls by 15 percent. Finally, 11 percent of workers were unionized in 2017, and labor's share of national income is about 0.7. Plugging these values into equation (10-2) implies that the efficiency loss as a fraction of national income is slightly less than 0.1 percent (or $\frac{1}{2} \times 0.15 \times 0.15 \times 0.11 \times 0.7$). National income was about $19 trillion in 2018. The efficiency loss attributable to unions then equals $19 billion, a relatively small amount in the context of a $19 trillion economy.

10-5 Efficient Bargaining

As we have seen, the wage–employment solution implied by monopoly unionism is inefficient because unions reduce national income. This fact suggests that perhaps the firm and the union could find—and agree on—an employment contract that does not lie on the labor demand curve and that would make at least one of the parties better off, without making the other party worse off.

Isoprofit Curves

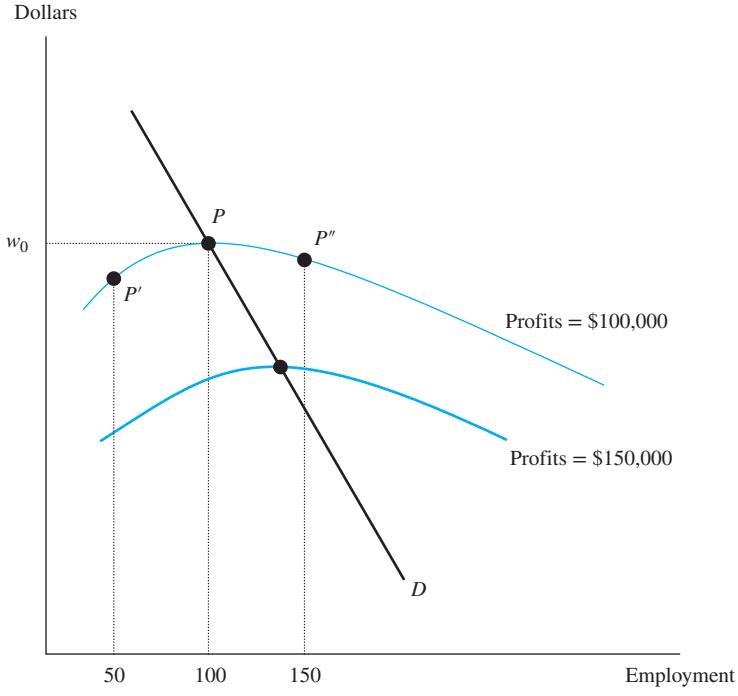
Before showing how both the union and the firm can benefit by moving off the labor demand curve, we first derive the firm's isoprofit curves. An isoprofit curve gives the wage–employment combinations that yield the same level of profits. A profit-maximizing firm would be indifferent among the various wage–employment combinations that lie on a single isoprofit curve.

Suppose that the wage is set at w_0 dollars. A profit-maximizing firm would then choose point *P* on the labor demand curve in Figure 10-5, hiring 100 workers. This wage–employment combination yields a particular level of profits, say, $100,000. It turns out that there are other wage–employment combinations that yield the same profits. Suppose, for instance, that the firm did not hire 100 workers, but hired fewer workers instead, say, 50.

[15] Equation (10-1) can be rewritten as

$$\frac{\text{Efficiency loss}}{GDP} = \frac{1}{2} \times \frac{w_U - w_N}{w_N} \times \frac{E_1 - E_1'}{E_1} \times \frac{E_1}{\overline{H}} \times \frac{w_N \overline{H}}{GDP}$$

FIGURE 10-5 The Labor Demand Curve and the Isoprofit Curves
If the wage is w_0, the firm maximizes profits (and earns $100,000) by hiring 100 workers. If the firm wants to hire 50 workers and hold profits constant, it must reduce the wage. Similarly, if the firm wants to hire 150 workers and hold profits constant, it must also reduce the wage. The isoprofit curve has an inverted-U shape. Lower isoprofit curves yield more profits.



If the wage remained constant at w_0, the firm would earn more profits by hiring 100 workers than by hiring 50 workers. After all, hiring 100 workers is the *right* (that is, profit-maximizing) thing to do at wage w_0. The firm can hire 50 workers and maintain profits constant only if it pays them a lower wage, as illustrated by point P' in the figure.

Suppose, instead, that the firm hired "too many" workers, say, 150. Again, at wage w_0 the firm earns higher profits by hiring 100 workers than by hiring 150 workers. The only way profits could remain constant if the firm hired 150 workers would be to pay a lower wage, as at point P'' in the figure. The firm's isoprofit curve, therefore, has an inverted-U shape, peaking where it intersects the demand curve for labor.

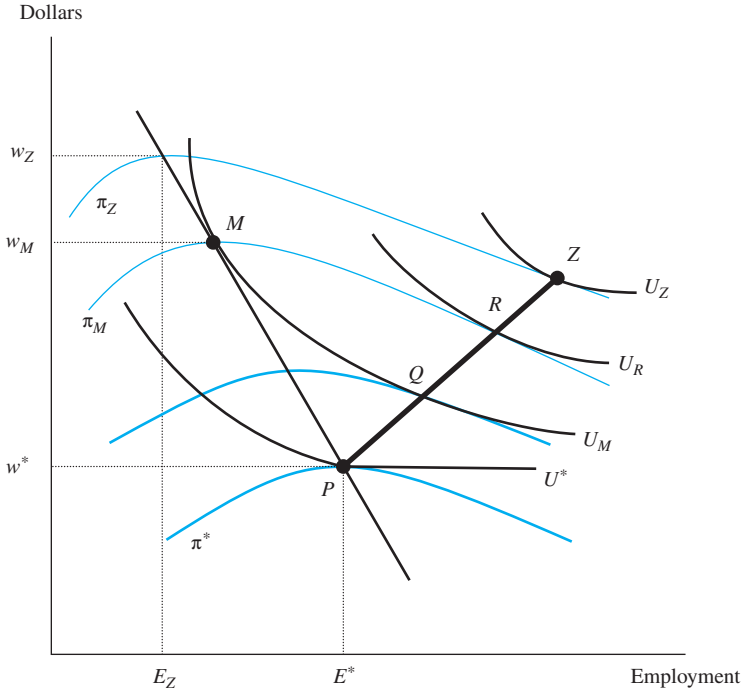
We can derive an entire family of isoprofit curves, one curve for each level of profits. Note that lower isoprofit curves are associated with *higher* profits. In Figure 10-5, for example, a firm hiring 100 workers would be better off if it could locate itself on a lower isoprofit curve (such as the one yielding $150,000); the firm would then be paying those workers a lower wage.

The Contract Curve

Figure 10-6 shows why both firms and unions have an incentive to move off the demand curve. The competitive wage is w^*. At that wage, the firm would employ E^* workers

FIGURE 10-6 Efficient Contracts and the Contract Curve

The employer hires E^* workers at the competitive wage w^*. A monopoly union moves the firm to point M, demanding wage w_M. Both the union and the firm are better off by moving off the demand curve. At point R, the union is better off and the firm is no worse off. At point Q, the employer is better off, but the union is no worse off. If all bargaining opportunities between the two parties are exhausted, the union and the firm agree to a wage–employment combination on the contract curve PZ.



(at point P) and earns π^* dollars in profits. If the union workers were to accept the wage–employment offer at point P, the union would get U^* units of utility.[16]

If the union were a monopoly union, it would pick point M on the demand curve (and get U_M utils). But the firm could try to talk the union into moving to point Q. The union would be indifferent between points M and Q (both points lie on the same indifference curve), but the firm would be better off because Q is on a lower isoprofit curve. By moving *off* the demand curve to point Q, therefore, the union would not be worse off than at the monopoly solution M, but the firm would be better off.

Similarly, the union could try to talk the firm into moving to point R. At this point, the firm would earn the same profits as at point M, but the union would be better off because it could jump to the higher indifference curve U_R. If the union and the firm could agree to move off the demand curve to any point between point Q and point R, then *both* the union and the firm would be better off than at the monopoly union solution M.

[16] The indifference curve U^* is drawn so that the union would not accept a wage below the competitive wage (note that the indifference curve becomes horizontal at w^*). This ensures that contract curve we are about to derive starts at the competitive solution (point P).

Suppose that the highest wage the firm can pay without incurring a loss is w_Z. At that wage, the firm would hire E_Z workers. The isoprofit curve going through this particular wage–employment combination is given by π_Z and gives all the wage–employment combinations that produce zero profits. This isoprofit curve provides the upper bound to the wage–employment combinations that the firm is willing to offer. If the firm chooses any point above the zero-profit isoprofit curve, it would incur a loss and go out of business.

Therefore, there are many off-the-demand-curve wage–employment combinations that could be beneficial to both the union and the firm. The curve *PZ* gives all the points where the union's indifference curves are tangent to the firm's isoprofit curves. These wage–employment combinations are **Pareto optimal**, because once a deal is struck anywhere on this curve, deviations from that particular deal can improve the welfare of one of the parties only at the expense of the other. The curve *PZ* is called the **contract curve**. If the union and the firm agree to a wage–employment combination on the contract curve, the resulting contract is called an **efficient contract**.[17]

The two extreme points on the contract curve bound the range of possible outcomes of the collective bargaining process. At point *P,* the union workers get paid the competitive wage and the firm keeps all the rents. At point *Z,* all the rents are transferred to the workers and the firm makes zero profits. The contract curve, therefore, provides the basis for negotiations between the union and the firm.

Note that the contract curve lies to the right of the labor demand curve. For any given wage, an efficient contract leads to more employment than would be observed with monopoly unions. Put differently, an efficient contract suggests that the employer–union relationship is not characterized by the union demanding a higher wage and by the firm responding by moving up the demand curve. Rather, efficient contracts imply that unions and firms bargain over both wages *and* employment.

An efficient contract also suggests that the unionized firm is overstaffed, hiring far more workers than it otherwise would at the "going" wage. Examples of overstaffing abound in unionized markets. For instance, even though airlines need only two pilots to fly a particular type of aircraft, the union contract may require they hire three. The firm and the union will then have to negotiate "make-work" or **featherbedding practices** to share the available tasks among the many workers.[18]

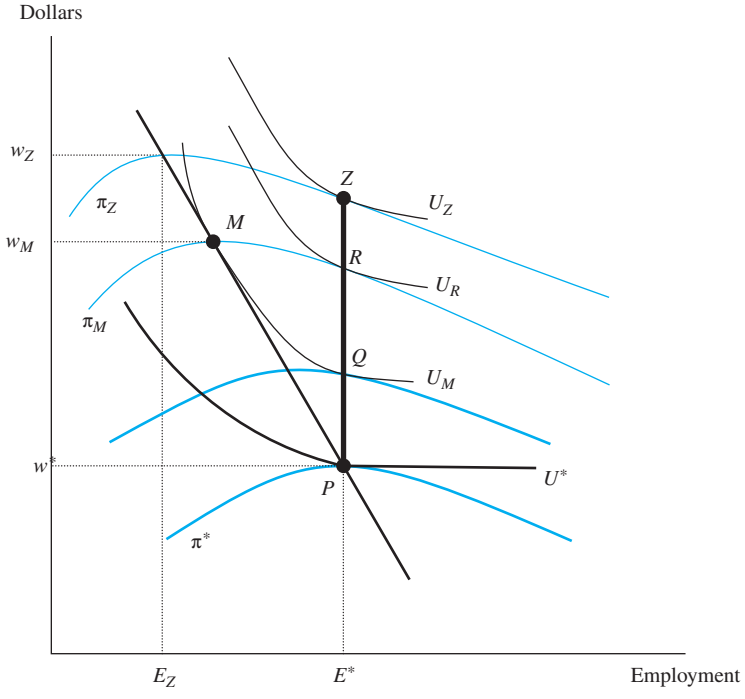
Strongly Efficient Contracts

Figure 10-7 illustrates an interesting possible shape for the contract curve, the vertical line *PZ.* The firm, therefore, hires the same number of workers, E^*, regardless of whether it is unionized or not. If the contract curve is vertical, the deal struck between the union and the worker is called a **strongly efficient contract** because the unionized firm is hiring the competitive level of employment.

[17] The model has its origins in Wassily Leontief, "The Pure Theory of the Guaranteed Annual Wage Contract," *Journal of Political Economy* 54 (February 1946): 76–79; Ian McDonald and Robert Solow, "Wage Bargaining and Employment," *American Economic Review* 71 (December 1981): 896–908.

[18] George E. Johnson, "Work Rules, Featherbedding, and Pareto-Optimal Union-Management Bargaining," *Journal of Labor Economics* 8 (January 1990, Part 2): S237–S259; Steven G. Allen, "Union Work Rules and Efficiency in the Building Trades," *Journal of Labor Economics* 4 (April 1986): 212–242.

FIGURE 10-7 Strongly Efficient Contracts
If the contract curve *PZ* is vertical, the firm hires the same number of workers that it would have hired in the absence of a union. The union and the firm are then splitting a fixed-size pie as they move up and down the contract curve. At point *P,* the employer keeps all the rents; at point *Z,* the union gets all the rents.



Because employment is the same regardless of which deal is struck on the vertical contract curve, the firm's output and revenue are also constant. As a result, a vertical contract curve essentially describes the many ways in which a fixed-size pie can be shared between the union and the worker. The firm's profits clearly depend on which particular point is chosen along *PZ.* At point *P,* the firm keeps all the excess profits. As the firm and the union move up the contract curve, the union gets more and more of the rents. The choice of a point along the vertical contract curve, therefore, is equivalent to a particular way of slicing the *same* pie.

It is unfortunate that the term "efficient contracts" is now commonly applied to all contracts that lie on the contract curve regardless of whether the contract curve is vertical or not. *Wage–employment combinations on an upward-sloping contract curve are efficient only in the sense that they exhaust all bargaining opportunities between the employer and the union.* In other words, any other wage–employment combinations can improve the welfare of one of the parties only at the expense of the other. These wage–employment combinations, however, are *not* efficient in an allocative sense because unionized firms are not hiring the number of workers they would have hired in a perfectly competitive market.

Wage–employment combinations on a vertical contract curve, however, are efficient in two distinct ways. First, they exhaust all bargaining opportunities between the employer and the union. Second, firms hire the "right" number of workers so that unions do not distort the allocation of labor, and there is no deadweight loss to the national economy.

Evidence on Efficient Contracts

The contract curve defines the range over which unions and firms can bargain over wages and employment. The process of collective bargaining narrows down the possibilities to a single point on the contract curve. The point that is chosen depends on the bargaining power of the two parties involved. Regardless of how the bargaining process ends, our analysis of efficient contracts suggests that both firms and unions will want to move off the demand curve.

There has been a lot of empirical research to determine if unions and firms indeed reach an efficient contract. Many of the studies in this literature estimate regressions that relate the employment in union firms to the union wage and to the competitive wage in the industry. If unions behaved like monopoly unions, the level of employment in union firms would depend only on the union wage and would not depend on the competitive wage in the industry. In contrast, if union contracts were strongly efficient, the level of employment in the union firm should be unrelated to the union wage but would depend instead on the competitive wage.

The available studies seem to indicate that wage–employment outcomes in unionized firms do not lie on the labor demand curve.[19] Detailed analysis of collective bargaining agreements with the International Typographical Union (ITU), where the data on union wages and employment date back to 1946, suggests that union employment depends on the competitive wage in the labor market, as implied by the efficient contracts model. There is, however, disagreement over whether the contract curve is vertical. Some studies find that union employment is also sensitive to the union wage, contradicting the hypothesis that the firm hires the competitive level of employment regardless of the union wage.

The strongest evidence in favor of a vertical contract curve is given by a study of the relationship between the timing of union contracts and the value of the firm in the stock market.[20] This analysis indicates that a $1 unexpected increase in the share of rents going to union workers reduces the value of the firm (that is, the shareholders' wealth) by exactly $1. This result is precisely what we would expect to find if the contract curve were vertical because a fixed-size pie is being shared, and there would be a dollar-for-dollar tradeoff in rents between workers and firms.[21]

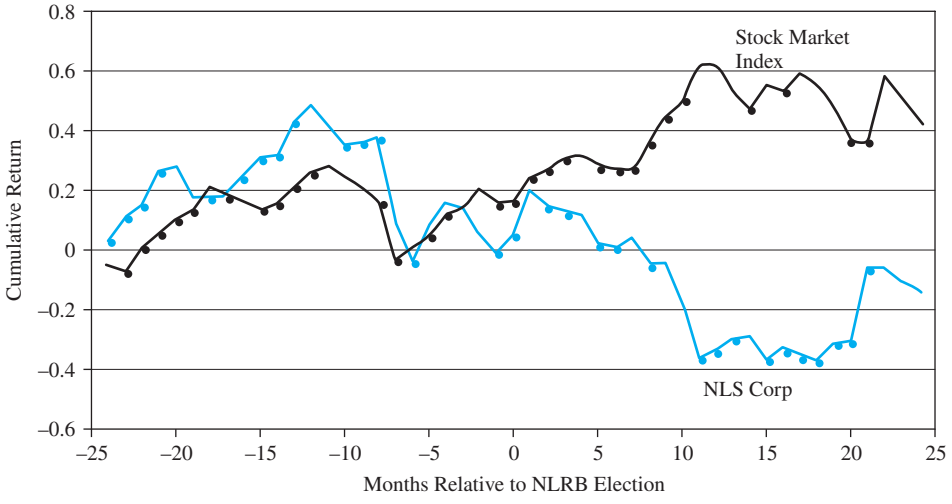
[19] Thomas E. MaCurdy and John H. Pencavel, "Testing between Competing Models of Wage and Employment Determination in Unionized Markets," *Journal of Political Economy* 94 (June 1986): S3–S39; James N. Brown and Orley Ashenfelter, "Testing the Efficiency of Employment Contracts," *Journal of Political Economy* 94 (June 1986): S40–S87; and Randall W. Eberts and Joe A. Stone, "On the Contract Curve: A Test of Alternative Models of Collective Bargaining," *Journal of Labor Economics* 4 (January 1986): 66–81.

[20] John M. Abowd, "The Effect of Wage Bargains on the Stock Market Value of the Firm," *American Economic Review* 79 (September 1989): 774–800.

[21] Several studies have tried to estimate the sharing ratio, the fraction of rents going to union workers. The estimates range from 0.1 to 0.7. See Jan Svejnar, "Bargaining Power, Fear of Disagreement and Wage Settlements: Theory and Evidence from U.S. Industry," *Econometrica* 54 (September 1986): 1055–1078; John M. Abowd and Thomas Lemieux, "The Effects of Product Market Competition on Collective Bargaining Agreements: The Case of Foreign Competition in Canada," *Quarterly Journal of Economics* 108 (November 1993): 983–1014; and Louis N. Christofides and Andrew J. Oswald, "Real Wage Determination and Rent-Sharing in Collective Bargaining Agreements," *Quarterly Journal of Economics* 107 (August 1992): 985–1002.

FIGURE 10-8 Stock Market Returns Before and After a Certification Election

Source: David S. Lee and Alexandre Mas, "Long-Run Impacts of Unions on Firms: New Evidence from Financial Markets, 1961–1999," *Quarterly Journal of Economics* 127 (February 2012): 333–378.



The strong link between unionization and shareholder wealth is easy to visualize by simply tracking what happens to the value of a particular firm as a result of a successful union certification election. In March 1999, workers at the National Linen Service Corp. (NLS), a large linen supplier, voted overwhelmingly to organize themselves into a local chapter of the Union of Needletrades, Industrial, and Textile Employees. Figure 10-8 shows the cumulative return to NLS stock during the 4-year period before and after the election. Prior to the election, the trend in the return to NLS stock was roughly similar to the trend in the overall stock market. Soon after the election, the returns to NLS stock began to fall behind. After 2 years, the price of NLS shares had fallen by about 25 percent, while the broad market index had risen by 50 percent.

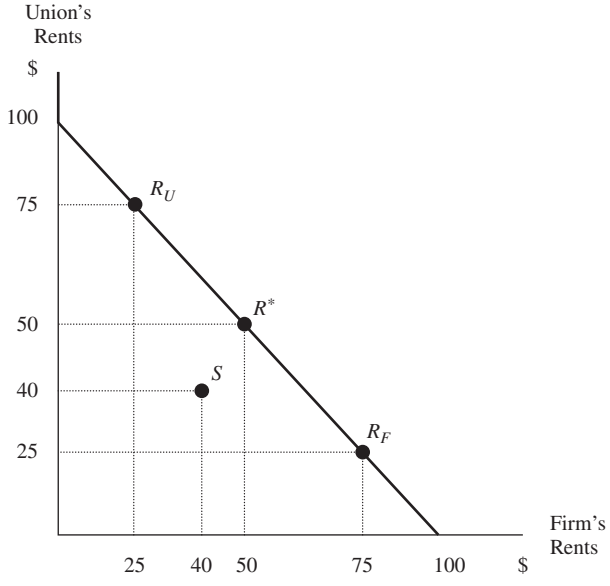
10-6 Strikes

Economists have had a difficult time explaining why strikes occur.[22] The problem can be easily described. Suppose that there are $100 worth of rents to be shared between the union and the firm. The downward-sloping line shown in Figure 10-9 illustrates the many ways in which those rents can be shared. The firm offers the division of rents indicated by point R_F, where the firm keeps $75 and the union gets $25. The union makes the counteroffer at R_U, where the union keeps $75 and the firm gets $25. Neither party gives in, and a strike occurs.

[22] John Kennan, "The Economics of Strikes," in Orley C. Ashenfelter and Richard Layard, editors, *Handbook of Labor Economics,* vol. 2, Amsterdam: Elsevier, 1986, pp. 1091–1137.

FIGURE 10-9 The Hicks Paradox

The firm makes the offer at point R_F, keeping \$75 and giving the union \$25. The union wants point R_U, getting \$75 for its members and giving the firm \$25. The parties do not come to an agreement and a strike occurs. The strike is costly, and the settlement is at point *S;* each party keeps \$40. Both parties could have agreed to a prestrike settlement at point *R**, and both parties would have been better off.



However, strikes are costly. The firm's profits fall; it may lose customers permanently; and a highly publicized strike may diminish the value of the brand. As a result, the size of the available pie shrinks and the two parties finally come to terms at point *S,* where each party gets \$40. The firm kept a bigger share of the pie than the union wanted to give (that is, \$40 versus \$25), so the firm can claim that it "won." And the union gets a bigger share of the pie than the firm was willing to give (again, \$40 versus \$25), and the union too can claim it "won."

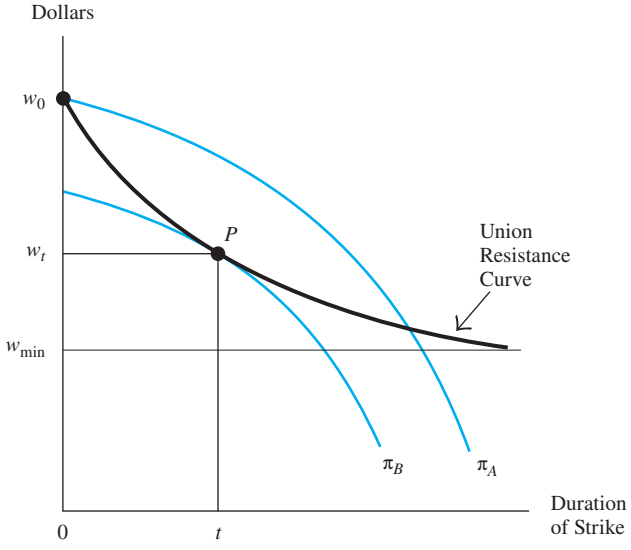
But it's a hollow victory for both sides. If everyone could have foreseen the end result, they could have initially agreed to other sharing solutions (such as point *R**, where each side keeps \$50) which would have made both parties better off relative to the poststrike outcome. In other words, strikes are not Pareto optimal. When the parties have reasonably good information about the cost and the likely outcome, it is irrational to strike. The irrationality of strikes has come to be known as the **Hicks paradox.**[23]

Strikes and Asymmetric Information

A number of ingenious models have been proposed to escape the Hicks paradox. The models typically emphasize that strikes occur because workers are not well informed about the firm's financial status and may have unreasonably optimistic expectations about the size of

[23] The irrationality of strikes was first noted by John R. Hicks, *The Theory of Wages,* London: Macmillan, 1932.

FIGURE 10-10 The Optimal Duration of a Strike
Unions moderate their wage demands the longer the strike lasts, generating a downward-sloping union resistance curve. The employer chooses the point on the union resistance curve that puts him on the lowest isoprofit curve. This occurs at point *P;* the strike lasts t periods and the settlement wage is w_t.



the pie and how much of it the firm is willing to give away. In effect, there is asymmetric information at the bargaining table. The firm knows more about the size of the pie than does the union or the workers.[24]

Because workers do not know the firm's true financial condition, the strike helps "teach a lesson" to the workers. Figure 10-10 illustrates the **union resistance curve** summarizing the lesson that is learned. Based on their incomplete information about the size of the pie prior to the strike, the union makes a perhaps unrealistic initial wage demand of w_0. The occurrence and duration of a strike signal to the union that perhaps the firm is not as profitable as the union thought it was and encourages the union to moderate its demands. Moreover, the union rank and file may find it difficult to pay their bills during a long strike, further moderating union wage demands. The longer the strike, the lower the wage the union demands. Eventually, the union wage demand falls to $w_{\min}$, the lowest wage the union is willing to accept.

The firm knows that the union will moderate its demands over time. Even though the firm would obviously have a lower payroll if it waited out the strike, strikes are costly. The firm will then want to compare the present value of profits if it gives in to the union's initial wage demands with the present value of profits if the strike lasts 1 week, or if the strike lasts 2 weeks, and so on. *The firm then chooses the strike duration that maximizes the*

[24] Orley C. Ashenfelter and George E. Johnson, "Bargaining Theory, Trade Unions, and Industrial Strike Activity," *American Economic Review* 74 (March 1969): 35–49. See also Beth Hayes, "Unions and Strikes with Asymmetric Information," *Journal of Labor Economics* 2 (January 1984): 57–83; and John Schnell and Cynthia Gramm, "Learning by Striking: Estimates of the Teetotaler Effect," *Journal of Labor Economics* 5 (April 1987): 221–241.

present value of profits. This choice is determined by a simple trade-off: If the firm gives in too quickly, the increased payroll costs eat away at profits; if the firm waits too long to settle, the costs of the strike can be substantial.

Figure 10-10 shows how the "optimal" length of the strike is determined. The firm's profit opportunities are summarized by isoprofit curves. The isoprofit curve labeled π_A gives the various combinations of wage settlements and strike durations that generate A dollars' worth of profits. The isoprofit curve must be downward sloping because the firm is indifferent between long and short strikes only if the long strikes lead to a lower settlement wage. And a lower isoprofit curve yields higher profits because, for any given strike duration, the firm is paying a lower wage. Hence, the isoprofit curve labeled π_B in the figure indicates a higher level of profits than the isoprofit curve π_A.

As drawn, the isoprofit curve π_A gives the firm's profits if the firm accepts the union's initial demand. Suppose that the firm knows the shape of the union's wage resistance curve. The firm will then choose the point along that curve that maximizes profits. The firm moves to the lowest possible isoprofit curve and maximizes profits by choosing the point of tangency between the isoprofit curve and the union resistance curve, or point P in the figure. The "optimal" strike—that is, the strike that maximizes the firm's profits for a given union resistance curve—lasts t weeks, and the settlement wage equals w_t dollars.

Empirical Determinants of Strike Activity

Although strikes receive a disproportionate amount of media attention, the extent of strike activity in the United States has declined precipitously since the 1950s. In 1955, for example, 2.1 million workers were involved in strikes. By 2010, only 45,000 workers were involved.

The existing research on strike activity, mostly using data from those decades when strike activity was more frequent, suggests that unions indeed moderate their demands in response to the length of the strike, as suggested by the union resistance curve. The settlement wage falls by about 2 percent after a 50-day strike and by about 4 percent after a 100-day strike.[25]

A key assumption of models used to understand why strikes occur is that the firm knows more about its financial conditions than the workers do. And it seems that indeed strikes are more likely to occur when unions are uncertain about the firm's finances. In particular, there is a positive correlation between strike activity and volatility in the firm's stock value.[26] Volatility in the stock market reflects the investors' (and, therefore, the workers') uncertainty about the firm's financial condition.

The cost of a strike, in terms of forgone output and revenues, is an important deterrent to strike activity. For the typical firm, the cost is substantial and quickly reflected in the firm's market value, as a strike reduces the value of shareholders' wealth.[27]

[25] Sheena McConnell, "Strikes, Wages, and Private Information," *American Economic Review* 79 (September 1989): 801–815; David Card, "Strikes and Wages: A Test of an Asymmetric Information Model," *Quarterly Journal of Economics* 105 (August 1990): 625–659.

[26] Joseph S. Tracy, "An Empirical Test of an Asymmetric Information Model of Strikes," *Journal of Labor Economics* 5 (April 1987): 149–173.

[27] Brian Becker and Craig Olson, "The Impact of Strikes on Shareholder Equity," *Industrial and Labor Relations Review* 39 (April 1986): 425–438; and John DiNardo and Kevin F. Hallock, "When Unions "Mattered": Assessing the Impact of Strikes on Financial Markets," *Industrial and Labor Relations Review* 55 (January 2002): 219–233.

Theory at Work

THE COST OF LABOR DISPUTES

In August 2000, Firestone and Ford recalled 14.4 million tires. At the time of the recall, more than 6 million of these tires were still on the road, mostly on Ford Explorers. The National Highway Traffic Safety Administration (NHTSA) reported that the tire models being recalled were associated with tire failures that had led to 271 fatalities and more than 800 injuries. The most common source of failure was tread separation, a defect that causes the tire to blow out when the rubber tread detaches from the steel belts.

It turned out that many of the recalled tires had been produced in Bridgestone/Firestone factories engulfed in a heated labor dispute. After Bridgestone/Firestone insisted on moving workers from an 8-hour to a 12-hour shift and on cutting pay for new hires by 30 percent, 4,200 workers went on strike in July 1994. The strike affected workers in three of Bridgestone/Firestone's 11 North American plants, including one in Decatur, Illinois. The company hired replacement workers. By May 1995, the Decatur plant employed 1,048 replacement workers and 371 permanent workers who had crossed the picket line. The Decatur plant is significant because it manufactured nearly a third of the tires involved in the recall, and its tires had the highest rate of defects. In May 1995, almost a year after the strike began, the union offered to return to work without a contract, but Firestone announced that it would permanently retain the replacement workers. A final agreement, which included provisions to recall all workers, was not reached until December 1996.

The working conditions for the recalled workers were difficult. A document produced by the United Steel Workers claims that "the strikers were assigned to the hardest jobs on the worst machines, rather than the jobs they had held for 10, 20, and even 30 years. The company supervisors had a field day harassing, intimidating, and firing union members for the smallest infractions." The bitterness was equally strong on the union side. The union imposed a $4,500 fine on workers who crossed the picket line if they wanted to rejoin the union.

Tire manufacturing is a complex, labor-intensive task. The production line at the Decatur plant was not automated, so workers had some discretion in determining how much effort to put into wrapping the steel belts. It turns out that "one of every 400 tires produced in the Decatur, IL, plant in 1995 was returned under warranty because of a tread separation by 2000." In fact, the tires manufactured at Decatur during the labor dispute had higher failure rates than tires produced at that facility before or after the dispute, and higher than tires produced at other plants.

Source: Alan B. Krueger and Alexandre Mas, "Strikes, Scabs, and Tread Separations: Labor Strife and the Production of Defective Bridgestone/Firestone Tires," *Journal of Political Economy* 112 (April 2004): 253–289.

However, it is important to stress the difference between the "private" cost of a strike, which is borne by the firm and the affected workers, and the "social" cost, which includes the forgone output in the aggregate economy and adverse spillover effects on other industries.

The perception that the social cost may be substantial led to the enactment of the "cooling-off provision" in the Taft–Hartley Act of 1947. This provision gives the president the power to declare an 80-day cooling-off period during which the union and firm can continue to negotiate and reach agreement. The most famous use of the provision was in 1959, when President Eisenhower invoked it to end a 116-day steel strike. Most recently, President George W. Bush invoked it in October 2002 when he ordered the Pacific Maritime Association to end its lockout of 10,500 dockworkers at 29 West Coast ports.

10-7 Union Wage Effects

By how much do unions increase the wages of their members?[28] We begin by defining precisely what we mean by a "union wage effect." Suppose that a particular worker i earns w_N^i if he works at a nonunion job but would earn w_U^i if the firm became unionized. The percent wage gain *for this worker* is

$$\Delta_i = \text{Percent wage gain for worker } i = \frac{w_U^i - w_N^i}{w_N^i} \qquad \textbf{(10-3)}$$

Suppose that there are k workers in the labor market. We could then calculate how much each of the workers would gain if they became unionized and define the **union wage gain** as the average

$$\text{Union wage gain} = \frac{\Delta_1 + \Delta_2 + \cdots + \Delta_k}{k} \qquad \textbf{(10-4)}$$

Although we are interested in knowing the size of the union wage gain in equation (10-4), this statistic is very difficult to calculate. We need to know how much a worker would earn in a nonunion job and how much he would earn if the job suddenly became unionized. Typically, we observe only one of these two wages (that is, the job is either unionized or not). As a result, we instead calculate a very different sort of union–nonunion wage differential. Suppose that the average wage in union jobs is $\overline{w}_U$ and the average wage in nonunion jobs is $\overline{w}_N$. The **union wage gap** is defined by

$$\text{Union wage gap} = \frac{\overline{w}_U - \overline{w}_N}{\overline{w}_N} \qquad \textbf{(10-5)}$$

which is the percent wage differential between union jobs and nonunion jobs.

Estimates of the union wage gap typically adjust for differences in socioeconomic characteristics (such as education, age, industry, and region of employment) between workers in union jobs and workers in nonunion jobs. These adjustments are similar to those used in the Oaxaca–Blinder decomposition introduced in the chapter on labor market discrimination. Although the union wage gap gives the wage differential between workers who are in union jobs and comparably skilled workers who are in nonunion jobs, we will see below that the union wage gap may have little to do with the union wage gain.

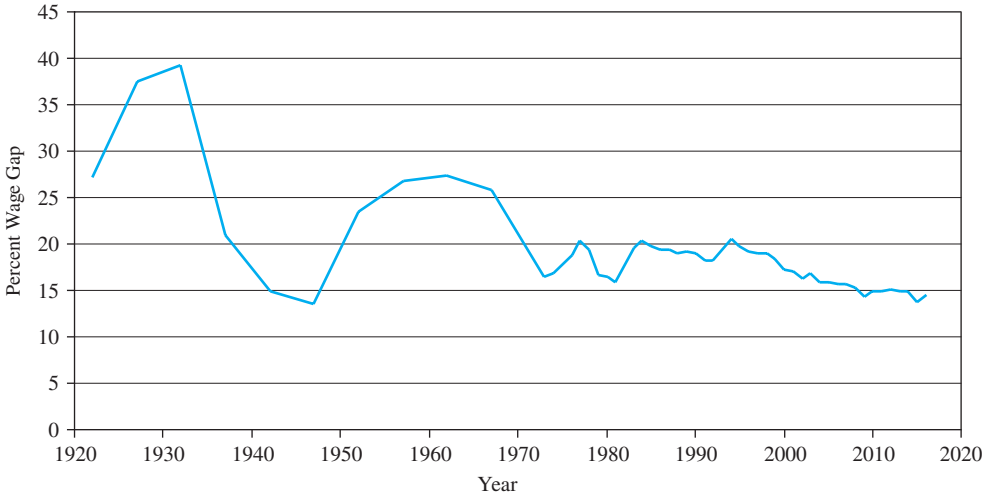
Figure 10-11 illustrates the trend in the union wage gap between 1920 and 2016. The union wage gap is large in some time periods but narrows substantially in others. During the early 1930s, union members earned about 39 percent more than nonunion members. Since the 1970s, the union wage gap has hovered in the 15–20 percent range. In 2016, the union wage gap stood at 14.5 percent.[29]

[28] A comprehensive summary of the large literature addressing this question is given by H. Gregg Lewis, *Union Relative Wage Effects: A Survey*, Chicago: University of Chicago Press, 1986.

[29] This statistic gives the wage gap between workers in union and nonunion firms, holding constant the worker's education, age, gender, region of residence, metropolitan status, industry of employment, and occupation.

FIGURE 10-11 The Union Wage Gap, 1920–2016

Sources: The pre-1970 data are from John Pencavel and Catherine E. Hartsog, "A Reconsideration of the Effects of Unionism on Relative Wages and Employment in the United States, 1920–1980," *Journal of Labor Economics* 2 (April 1984): 193–232. The post-1970 data are from Barry T. Hirsch and David A. Macpherson, *Union Membership and Earnings Data Book: Compilations from the Current Population Survey (2017 Edition),* Washington, DC: Bureau of National Affairs, 2017, Table 2a.



The wage distribution of unionized workers not only has a higher mean than that of nonunion workers, but it has less variance as well.[30] The compression of the wage distribution in the union sector arises partly because union workers are a more homogeneous group (in terms of education and other observable characteristics) than nonunion workers. Unionized firms, however, also offer their workers a lower payoff for skills, probably because unions emphasize equity in collective bargaining negotiations. This emphasis prohibits employers from making wage-setting decisions that reward very productive workers and penalize less-productive ones.

Finally, the wage is only part of the worker's compensation package. Unions also affect the value of the fringe benefits offered by firms, such as health and life insurance, vacation and sick days, pensions, and bonuses. The ratio of the value of fringe benefits to the wage is 20 percent in unionized firms and 15 percent in nonunion firms.[31] Because union wages are higher than nonunion wages, the value of the fringe benefit package is larger for union workers than for nonunion workers. As a result, the "union compensation gap" (that is, the percent difference in total compensation) may be about 2–3 percentage points higher than the union wage gap.

[30] Richard B. Freeman, "Unionism and the Dispersion of Wages," *Industrial and Labor Relations Review* 34 (October 1980): 3–23; Richard B. Freeman, "Union Wage Practices and Wage Dispersion within Establishments," *Industrial and Labor Relations Review* 36 (October 1982): 3–21; David Card, "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis," *Econometrica* 64 (July 1996): 957–979.

[31] Richard B. Freeman, "The Effect of Unionism on Fringe Benefits," *Industrial and Labor Relations Review* 34 (July 1981): 489–509; Thomas C. Buchmueller, John DiNardo, and Robert G. Valletta, "Union Effects on Health Insurance Provision and Coverage," *Industrial and Labor Relations Review* 55 (July 2002): 610–627.

Does the Union Wage Gap Measure the Union Wage Gain?

The union wage gap measures the wage differential between similarly skilled workers in the union and nonunion sectors. But can this wage gap be interpreted as a measure of the union wage gain? In other words, does the fact that the typical union worker earns about 15 percent more than the typical nonunion worker imply that if *we* became unionized, we also would earn 15 percent more? The answer is no.

Suppose that a union contract forces the firm to pay 15 percent more than the competitive wage. The collective bargaining agreement typically makes it difficult for the firm to fire or lay off workers. Because of the high cost of labor and because the firm is stuck with the workers it hires, the unionized firm will want to screen job applicants very carefully. Moreover, the 15 percent wage premium encourages many workers to apply for jobs at the unionized firm. The firm can then pick and choose from the applicant pool. Over time, the firm's workforce will be composed of workers who are relatively more productive than workers in nonunion firms.

The union wage gap is typically estimated by comparing observationally equivalent workers in union and nonunion jobs. Because observable measures of skills do not completely account for skill differences among workers, the typical worker in a union job may well be more productive than a seemingly comparable worker in a nonunion job. The union wage gap, therefore, overestimates the union wage gain.

There have been many attempts to address the problem that different types of workers end up in union and nonunion jobs. For example, we can estimate "selectivity-corrected" earnings regressions to calculate the union wage gap.[32] In theory, this methodology lets us to predict what a union worker would earn if he were to work in a nonunion job and what a nonunion worker would earn if he were to join a union. But the evidence provided by this method is mixed; sometimes the union wage gain is improbably high (greater than 50 percent) and sometimes it is ridiculously low (suggesting that unions reduce wages).

An alternative approach estimates the union wage gain by tracking a particular worker over time, so that a worker can be observed either entering or leaving a union job.[33] The union wage gain is then given by the observed wage change as the worker changes jobs. The union wage gain implied by this exercise tends to be smaller than the union wage gap (10 percent versus 15 percent).

However, this type of tracking analysis views a worker's move between union and nonunion jobs as if it were a natural experiment, with a person randomly switching from one job to another. But workers are picky when they decide which job offers to accept and which to reject. A worker who trades a highly paid union job for a lower-wage nonunion job is providing very relevant information about other job characteristics, making it unlikely that the tracking of workers actually estimates the union wage gain.

[32] Greg Duncan and Duane Leigh, "Wage Determination in the Union and Nonunion Sectors: A Sample Selectivity Approach," *Industrial and Labor Relations Review* 34 (October 1980): 24–34; Chris Robinson and Nigel Tomes, "Union Wage Differentials in the Public and Private Sectors: A Simultaneous Equations Specification," *Journal of Labor Economics* 2 (January 1984): 106–127.

[33] Richard B. Freeman, "Longitudinal Analysis of the Effects of Trade Unions," *Journal of Labor Economics* 2 (January 1984): 1–26; George Jakubson, "Estimation and Testing of the Union Wage Effect Using Panel Data," *Review of Economic Studies* 58 (October 1991): 971–992.

A recent study uses data from certification elections in the 1980s and 1990s to present a novel way of estimating the wage impact of unions.[34] In particular, it compares the wage evolution in firms where the union barely won the certification election with the wage evolution in firms where the union barely lost the election. Surprisingly, the comparison of the pre- and post-election wages in both types of firms suggests that the wage impact of unions is very small. Wages changed by roughly the same amount regardless of the election outcome, implying a union wage gain close to zero.

Threat and Spillover Effects

The calculation of union wage effects typically assumes that unions do not affect the wage in the nonunion sector. Unions, however, are likely to influence the wage of both union and nonunion workers. As a result, calculating the wage differential between union jobs and nonunion jobs will not measure the union wage gain (even in the absence of selection biases).

Threat effects provide one channel through which unions influence wages in the nonunion sector. Profit-maximizing employers in the industry have an incentive to keep the union out and might be willing to share some of the excess rents in the hope that the workers will not unionize.[35] Threat effects imply that unions, through their mere existence, have a positive impact on nonunion wages. The wage differential between union and nonunion jobs would then underestimate the true impact of the union on the wage.

Unions also might have **spillover effects** on the nonunion sector. As workers lose their jobs in unionized firms (perhaps because firms move up along the demand curve in response to the union-mandated wage increase), the supply of workers in the nonunion sector increases and the competitive wage falls. A comparison of the wage between union and nonunion jobs would overestimate the impact of the union on the wage of unionized workers.

There is evidence for both threat effects and spillover effects.[36] The wage of nonunion police is higher in metropolitan areas where a powerful police union exists, suggesting the existence of threat effects. At the same time, the wage of nonunion workers is lower in cities that have high unionization rates, suggesting the existence of spillover effects.

An extreme example of how the union affects the wage of nonunion workers is given by the provision in the Davis–Bacon Act of 1931 requiring that workers employed in federally subsidized construction projects be paid a "prevailing wage." The U.S. Department of Labor has typically interpreted the prevailing wage to be the union wage. The prevailing wage provision may have increased the cost of construction projects by perhaps as much as 25 percent.[37]

[34] John DiNardo and David S. Lee, "Economic Impacts of New Unionization on Private Sector Employers: 1984–2001," *Quarterly Journal of Economics* 119 (November 2004): 1383–1441.

[35] Sherwin Rosen, "Trade Union Power, Threat Effects, and the Extent of Organization," *Review of Economic Studies* 36 (April 1969): 185–196.

[36] Richard B. Freeman and James L. Medoff, "The Impact of the Percentage Organized on Union and Nonunion Wages," *Review of Economics and Statistics* 63 (November 1981): 561–572; Casey Ichniowski, Richard Freeman, and Harrison Lauer, "Collective Bargaining Laws, Threat Effects and the Determinants of Police Compensation," *Journal of Labor Economics* 7 (April 1989): 191–209; and Henry Farber, "Nonunion Wage Rates and the Threat of Unionization," *Industrial and Labor Relations Review* 58 (April 2005): 335–352.

[37] Martha Fraundorf, John Farrell, and Robert Mason, "The Effect of the Davis-Bacon Act on Construction Costs in Rural Areas," *Review of Economics and Statistics* 66 (February 1984): 142–146; see also Steven Allen, "Much Ado about Davis-Bacon: A Critical Review and New Evidence," *Journal of Law and Economics* 6 (October 1983): 707–736.

Theory at Work

OCCUPATIONAL LICENSING

The precipitous decline in the fraction of workers who are unionized in the United States does not necessarily imply that American workers are less protected against the vagaries of labor market competition. At the same time that unions were collapsing in the private sector, there was a substantial increase in the number of workers who were required by the federal, the state, or the local government to obtain a license to do their work. Examples of jobs that require a license include such diverse occupations as physician, accountant, and attorneys, as well as barbers, manicurists, and massage therapists.

In *Capitalism and Freedom,* Milton Friedman proposed an influential theory of licensing in the labor market. Friedman emphasized that the incumbents in a particular occupation have an incentive to create a set of formal standards that limit entry into the occupation, and to lobby legislatures to enact such barriers. The licensing agency, in effect, is "captured" by the incumbents in the occupation. As a result, the agency will take actions that restrict entry and that raise the occupation's wage.

Fewer than 5 percent of workers were required to be licensed in the early 1950s. Remarkably, almost 30 percent of workers are now required to have a license to perform their jobs. The entry barriers raised the wage of the protected incumbents by 10–15 percent, even after adjusting for differences in skills between licensed and unlicensed workers. It is interesting that the wage effect resulting from licensing is identical to the union wage gap.

Source: Morris M. Kleiner and Alan B. Krueger, "Analyzing the Extent and Influence of Occupational Licensing on the Labor Market," *Journal of Labor Economics* 31 (April 2013): S173–S202.

10-8 Policy Application: Public-Sector Unions

There has been a rapid increase in the proportion of public-sector workers in the United States who are unionized. Much of the research on the economic impact of public-sector unions is motivated by the fact that labor demand curves for many essential public-sector workers—such as police officers, firefighters, and teachers—may be inelastic. If public-sector unions behaved like monopoly unions (so that wage–employment outcomes lie on the labor demand curve), Marshall's rules of derived demand imply that public-sector unions could "extort" very high wages from taxpayers. And because public-sector workers often become a potent political force, some politicians might be willing to grant high wages to those workers in exchange for political support.

However, state and local governments *do* face constraints. A wage increase for public-sector workers has to be funded by taxpayers, and higher taxes will encourage the outmigration of jobs and workers from the locality. In effect, governmental units compete with each other to attract residents and business opportunities, and this competition helps keep down the cost of public services.

The evidence suggests that the union wage effect in the public sector differs greatly across occupations, depending on whether the worker is, for example, a police officer or a

teacher. The union wage gap for police officers and firefighters is about 14 percent, while the gap for teachers and other public-sector workers is 4 percent.[38]

Teacher Unions and Student Outcomes

An important focus of the policy debate over public sector unions has been the teachers' unions. Depending on one's perspective, teachers' unions either help provide quality education to millions of students or funnel millions of dollars in contributions to politicians who just happen to back the union objectives after they get elected. And those objectives are often to set higher wages and benefits for teachers.

Most states explicitly prohibited collective bargaining by teachers in 1960. Between 1960 and 1990, some states extended collective bargaining rights to teachers. But there were significant differences in the timing of the liberalizing legislation, with some states (such as California and New York) granting collective bargaining rights before 1970, while other states (such as Connecticut and Illinois) granting such rights only after 1980.

An influential study exploits the differences in the timing of these laws to determine how teachers' unions affect a variety of outcomes in the education system.[39] Not surprisingly, the creation of a teachers' union has widespread effects. For instance, per-pupil spending increases by about 12 percent. Some of this increase goes to teachers directly through a pay raise of about 5 percent (presumably allowing the hiring of better teachers). Some of the increased spending goes to hiring more teachers, so that the pupil–teacher ratio falls.

Surprisingly, the data also suggest that despite the fact that there are more teachers and that these teachers are paid more, the academic achievement of students does not improve. Instead, the dropout rate *increases* by 2 percentage points. It seems that adding more inputs to the education production function, such as more and better teachers or higher per-pupil spending, is not very effective in the rigid work environment implied by a unionized labor market.[40]

Arbitration

The power of public-sector unions is also constrained because most states prohibit strikes by public-sector workers. Public-sector unions often use binding arbitration as a way of resolving collective bargaining disputes.

Two types of arbitration procedures are in widespread use. Under **conventional arbitration**, the two parties to the dispute present their offers to an objective arbitrator. The arbitrator, who is effectively the judge in the case, compares the two offers. After studying the facts, he comes up with a solution that both sides must accept. The arbitrator's solution might lie anywhere in between the two offers or might even lie outside this range. In **final-offer arbitration**, both sides again present their offers to the arbitrator,

[38] Richard B. Freeman and Eunice S. Han, "Public Sector Unionism without Collective Bargaining," Harvard University Working Paper, 2012, Table 3. See also Robert G. Valletta, "Union Effects on Municipal Employment and Wages: A Longitudinal Approach," *Journal of Labor Economics* 11 (July 1993): 545–574; and Jan Brueckner and David Neumark, "Beaches, Sunshine, and Public-Sector Pay: Theory and Evidence on Amenities and Rent Extraction by Government Workers," *American Economic Journal: Economic Policy* 6 (May 2014): 198–230.

[39] Caroline Minter Hoxby, "How Teachers' Unions Affect Education Production," *Quarterly Journal of Economics* 111 (August 1996): 671–718.

[40] Some conflicting evidence is presented in Michael F. Lovenheim, "The Effect of Teachers' Union on Education Production: Evidence from Union Election Certifications in Three Midwestern States," *Journal of Labor Economics* 27 (October 2009): 525–587.

but the arbitrator chooses from one of the two offers. Again, both sides are bound to accept the arbitrator's decision.

Because wage settlements in the public sector depend so heavily on the arbitrator's judgment, both employers and unions have incentives to make strategic offers designed to influence the arbitrator's behavior.[41]

In the typical model of conventional arbitration, employers and unions have beliefs about what the arbitrator considers to be a reasonable outcome. Both parties suspect that if they present an outlandish offer to the arbitrator (too high a wage demand in the case of the union or too low a wage offer in the case of the employer), the arbitrator will disregard their position and the arbitrator's decision will be greatly influenced by the other party's offer. Both parties will then position themselves around what they believe to be the arbitrator's desired outcome. The arbitrator, in effect, only needs to "split the difference" between the offers.

Final-offer arbitration introduces different incentives. After studying the facts of the case, the arbitrator again has a notion of what constitutes a fair settlement. The arbitrator will then choose whichever offer comes closest to his or her assessment. Obviously, both the employer and the union will avoid making offers that deviate greatly from the arbitrator's preferred outcome. After all, arbitrators will completely ignore outlandish offers. But parties who are risk-averse and are not willing to take a chance with the arbitrator will make offers that are very close to the arbitrator's preferred position and will tend to "win" a higher fraction of final-offer awards. As a result, the fact that one party, say, the union, wins most of the cases need not indicate a systematic bias on the part of the arbitrator. It might just indicate that unions are more risk-averse than firms.

A number of studies have analyzed how arbitration affects the wage of police officers in New Jersey.[42] Among the disputes that reached mandated final-offer arbitration, the typical employer offered only a 5.7 percent increase in compensation, whereas the typical union wanted an 8.5 percent wage increase. The union "won" about two-thirds of the time.

It is useful to compare this track record with settlements reached in comparable disputes under conventional arbitration. In those disputes, the arbitrator typically awarded the union an 8.3 percent wage increase. There is little difference, therefore, in the average award made under conventional and final-offer arbitration. If we interpret the conventional arbitration award as a measure of the "preferred" settlement, it is evident that the union was more risk-averse than the firm and made more reasonable offers to the arbitrator (if the dispute was settled through final arbitration).

Summary

- There has been a precipitous decline in private-sector union membership in the United States since the mid-1960s. This decline is attributable partly to structural changes in the U.S. economy, including the shrinking of the manufacturing sector and the movement of the population to southern and western states. At the same time, union membership in the public sector rose rapidly.

[41] Henry S. Farber and Harry C. Katz, "Interest Arbitration, Outcomes, and Incentives to Bargain," *Industrial and Labor Relations Review* 33 (October 1979): 55–63; Henry S. Farber, "Splitting-the-Difference in Interest Arbitration," *Industrial and Labor Relations Review* 35 (October 1981): 70–77.

[42] Orley C. Ashenfelter and David E. Bloom, "Models of Arbitrator Behavior: Theory and Evidence," *American Economic Review* 74 (March 1984): 111–124; Janet Currie, "Arbitrator Behavior and the Variances of Arbitrated and Negotiated Wage Settlements," *Journal of Labor Economics* 12 (January 1994): 29–39.

- Monopoly unions choose a wage, and firms respond to that wage demand by moving along the labor demand curve.
- The wage–employment outcome in the model of monopoly unions is inefficient in two distinct ways. First, unions distort the allocation of labor in the economy. The deadweight loss created by this distortion in the allocation of resources is small, perhaps on the order of $19 billion annually. A second type of inefficiency arises because both firms and workers can be made better off by moving off the demand curve.
- The contract curve summarizes the wage–employment combinations that are off the demand curve and that exhaust the gains from bargaining. Once a deal is struck on the contract curve, deviations from this point improve the welfare of one of the parties only at the expense of the other.
- If contract curves are not vertical, unionized firms will still distort the allocation of labor in the economy. If contract curves are vertical, unionized firms hire the "right" number of workers and the only impact of unions is to transfer part of the firms' rents to workers.
- Strikes are irrational if both parties have reasonably good information about the costs and the likely outcome of the strike. Strikes might nevertheless occur if one of the parties is better informed about the financial conditions of the firm.
- The union wage gain gives the percentage wage increase if a randomly chosen worker in the economy were to join a union. The union wage gap gives the percentage wage differential between workers in union firms and workers in nonunion firms.
- The union wage gap is around 15 percent, but the union wage gap may not provide a good estimate of the union wage gain.

Key Concepts

certification elections, *342*
contract curve, *355*
conventional arbitration, *368*
decertification elections, *342*
efficient contract, *355*
featherbedding practices, *355*
final-offer arbitration, *368*
Hicks paradox, *359*
monopoly unions, *350*
Pareto optimal, *355*
right-to-work laws, *342*
spillover effects, *366*
strongly efficient contract, *355*
threat effects, *366*
unfair labor practices, *342*
union resistance curve, *360*
union wage gain, *363*
union wage gap, *363*
yellow-dog contracts, *342*

Review Questions

1. What factors account for the decline in private-sector unionism in the United States since the mid-1960s? What factors account for the rapid increase in public-sector unionism during the same period?
2. What does it mean to say that a union has a utility function? How exactly is this utility function derived from the preferences of the workers?
3. Describe the wage–employment outcome in a model of monopoly unions. Explain why (and in what sense) this wage–employment outcome is inefficient.

4. Describe how we calculate the percentage decline in national income resulting from the misallocation of labor in a model of monopoly unions. What is the dollar value of this allocative inefficiency if unions and firms negotiate an efficient contract and the contract curve is vertical?
5. Discuss how both unions and firms can be better off if they move off the demand curve. Derive the contract curve.
6. Discuss the difference between efficient contracts and strongly efficient contracts.
7. What is the Hicks paradox?
8. Describe how employers "choose" the optimal length of a strike in a model where there is asymmetric information.
9. Define the union wage gain and the union wage gap. Why should we care about the magnitude of the union wage gain? Why should we care about the magnitude of the union wage gap? Under what conditions will the union wage gap provide a reasonable estimate of the union wage gain?
10. What are threat and spillover effects? How do they bias our estimates of the union wage effect?
11. What is conventional arbitration? What is final-offer arbitration? How do the union and firm take into account the arbitrator's behavior when deciding which wage offers to put on the table?

Problems

10-1. Suppose the firm's labor demand curve is given by

$$w = 20 - 0.01E,$$

where w is the hourly wage and E is the level of employment. Suppose also that the union's utility function is given by

$$U = w \times E.$$

It is easy to show that the marginal utility of the wage for the union is E and the marginal utility of employment is w. What wage would a monopoly union demand? How many workers will be employed under the union contract?

10-2. Suppose the union in problem 10-1 has a different utility function. In particular, its utility function is given by

$$U = (w - w^*) \times E$$

where w^* is the competitive wage. The marginal utility of a wage increase is still E, but the marginal utility of employment is now $w - w^*$. Suppose the competitive wage is \$8 per hour. What wage would a monopoly union demand? How many workers will be employed under the union contract? Contrast your answers to those in problem 10-1. Can you explain why they are different?

10-3. Figure 10-2 demonstrates some of the tradeoffs involved when deciding to join a union. Suppose in addition to higher wages the union negotiates a 10 percent employer contribution to a defined contribution pension plan. Provide a graph similar to Figure 10-2 that incorporates this retirement benefit into the decision of whether to join a union. Show on your graph how additional fringe benefits such as a retirement plan may cause the worker to be more inclined to join the union.

10-4. Consider a two-sector economy with homogeneous labor and jobs in both sectors. Two million workers supply their labor perfectly inelastically. Labor demand in both sectors can be written as:

$$E_1 = 1{,}800{,}000 - 100{,}000\, w_1 \text{ and } E_2 = 1{,}800{,}000 - 100{,}000\, w_2.$$

(a) If both sectors are competitive, what is the market-clearing wage and how many workers are employed in both sectors?

(b) Suppose a labor union forms in sector 1. The union negotiates a wage of $12 per hour, and firms choose how much labor to employ. Anyone not employed in sector 1 is relegated to sector 2. How many workers will be employed in sector 1 (unionized)? How many workers will be employed in sector 2, and what wage will they receive?

(c) What is the union-wage gap in part (b)? What would the union-wage effect be if one controlled for the spillover effect?

10-5. Consider a firm that faces a constant per unit price of $1,200 for its output. The firm hires workers, E, from a union at a daily wage of w, to produce output, q, where

$$q = 2E^{1/2}.$$

Given the production function, the marginal product of labor is $1/E^{1/2}$. There are 225 workers in the union. Any union worker who does not work for the firm can find a nonunion job paying $96 per day.

(a) What is the firm's labor demand function?

(b) If the firm is allowed to specify w and the union is then allowed to provide as many workers as it wants (up to 225) at the daily wage of w, what wage will the firm set? How many workers will the union provide? How much output will be produced? How much profit will the firm earn? What is the total income of the 225 union workers?

10-6. Consider the same set-up as in problem 10-5, but now the union is allowed to specify any wage, w, and the firm is then allowed to hire as many workers as it wants (up to 225) at the daily wage of w. What wage will the union set in order to maximize the total income of all 225 workers? How many workers will the firm hire? How much output will be produced? How much profit will the firm earn? What is the total income of the 225 union workers?

10-7. Suppose the union's resistance curve is summarized by the following data. The union's initial wage demand is $10 per hour. If a strike occurs, the wage demands change as follows:

Length of Strike:	Hourly Wage Demanded
1 month	9
2 months	8
3 months	7
4 months	6
5 or more months	5

Consider the following changes to the union resistance curve and state whether the proposed change makes a strike more likely to occur, and whether, if a strike occurs, it is a longer strike.

(a) The drop in the wage demand from \$10 to \$5 per hour occurs within the span of 2 months, as opposed to 5 months.

(b) The union is willing to moderate its wage demands further after the strike has lasted for 6 months. In particular, the wage demand keeps dropping to \$4 in the 6th month, \$3 in the 7th month, etc.

(c) The union's initial wage demand is \$20 per hour, which then drops to \$9 after the strike lasts 1 month, \$8 after 2 months, and so on.

10-8. At the competitive wage of \$20 per hour, firms A and B both hire 5,000 workers (each working 2,000 hours per year). The elasticity of demand is −2.5 and −0.75 at firms A and B, respectively. Workers at both firms then unionize and negotiate a 12 percent wage increase.

(a) What is the employment effect at firm A? How has total worker income changed?

(b) What is the employment effect at firm B? How has total worker income changed?

(c) How much would the workers at each firm be willing to pay in annual union dues to achieve the 12 percent gain in wages?

10-9. Several states recently passed laws restricting bargaining rights for public employees. Most notably the changes tended to restrict the union's right to negotiate over fringe benefits such as health care and retirement benefits. What problems were these legislative changes trying to address? Even assuming such a law survives a constitutional challenge (which some did not), why might restricting bargaining rights not fully address the problems lawmakers were aiming to solve?

10-10. Suppose the economy consists of a union and a nonunion sector. The labor demand curve in each sector is given by $L = 1{,}000{,}000 - 20w$. The total (economy-wide) supply of labor is 1,000,000, and it does not depend upon the wage. All workers are equally skilled and equally suited for work in either sector. A monopoly union sets the wage at \$30,000 in the union sector. What is the union wage gap? What is the effect of the union on the wage in the nonunion sector?

10-11. In Figure 10-6, the contract curve is *PZ*.

(a) Does point *P* represent the firm or the workers having all of the bargaining power? Does point *Z* represent the firm or the workers having all of the bargaining power? Explain.

(b) Suppose the union has the power to be a monopoly union in setting wages if it chooses, but it doesn't have the power to force a wage and an employment level on the firm. On what portion of the contract curve *PZ* would you expect the bargained wage-employment contract to occur?

10-12. Consider the following data on union versus nonunion wage and fringe benefit compensation.

	Average Hourly Wage	Average Hourly Fringe Benefit	Total Hourly Compensation
Union Workers	\$21.91	\$13.69	\$35.60
Nonunion Workers	\$17.66	\$6.85	\$24.51

Calculate the union effect for hourly wages, hourly fringe benefits, and total hourly compensation. What might you infer from the various union-negotiated effects?

10-13. Use a graph similar to Figure 10-10 to demonstrate the likely bargaining outcomes of three industries, all with identical union resistance curves.

(a) Firm A has been losing money recently as wages and fringe benefits have risen from 63 to 89 percent of all costs in just the last 3 years.

(b) Most of firm B's revenues come from supplying a product to three customers who use the product in their manufacturing of computers using a just-in-time inventory system.

(c) Firm C is a local government that finds itself negotiating with its unionized employees. Government officials are pleased with the employees' productivity, but they also face local pressure to keep taxes low.

10-14. Major League Baseball players are not eligible for arbitration or free agency until they have been in the league for several years. During these "restricted" years, a player can only negotiate with his current team. Consider a small-market team that happens to own the rights to last year's Rookie-of-the-Year. This player is currently under contract for $500,000 for the next 3 years. Because his current team is in a small market, the player's value to his current team is $6 million per year (now and in the future). When the player becomes eligible for free agency, he will likely command $10 million per year for 7 years in free agency from competing large-market teams. In the questions below, assume the player wants to maximize his lifetime earnings.

(a) What is the worst 10-year contract extension from the player's point of view that the player would accept from his current team?

(b) What is the best 10-year contract extension from the player's point of view that his current team would offer him?

(c) Would you expect this player to sign a contract extension or to play out his contract and enter free agency 3 years from now?

10-15. Recently the National Football League Players Association (NFLPA), which is the union for the players in the National Football League (NFL), and the team owners (the NFL) experienced a labor impasse in the form of a lockout. For the record, each year about 150 players (called rookies) enter the NFL and 150 veteran players exit the league (via retirement or not making a team roster). While renegotiating the most recent labor settlement, the union took several stances. Explain why a union of players would advocate against:

(a) Expanding the number of games played.

(b) Expanding the size of team rosters.

(c) A team salary cap.

(d) A rookie salary cap.

Selected Readings

John M. Abowd, "The Effect of Wage Bargains on the Stock Market Value of the Firm," *American Economic Review* 79 (September 1989): 774–800.

Orley C. Ashenfelter and George E. Johnson, "Bargaining Theory, Trade Unions, and Industrial Strike Activity," *American Economic Review* 74 (March 1969): 35–49.

John DiNardo and David S. Lee, "Economic Impacts of New Unionization on Private Sector Employers: 1984–2001," *Quarterly Journal of Economics* 119 (November 2004): 1383–1441.

Henry S. Farber and Bruce Western, "Accounting for the Decline of Unions in the Private Sector, 1973–1998," *Journal of Labor Research* 22 (Summer 2001): 459–486.

Caroline Minter Hoxby, "How Teachers' Unions Affect Education Production," *Quarterly Journal of Economics* 111 (August 1996): 671–718.

Alan B. Krueger and Alexandre Mas, "Strikes, Scabs, and Tread Separations: Labor Strife and the Production of Defective Bridgestone/Firestone Tires," *Journal of Political Economy* 112 (April 2004): 253–289.

David S. Lee and Alexandre Mas, "Long-Run Impacts of Unions on Firms: New Evidence from Financial Markets, 1961–1999," *Quarterly Journal of Economics* 127 (February 2012): 333–378.

Thomas E. MaCurdy and John H. Pencavel, "Testing between Competing Models of Wage and Employment Determination in Unionized Markets," *Journal of Political Economy* 94 (June 1986): S3–S39.

Chapter 11

Incentive Pay

I like work; it fascinates me. I can sit and look at it for hours.

—Jerome K. Jerome

Throughout this book, we have examined the nature of the employment contract in what are called **spot labor markets**. In each period, firms decide how many workers to hire at given wages; workers decide how many hours to work; and the interaction of workers and firms determines the equilibrium wage and employment. Once the market "shouts out" the equilibrium wage, workers and firms make the relevant labor supply and labor demand decisions. In these spot labor markets, the wage equals the worker's value of marginal product.

One problem with this simple story of how spot labor markets work is that the nature of the employment contract affects both the worker's productivity and the firm's profits. The details of the contract matter because employers often do not know the worker's true productivity and workers would like to get paid a high salary while putting in as little effort as possible. This chapter analyzes the employment contracts that arise to tackle the problems of incomplete information and worker shirking.

Some firms, for instance, might choose to offer workers a piece rate for their efforts, whereas other firms offer workers a fixed salary. Because the piece-rate worker's income depends strictly on how much output is produced, "she works hard for the money." If it is difficult for an employer to monitor a worker's activities, a salaried worker can get away with daydreaming, Web surfing, and endless texting.

Labor markets use a wide array of compensation systems, with piece rates and fixed salaries being only the tip of the iceberg. The employer will naturally view **incentive pay**, a compensation package designed to elicit particular levels of effort from the worker, as yet another tool it can use to increase its profits.

11-1 Piece Rates and Time Rates

The simplest way of showing the link between work incentives and the method of compensation is to compare two widely used pay systems: **Piece rates** and **time rates**.

A piece-rate system compensates the worker according to some measure of the worker's output. For example, garment workers might be paid on the basis of how many pairs of pants they produce; salespersons are paid a commission based on the volume of sales;

and California strawberry pickers are paid according to how many boxes of strawberries they fill. In 1987, "Junk Bond King" Michael Milken's salary at Drexel Burnham Lambert totaled $550 million (almost $1.2 billion in inflation-adjusted 2017 dollars). Most of this salary came from a 35 percent commission (or a piece rate) on the profits generated by his junk bond group.[1]

In contrast, the compensation of time-rate workers depends only on the number of hours the worker spends at the job and has nothing to do with the number of units she produces, at least in the short run. For example, the weekly income of a time-rate worker paid by the hour depends on the number of hours worked during the week. Over the long run, of course, the firm will look at the worker's performance record to make decisions on retention and promotion. For simplicity, we focus on the short run, assuming that the earnings of time-rate workers depend only on hours worked and not on the worker's performance.

Should a Firm Offer Piece Rates or Time Rates?

Workers differ in their productivity either because there are ability differences across workers or because some workers devote a lot of effort to the job and others do not.

Consider a firm deciding whether to offer piece rates or time rates.[2] If the firm offers a piece rate, the worker's wage will exactly equal her value of marginal product. If a translator's income is based on a payment of 10 cents per word translated, her salary would be $10,000 after translating a 100,000-word manuscript. If the firm were to offer the translator a lower piece rate per word translated, she would simply find another firm paying the competitive rate and move on.

Sometimes, however, the worker may know precisely how much she has produced, but the firm may not be so sure. Put differently, the firm cannot measure the worker's output precisely *and* cannot expect the worker to report that number truthfully. If the firm wishes to adopt a piece-rate system, the firm will have to monitor the worker constantly. But monitoring is costly, as the firm could have used these resources in other ways, such as adding capital to the production line. The monitoring cost will typically vary from firm to firm, depending on how easy or how hard it is to monitor, but could be substantial for some firms.

The firm could avoid the monitoring cost entirely by simply adopting a time-rate system, such as paying the worker a fixed salary of, say, $500 per week. At least in the short run, a firm that chooses a time-rate system does not have to monitor the worker's performance constantly.

Competitive firms choose whichever system is most profitable. Regardless of whether the monitoring costs are borne by the firm or by the worker (through a lower piece rate), firms that have very high monitoring costs will not be able to offer piece-rate systems because few workers would want to receive such low take-home pay. Firms with high monitoring costs, therefore, opt for time rates, and firms with low monitoring costs choose piece rates.

[1] Connie Bruck, *The Predators' Ball,* New York: Penguin Books, 1989, pp. 31–32.

[2] Charles Brown, "Firms' Choice of Method of Pay," *Industrial and Labor Relations Review* 43 (February 1990, Special Issue): 165S–182S; Edward P. Lazear, "Salaries and Piece Rates," *Journal of Business* 59 (July 1986): 405–431; and Robert Gibbons, "Piece-Rate Incentive Schemes," *Journal of Labor Economics* 5 (October 1987): 413–429.

It is then not surprising that piece rates are often paid to workers whose output can be observed easily (the number of pants produced, the number of boxes of strawberries picked, the dollar volume of sales). And time rates are offered to workers whose output is more difficult to measure (such as college professors or workers on a software production team).

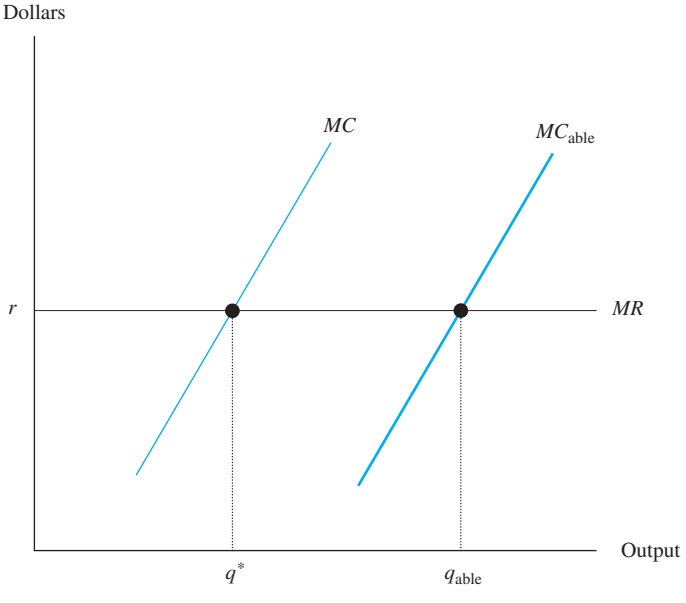
How Much Effort Do Workers Allocate to Their Jobs?

A piece-rate worker chooses how much output to produce at the firm. She produces the amount that maximizes her utility. The more output she produces, the greater her take-home salary, and the better off she is. But it takes effort to produce output and working hard causes disutility or "pain." The worker would rather surf the Web or socialize than write endless strings of computer code.[3]

Figure 11-1 illustrates the worker's decision when she is paid a constant piece rate of r dollars per unit produced. The marginal revenue of effort curve MR gives the additional income generated by producing one additional unit of output. The MR curve is horizontal because the piece rate r is constant. But producing an additional unit of output also causes pain, and this pain rises as the worker devotes more effort to the job. The marginal cost of effort curve MC is then upward sloping. A worker who wants to maximize utility produces up to the point where marginal revenue equals marginal cost, or q^* in the figure.

FIGURE 11-1 Determination of Output by Piece-Rate Workers

The piece rate is r dollars, so the marginal revenue of an additional unit of output equals r. The worker gets disutility from producing output, as indicated by the upward-sloping marginal cost of effort curve. The optimal output equates marginal revenue to marginal cost, or q^* units. If it is easier for more able workers to produce output, they face lower marginal cost curves and produce more.



[3] A simple specification of the worker's utility function could then be written as $U = rq - C(q)$, where q gives the output produced and $C(q)$ gives the psychic cost of producing that output.

Workers differ in their innate ability, so different workers behave differently. Suppose that more able workers find it easier to produce output. More able workers would then have a lower marginal cost curve (such as MC_{able} in the figure), and produce more output.

Now consider the effort decision faced by time-rate workers. Suppose there is a minimum level of output, call it $\overline{q}$, that the firm can easily monitor. For instance, the firm knows if the worker shows up for work and sits at her desk or takes her spot on the assembly line. If the worker does not achieve this minimum level of effort, she is fired. A time-rate worker will then produce $\overline{q}$ units of output, *and no more.* After all, it is painful to produce output, and the time-rate worker knows she can get away with producing the minimum amount.

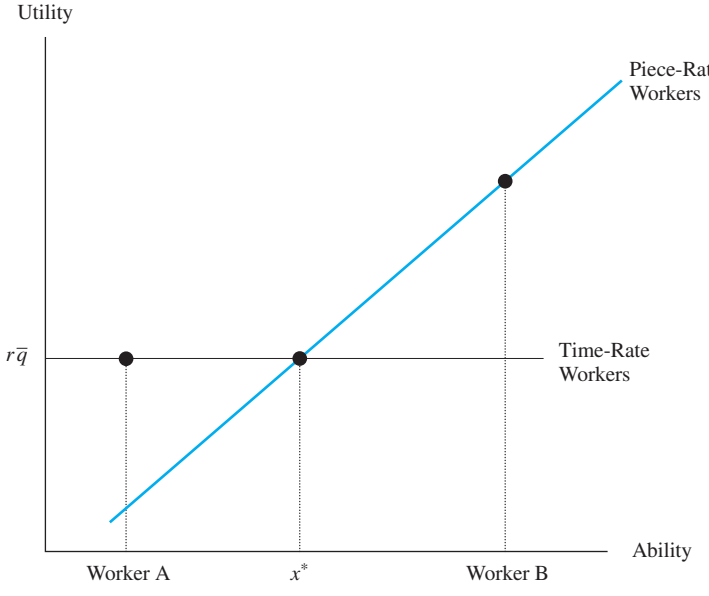
Of course, firms know that if they offer a time-rate pay system, the worker produces $\overline{q}$ units of output, and time-rate workers will then be paid the salary $r \times \overline{q}$. If we assume that there is no pain associated with simply showing up at the workplace and doing the very minimum that is expected, the utility of a time-rate worker is given by $r \times \overline{q}$.

The Sorting of Workers across Firms

Figure 11-2 illustrates the relation between a worker's utility and her ability. In the time-rate job, the worker's utility equals her income in that job (or $r \times \overline{q}$ dollars). Note that all time-rate workers, *regardless of their abilities,* get the same utility (because all workers devote the same minimal level of effort to time-rate jobs). If the worker is paid by the piece, however, her utility depends on her ability. High-ability workers produce more output, have higher incomes, and have higher utilities.

FIGURE 11-2 Sorting of Workers in Piece-Rate and Time-Rate Jobs

All workers, regardless of their ability, allocate the same minimal level of effort to time-rate jobs. Higher levels of ability increase output, income, and utility in piece-rate jobs. Workers with more than x^* units of ability choose piece-rate jobs, and less able workers choose time-rate jobs.



Workers are not indifferent between the two types of employment contracts and will sort themselves according to what is best for them. Consider the choice of a less able worker, such as worker A in Figure 11-2. This worker is better off accepting a time-rate job offer. But a very able worker (worker B) is better off working for a piece-rate firm. In fact, the figure indicates that all workers with fewer than x^* units of ability work for time-rate firms and workers with more than x^* units work for piece-rate firms.

The sorting by ability is not surprising. More able workers want to separate themselves out of the pack and choose firms that offer piece-rate systems, where their talent for producing output is better rewarded. Less able workers choose time-rate firms, where their low productivity is less easily discernible.

The available evidence indeed suggests that piece-rate workers are more productive and earn more than time-rate workers.[4] In the footwear industry, for example, piece-rate workers earn 13 percent more per hour than time-rate workers; among garment workers producing men's and boys' suits and coats, piece-rate workers earn 15 percent more; and among workers in auto repair shops, piece-rate workers earn at least 20 percent more.

Disadvantages of Using a Piece-Rate Compensation System

There are many advantages to piece-rate incentive pay. A piece rate attracts the most able workers, elicits a lot of effort from the workforce, ties pay directly to performance, minimizes the role of discrimination and nepotism, and increases the firm's productivity.

In view of these benefits, why are piece rates not used more often in the labor market? One obvious reason is that the work incentives introduced by piece rates are of little use when the firm's production depends on team effort as opposed to individual effort. Offering piece rates to one of the workers along an automobile production line would have little impact on her productivity because the speed at which the line moves depends on the productivity of all the other workers on the line. Although the firm could structure compensation in a way that offers a piece rate to the entire team based on the team's output, there is always the possibility that some team members will shirk, introducing the **free-riding problem**. Because a single worker's pay is only distantly related to her productivity, a single worker does not have much incentive to allocate effort to her job and will instead depend on the "kindness of others." Piece-rate systems, therefore, work best when the worker's own pay can be tied directly to her own productivity.

A piece-rate compensation system also tends to overemphasize the *quantity* of output produced, introducing a tradeoff between quantity and quality. This problem could be abated if the worker's earnings depend on the number of units produced that meet a well-defined quality standard. But incorporating both quality and quantity increases monitoring costs, and reduces the likelihood that firms offer piece-rate systems in the first place.

Many workers frown on piece-rate systems because their salaries might fluctuate a lot over time. The daily earnings of a strawberry picker will depend on weather conditions, and the earnings of a salesperson working on commission will depend on the aggregate

[4] Eric Seiler, "Piece Rate vs. Time Rate: The Effect of Incentives on Earnings," *Review of Economics and Statistics* 66 (August 1984): 363–376; Harry J. Paarsch and Bruce S. Shearer, "The Response of Worker Effort to Piece Rates: Evidence from the British Columbia Tree-Planting Industry," *Journal of Human Resources* 34 (Fall 1999): 643–667; and Jean-Marie Baland, Jean Dreze, and Luc Leruth, "Daily Wages and Piece Rates in Agrarian Economies," *Journal of Development Economics* 59 (August 1999): 445–461.

Theory at Work

WINDSHIELDS BY THE PIECE

The Safelite Glass Corporation is the largest installer of automobile glass in the United States. Until January 1994, glass installers were paid an hourly wage rate that was unrelated to the number of windows they installed. In 1994 and 1995, the company shifted its pay structure to a piece-rate plan. On average, installers were paid about $20 per window installed.

The company adopted an incentive pay system because it believed that the piece rate would increase worker productivity. Moreover, it was easy to monitor the actual output of each installer. A computerized system kept track of how many units a worker installed in any given week. In fact, the very detailed records mean that we have information on the number of windows *a particular worker* installed both under the old time-rate system and under the new piece-rate system.

The number of windows installed by a particular worker increased by around 20 percent after the piece-rate system went into effect. In other words, a key prediction of the theory—that piece rates elicit more effort from a worker—is strongly confirmed by Safelite's experience.

The data also reveal that there are strong sorting effects among new workers hired. The piece-rate system tends to attract high-productivity workers because these are the workers who have the most to gain from being paid their actual marginal product. Workers hired by Safelite after the piece-rate system went into effect are about 20 percent more productive than workers hired under the old pay regime.

Finally, not only were workers more productive and had higher earnings, but the firm's profits also increased.

Source: Edward P. Lazear, "Performance Pay and Productivity," *American Economic Review* 90 (December 2000): 1346–1361.

economy. If workers are risk-averse, they dislike such fluctuations. Workers will instead prefer a pay system where they can feel "insured" against these events and can be guaranteed a steady salary stream. Risk-averse workers, therefore, prefer to work in firms that offer time-rate systems. In order to attract workers, piece-rate firms would then have to compensate workers for the disutility caused by the volatility in their salaries. But this compensating differential will reduce the firm's profits, and fewer firms will choose to offer piece rates.

Finally, workers in piece-rate firms fear the well-known **ratchet effect**. Suppose that a piece-rate worker produces more output than the firm expected. The firm's managers might interpret the high level of production as evidence that the job was not quite as difficult as they thought and that they are paying too much. They respond by lowering the piece rate r and workers will then have to work harder just to keep even. For example, Soviet managers who posted high levels of productivity in response to a particular set of incentives were often accused of being lazy or "counterrevolutionary" in earlier years, with dire consequences. The ratchet effect discourages workers from accepting piece-rate jobs.

11-2 Tournaments

Most economic models of the labor market typically assume that the worker is paid according to an *absolute* measure of performance on the job. If the worker's value of marginal product is $15 an hour, the worker's wage equals $15.

In some settings, however, the labor market does not reward workers according to an absolute measure of productivity. The rewards are instead based on how a worker performs

relative to other workers. In effect, the firm holds a **tournament**, or a contest, to rank workers according to their productivity. The rewards are then distributed according to rank, with the winner receiving a sizable reward and the losers receiving much smaller payoffs.[5]

The reward structure in professional sports illustrates this type of labor market arrangement. The winner of the 2017 British Open (Jordan Spieth) took home $1.85 million, the runner-up (Matt Kuchar) earned $1.1 million, and the third-place golfer (Li Haotong) got $684,000. The wage gap among the players had nothing to do with the absolute difference in the quality of play. The compensation was determined solely by the relative standing.

Similarly, the financial rewards in the competitive world of ice skating are determined by the color of the medal won in the Olympics. A popular winner of an Olympic gold medal can earn millions by endorsing products, charging fees for personal appearances, and participating in touring ice shows. The winner of the bronze medal will have a much smaller paycheck. The actual difference in productivity between the gold and bronze medal winners is hard to discern. In fact, the judges often disagree over the ranking. Nevertheless, to the winner go the spoils.

Competitive sports are not the only setting where rewards are allocated according to relative performance. The senior vice presidents of large corporations compete fiercely for promotion to the position of president or chief executive officer (CEO). This competition can be viewed as a tournament. A survey of 200 large American firms indicated that the promotion from vice president to CEO involved a pay increase of 142 percent.[6] It is hard to believe that a worker's value of marginal product increases that much overnight. The salary structure of vice presidents and CEOs is probably best understood as a compensation package where salaries are determined by the rank in a tournament, rather than by absolute performance.

Why do some firms rely on tournament-type contracts, as opposed to using piece-rate or time-rate systems? It may be easier for the firm to observe a worker's rank in the "pecking order" than to measure the worker's actual contribution to the firm. A game will decide which football team is better (at least on that particular day). It is difficult, however, to determine precisely how much better the winning team is. Similarly, a tournament among vice-presidents will determine which of them should be promoted, but the actual contribution of each vice president to the firm's output may be much more difficult to assess.

How Much Effort Do Tournaments Elicit?

Why do some firms choose tournaments to determine promotions and salaries, but other firms pay workers according to their actual value of marginal product? Why do the winners of these tournaments earn many times the salary of the losers, even though the difference in marginal product between winners and losers is often negligible? As we will see, tournaments exist because they elicit the "right" amount of effort when it is difficult to measure a worker's actual productivity, but it is easier to contrast the productivity of one worker with that of another.

To illustrate, suppose two workers, Andrea and Bea, are competing for one of two prizes. The firm announces that the first-prize winner will receive a substantial financial reward

[5] Edward P. Lazear and Sherwin Rosen, "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy* 89 (October 1981): 841–864; Sherwin Rosen, "Prizes and Incentives in Elimination Tournaments," *American Economic Review* 76 (September 1986): 701–715.

[6] Brian G. M. Main, Charles A. O'Reilly III, and James Wade, "Top Executive Pay: Tournament or Team Work?" *Journal of Labor Economics* 4 (October 1993): 606–628.

of Z_1 dollars, while the second-prize winner gets only Z_2 dollars. Workers in this tournament know that they are more likely to win if they devote a lot of effort to the task.

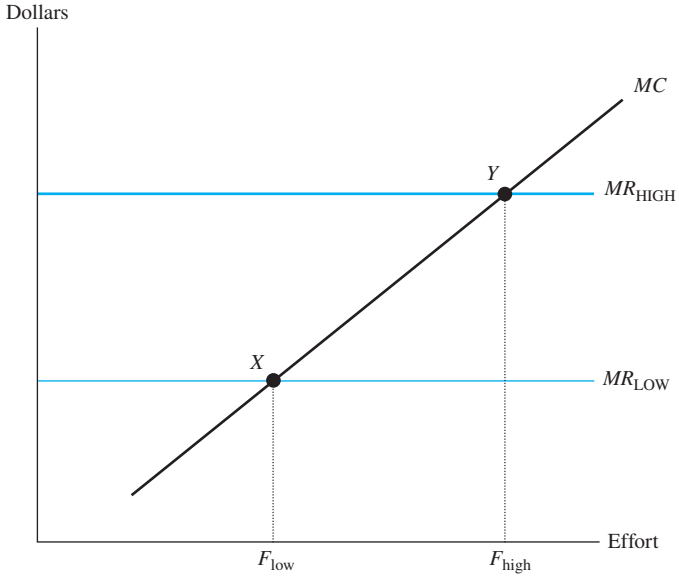
Figure 11-3 illustrates how Andrea decides the amount of effort she will devote to winning the tournament. The decision depends on a comparison of the marginal cost of effort to the marginal revenue.

The marginal cost of effort curve is upward sloping (as illustrated by the curve *MC*), as each additional unit of effort causes more "pain" than earlier units. The marginal revenue of a unit of effort depends on the difference in rewards between the first and second prize, or the prize spread $Z_1 - Z_2$. When the prize spread is narrow, the marginal revenue from an additional unit of effort is low (as in MR_{LOW}). A worker chooses the level of effort at point *X,* where the marginal cost of the effort equals the marginal revenue, and devotes F_{low} units of effort to winning the tournament. If the prize spread is large, the marginal revenue of effort is substantial (as in MR_{HIGH}), and the worker will try very hard to win by devoting F_{high} units of effort to the job.

Suppose that Andrea and Bea are equally able and that both workers "suffer" equally from devoting effort to the job (so that both players have the same marginal cost curve). Andrea and Bea will then behave in exactly the same way and allocate the same amount of effort in the contest. As a result, they have an equal chance of winning the tournament. The winner will be determined by random events at the time the game is played and may depend on such factors as where the game is played (Are the fans rooting wildly for the home-team player?) or on personalities (Do key members of the board of directors particularly like Andrea or Bea?).

FIGURE 11-3 The Allocation of Effort in a Tournament

The marginal cost curve gives the "pain" of allocating an additional unit of effort to a tournament. If the prize spread between first and second place is large, the marginal revenue to an additional unit of effort is very high (MR_{HIGH}) and the worker allocates a lot of effort to the tournament.



Suppose that Andrea and Bea are playing a tennis tournament. The winner takes home $10,000,000 and the loser takes home nothing. Each will play very hard to make sure that she is the winner at the end of the game. Because both are equally adept at tennis, the outcome of the game is unpredictable—perhaps a small wind gust slightly changes the direction and speed of the ball during a crucial play. But both Andrea and Bea know that if they do not give it their all, the other player will win. So they both work very hard at winning, even though that allocation of effort only helps them keep up with the other player.

It is obvious that the prize spread is a key determinant of the amount of effort that tournaments elicit: A very large prize spread elicits a lot of effort and keeps the game interesting. This explains why there is such a large disparity in prizes between winners and losers in sports tournaments. We all like to watch a good game. If both sides do not give it their all, spectators will leave the stadium or we will turn off the television. If both sides play at their peak ability, however, the game will be close throughout, with the final outcome being determined by random events in the last few minutes or even seconds of play. A large prize spread motivates both sides to play to their limit until the very end.

Disadvantages of Tournaments

Suppose again that two tennis players are competing for a particularly large prize. The winner will earn $10 million for her efforts; the loser gets nothing. These players have participated in many prior tournaments and have learned that they are roughly of equal ability. No matter how hard they play, the winner is typically determined by purely random events.

Both players quickly realize that they can get together prior to the tournament and agree to split the prize. They would then go through the motions of a game during the actual tournament and each would take home $5 million. Because workers can collude, tournaments may not elicit the right level of work effort.[7]

A related example of this type of corruption occurred in France in the early 1990s.[8] The local soccer team in Marseilles, the Olympique Marseilles, allegedly paid $42,000 to players of a competing team, the Valenciennes. In return, the Valenciennes would throw the game so that Marseilles could save its strength for an even bigger match that was scheduled within a week. The Marseilles team indeed won the match against the Valenciennes and then went on to capture the European Club Championship.

Tournaments also can encourage "too much" competition. The larger the prize spread, the higher the incentives for one player to take actions that *reduce* the chance other players win. A frequently heard rumor is that premed students often contaminate or destroy the experiments of other premed students in their chemistry and biology classes. Because the number of entry slots to medical schools is tightly rationed by the American Medical Association, the financial rewards to a medical degree can be considerable. The "winner" of a medical school slot is assured financial comfort and professional prestige. A large prize spread can then be a double-edged sword. It not only elicits a lot of work effort from the participants but also encourages participants to sabotage the work of others.

[7] The colluding equilibrium, however, is not very stable. After the players decide to split the prize and not to play "very hard," each of them realizes that she can win the game by putting in just a tiny bit of effort, and keep the entire $10 million.

[8] Roger Cohen, "A Soccer Scandal Engulfs All France," *New York Times,* September 6, 1993, p. 4

11-3 Policy Application: The Compensation of Executives

There is a lot of interest in the salaries of high-level executives, such as chief executive officers, or CEOs.[9] Table 11-1 lists the highest-paid CEOs in the United States. Their salaries sometimes reach dizzying heights. A few of the CEOs on the list earned in excess of $50 million annually.

Our interest in CEO salaries is only partly due to our fascination with persons who earn what most of us would consider to be extravagant salaries. The analysis of CEO compensation also raises a number of important questions in economics. Most important, *what should be the compensation package of a person who runs the firm, yet does not own it?*

The CEO is an "agent" for the owners of the firm (the owners are also called the "principals"). The owners of the firm, who are typically the shareholders, want the CEO to conduct the firm's business in a way that increases shareholder wealth. The CEO instead might want to decorate her office with expensive Impressionist originals. The purchase of these paintings reduces shareholder wealth but increases the CEO's utility. The inevitable conflict between the interests of the principals and the interests of the agent is known as the **principal–agent problem**.

The structure of executive compensation is best interpreted as a tournament in which the vice presidents compete for promotion, and the winner gets to run the company. There is a very large prize spread; executives promoted to CEO get an average 142 percent wage increase. Interestingly, the promotion from one level of vice-president to the next higher level of vice-president involves a much lower pay increase, about 40 percent.[10] In short, the prize spread gets larger the further one goes up the pyramid.

This compensation structure is implied by the theory of tournaments. Suppose there are three levels of management: The CEO, senior vice presidents, and junior vice presidents. Junior vice presidents compete among themselves for promotion to one of the senior vice

TABLE 11-1 The Highest Paid CEOs in the United States, 2012

Source: Jon Huang and Karl Russell, "The Highest-Paid C.E.O.s in 2016," *New York Times,* May 26, 2017.

Rank	Name	Company	Total Compensation (millions)
1.	Thomas M. Rutledge	Charter Communications	98.0
2.	Leslie Moonves	CBS	68.6
3.	David O'Connor	Madison Square Garden	54.0
4.	Fabrizio Freda	Estee Lauder	47.7
5.	Mark G. Parker	Nike	47.6
6.	Mark V. Hurd	Oracle	41.1
7.	Robert A. Iger	Walt Disney	41.0
8.	Safra A. Catz	Oracle	40.9
9.	David M. Zaslav	Discovery Communications	37.2
10.	Robert A. Kotick	Activision Blizzard	33.1

[9] A good survey of the literature is given by Kevin J. Murphy, "Executive Compensation," in Orley C. Ashenfelter and David Card, editors, *Handbook of Labor Economics,* vol. 3B, Amsterdam: Elsevier, 1999, pp. 2485–2563.

[10] Main, O'Reilly, and Wade, "Top Executive Pay: Tournament or Team Work?".

Theory at Work

HOW MUCH IS A SOUL WORTH?

Most of us expect that persons who work harder in private-sector jobs and bring in more business are compensated more handsomely. We all know from experience that's what makes the world go round. Remarkably, there is evidence that hard work and effort—and bringing in business—has monetary rewards in situations where one would think such considerations would be too crass to consider.

The United Methodist Church has roughly 8 million members in the United States, including such luminaries as George W. Bush and Hillary Clinton, and is known for its mainstream Christian beliefs.

A recent study was able to examine a 43-year time series (from 1961 to 2003) of all financial and hiring data for every local parish in the United Methodist Church's Oklahoma Annual Conference. This conference, led by a bishop and officials, controls the hiring and assignment of individual ministers for the parishes within its jurisdiction. A minister usually serves a local congregation for a few years and then rotates on a mandatory basis across parishes within the conference.

Local parishes and potential ministers cannot screen or select each other, because this sorting is done at the conference level. But officials at the local parish, through the Pastor Parish Relations Committee, meet annually with the minister and set pay for the next year. Median minister compensation in the parishes of the Oklahoma Annual Conference was around $37,000 (in 2008 dollars).

Among a pastor's many responsibilities, of course, is attracting new members to the parish. For example, a pastor may devote some effort to identifying nonbelieving members of the community who may be receptive to Methodist beliefs and traditions, or perhaps even compete for membership with other Christian denominations by stressing the benefits accruing from membership in the Methodist church.

The Oklahoma data reveals a systematic relationship between a minister's salary and the size of the membership of the local congregation. When a new member joins the congregation, the minister's annual salary increases by $15, while if a member leaves a congregation the salary falls by $7. The implied elasticity between a minister's salary and membership is about 0.2, about half the elasticity of CEO compensation with respect to firm size in the private sector.

Source: Jay C. Hartzell, Christopher A. Parsons, and David L. Yermack, "Is a Higher Calling Enough? Incentive Compensation in the Church," *Journal of Labor Economics* 28 (July 2010): 509–539.

president slots, who in turn compete among themselves for promotion to CEO. Executives who won the first-level tournament and were promoted to high-paying jobs as senior vice presidents may find that the compensation in their current position "meets all their needs," and therefore, may not want to compete for promotion to CEO. In order to elicit work effort from the senior vice presidents, the prize associated with becoming a CEO must be even larger than the prize associated with becoming a senior vice president.

Even after a person wins the tournament and gets promoted to CEO, the compensation package needs to be structured in a way that continues to elicit effort from the lucky executive who won the tournament. The CEO's compensation, therefore, will need to be tied to the firm's economic performance. The CEO would then hesitate to take actions that reduce shareholder wealth—because those actions would also reduce her wealth.

There is indeed a positive correlation between firm performance and CEO compensation, although the elasticity of CEO pay with respect to the rate of return to shareholders is small. In particular, a 10-percentage-point increase in the shareholder's rate of return

increases CEO pay by only 1 percent. In more straightforward terms, the CEO's salary increases by about 2 cents for every $1,000 increase in shareholder wealth.[11]

This elasticity is probably too small to impose real constraints on the CEO's behavior. Consider a CEO who wants to decorate her office with an Impressionist painting valued at $50 million. The purchase of this luxury good has no impact on the firm's bottom line and serves simply to further inflate the CEO's ego. The weak correlation between firm performance and CEO salaries implies that a $50 million reduction in shareholder wealth reduces the CEO's salary by only $1,000 a year. In effect, the CEO is giving up the equivalent of a few minutes' pay when decorating her office with an Impressionist painting. A study of the compensation of 16,000 managers at 250 large American corporations suggests that increasing the sensitivity of executive pay to firm performance would improve the profitability of the firm.[12]

11-4 Policy Application: Incentive Pay for Teachers

To improve educational outcomes in the United States, there is increased interest in trying out compensation systems that grant financial rewards to successful teachers. These rewards are often tied to the teacher's "output," as measured by student academic achievement. Similar programs have been implemented in many other countries, including Australia, India, Mexico, and Portugal. The international evidence suggests that financial incentives targeted to teachers can lead to the desired outcome of improved student outcomes.[13]

The ongoing research attempts to determine if incentive pay can work in the U.S. context, a context that is heavily influenced by the demographics of American schools and by teacher unions that strongly oppose the concept of "merit pay." The available evidence is mixed. It is far from clear that the additional financial incentives have the desired outcome of increasing "knowledge" in the student population.

One influential study of New York City public schools documented the impact of an experiment conducted between 2007 and 2010.[14] In this experiment, around

[11] Michael C. Jensen and Kevin J. Murphy, "Performance Pay and Top-Management Incentives," *Journal of Political Economy* 98 (April 1990): 225–264. The finding of a small positive correlation between CEO compensation and firm performance may be sensitive to how one defines compensation. The increasing use (and dollar value) of stock options as part of the typical CEO's employment package seems to have considerably increased the size of the correlation; see Brian J. Hall and Jeffrey B. Liebman, "Are CEOs Really Paid Like Bureaucrats?" *Quarterly Journal of Economics* 113 (August 1998): 653–692.

[12] John M. Abowd, "Does Performance-Based Management Compensation Affect Corporate Performance," *Industrial and Labor Relations Review* 43 (February 1990, Special Issue): 52S–73S; Ulrike Malmendier and Geoffrey Tate, "Superstar CEOs," *Quarterly Journal of Economics* 124 (November 2009): 1593–1638.

[13] See Victor Lavy, "Performance Pay and Teachers' Effort, Productivity, and Grading Ethics," *American Economic Review* 99 (December 2009): 1979–2011; Paul Glewwe, Nauman Ilias, and Michael Kremer, "Teacher Incentives," *American Economic Journal: Applied Economics* 2 (July 2010): 205–227. A survey is given by Derek Neal, "The Design of Performance Pay for Education," in *Handbook of Economics of Education,* vol. 4, edited by Eric A. Hanushek, Stephen Machin and Ludger Woessmann, Amsterdam: Elsevier, 2011.

[14] Roland G. Fryer, "Teacher Incentives and Student Achievement: Evidence from New York City Schools," *Journal of Labor Economics* 31 (April 2013): 373–407.

200 "high-need" schools—where students were likely to be poor, or there were many English Language Learners or special education students—participated in a program that granted financial awards if the school met preset achievement goals, and around 200 other comparable schools were selected to form a control group.

Each of the "treated" schools would receive $1,500 *per teacher* if the school reached 75 percent of its achievement target and would receive $3,000 *per teacher* if the school met or surpassed its target. The target was a composite index that depended on progress in state assessment tests, student attendance, and graduation rates. If the school qualified for the financial award, a committee at the school would then decide how to divide the lump sum given by the program among the school's teachers. The program ended up distributing $75 million to over 20,000 teachers.

But the program did not change teacher behavior, at least in terms of such observed measures as absenteeism and retention. And, as Table 11-2 shows, it also did not change student achievement. Test scores were about the same or perhaps even *lower* in the high-need schools that were selected to receive the financial incentives.

It seems that teachers in the New York City system, unlike their counterparts in other countries, failed to respond to incentives in the expected fashion. Although the reason for this discrepancy is unknown, it may be that the "prize" was much too small to make a difference, making up only about 4 percent of the average teacher's salary. It could also be that the program was too complex, with the financial incentives being filtered down to the teachers through a committee layer.

Interestingly, there is evidence that teachers in another large American city *did* respond to the economic pressures imposed by the increasing use of test scores to reward or punish teachers.[15] Specifically, the increased scrutiny that school districts place on the academic achievement of a particular teacher's students has given rise to teacher cheating, wherein dishonest teachers "revise" a student's answer sheet or obtain an early copy of the test and teach the test questions prior to the administration of the exam.

Elementary students in Chicago take a standardized, multiple-choice achievement test known as the Iowa Test of Basic Skills. The test consists of both reading and math sections, and all Chicago students between the third and eighth grade are required to take the test each year.

To detect the possibility of teacher cheating, a well-known study examined the answer sheets from all tests given in the 1993–2000 period and discovered a higher-than-expected

TABLE 11-2 Impact of Financial Incentives on Student Test Scores

Source: Roland G. Fryer, "Teacher Incentives and Student Achievement: Evidence from New York City Schools," *Journal of Labor Economics* 31 (April 2013), Tables 4, 5.

	English Score	Math Score
Elementary school	–0.013	–0.020
Middle school	–0.031	–0.051
High school	+0.009	–0.019

[15] Brian A. Jacob and Steven D. Levitt, "Rotten Apples: An Investigation of the Prevalence and Prediction of Teacher Cheating," *Quarterly Journal of Economics* 118 (August 2003): 843–877.

number of "runs" of specific answers within a classroom. In other words, the same combination of right/wrong answers would show up in exactly the same order.

The increased incidence of teacher cheating seems to be a response to specific changes in teacher incentives introduced in the Chicago school system. Specifically, the school would be put "on probation" if fewer than 15 percent of its students did not perform at the national norm in the reading portion of the exam. Probation would potentially expose the teachers to a school closing, dismissal, or reassignment. The evidence indicates that cheating was much more common in classrooms where the students' previous performance suggested there would be a poor performance in the current round. In other words, the cheating rate increased most in those poor-performing classrooms that were in schools most at risk of being put in probation.

11-5 Work Incentives and Delayed Compensation

Worker shirking, the allocation of work hours to activities other than work, is very costly in many industries. As much as 80 percent of shipping losses in the freight and airport cargo-handling industries arise from employee theft; 30 percent of retail employees steal merchandise from the workplace or misuse discount privileges; 27 percent of hospital employees steal hospital supplies; and 9 percent of workers in manufacturing falsify their time cards.[16] Firms clearly want to offer compensation packages that discourage workers from misbehaving.

It turns out that upward-sloping age–earnings profiles can discourage workers from shirking.[17] Figure 11-4 illustrates the intuition behind this insight. Suppose that the worker's value of marginal product over the life cycle is constant. The age–earnings profile in a spot labor market where the worker's effort can be measured easily would then be horizontal, as illustrated by the line *VMP* in the figure.

But the worker's effort and output are hard to observe, and it is very expensive for the firm to monitor the worker constantly. At best, the firm can make only random observations and take appropriate action if and when the worker is caught shirking. The worker stealing supplies from her employer knows that the chances of getting caught and fired are remote. She will then behave in ways that reduce productivity below her potential (so that the worker's actual contribution to the firm is less than *VMP*).

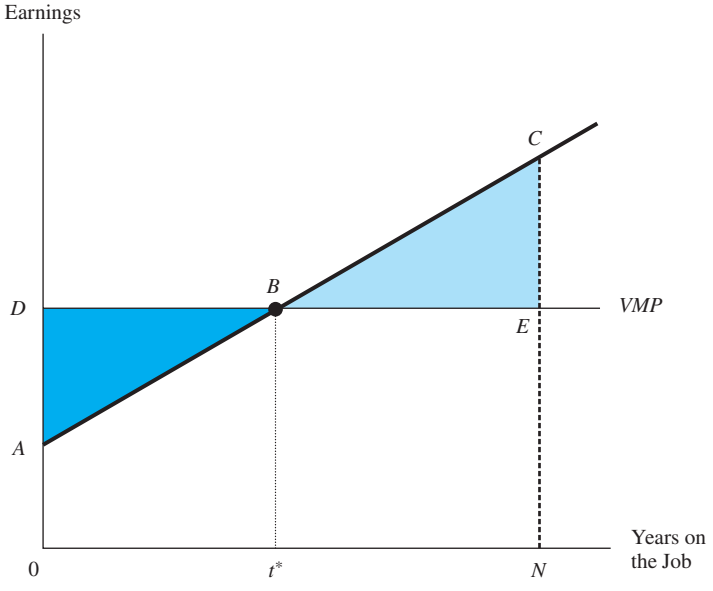
But there is a contract that will encourage the worker to *voluntarily* produce the right level of output (that is, her *VMP*) even if the firm cannot constantly monitor her performance. Suppose the firm offered to pay the worker a wage below her marginal product during the initial years on the job and a wage above her marginal product in the later years. The curve *AC* in Figure 11-4 illustrates this alternative offer.

The worker would be indifferent between the **delayed-compensation contract** given by the age–earnings profile *AC* and a contract that paid *VMP* in each time period if the

[16] William T. Dickens, Lawrence F. Katz, Kevin Lang, and Lawrence H. Summers, "Employee Crime and the Monitoring Puzzle," *Journal of Labor Economics* 7 (July 1989): 331–347.

[17] Edward P. Lazear, "Why Is There Mandatory Retirement?" *Journal of Political Economy* 87 (December 1979): 1261–1264; and Edward P. Lazear, "Agency, Earnings Profiles, Productivity, and Hours Restrictions," *American Economic Review* 71 (September 1981): 606–620.

FIGURE 11-4 Delayed Compensation and Upward-Sloping Age–Earnings Profiles
If the firm could monitor a worker easily, she would get paid her constant value of marginal product (*VMP*) over the life cycle. If it is difficult to monitor output, workers will shirk. An upward-sloping age–earnings profile (such as *AC*) discourages workers from shirking. Workers get paid less than their value of marginal product during the first few years on the job, and this "loan" is repaid in later years.



two earnings streams had the same present value. In other words, the worker would be indifferent as long as the triangle *DBA* in the figure has the same present value as the triangle *BCE.* The relatively low wage that the worker would receive initially is compensated by the high wage that the worker would eventually earn.

The two contracts, however, have very different implications for work incentives. If the worker is offered a constant wage equal to *VMP* in each period, the worker knows that the firm cannot monitor her activities constantly and has an incentive to shirk. At worst, the worker gets caught shirking, is fired, and moves on to another job paying exactly the same competitive wage.

In contrast, if the firm offers the age–earnings profile *AC,* the worker will refrain from shirking. Shirking activities now carry the risk of a substantial loss in income. If the worker is caught shirking and fired prior to year t^*, the worker has contributed more to the firm's output than she has received in compensation. In a sense, the worker made a loan to the firm, and if she gets fired, the loan is lost with no chance of its being repaid. The same logic applies if the worker is caught shirking anytime between year t^* and year *N.* Even though the worker is getting paid more than her value of marginal product, the firm still owes her money. By delaying compensation into the future, the firm elicits greater work effort and higher productivity. In a sense, the worker posts a bond with the firm during the initial years on the job, and the bond is repaid during the later years.

The delayed compensation contract in Figure 11-4 has one interesting implication for the firm's retirement policy. The firm will want the employment contract to end in year *N.*

At that time, the firm has paid off the loan, and there is no further financial gain from employing the worker at a wage exceeding her value of marginal product. The firm, therefore, will want the worker to leave the firm. The worker will not want to do so because she is getting "overpaid." This conflict might explain the origin of mandatory retirement clauses in employment contracts.[18]

The delayed compensation hypothesis faces an important conceptual obstacle. A worker would be willing to accept such offers only if she knows that she would not be fired after accumulating t^* years of seniority. As shown in Figure 11-4, this is the point at which the firm begins to repay the loan. Once the worker has put in t^* years on the job, the firm may want to renege on the contract and fire the worker. This type of firm misbehavior, however, may not occur very often. If it becomes known that the firm exploits workers by paying them less than their lifetime value of marginal product, the firm will have a hard time recruiting workers.

Even if the firm keeps its word and pays back the loan, there is always the chance that the firm will go out of business and that the worker ends up on the losing side of the deal. A delayed-compensation contract, therefore, is more likely to be offered by firms where the chances of bankruptcy are remote. As a result, delayed-compensation contracts, if they are observed at all, will tend to be observed in large and established firms.

It is worth noting that the delayed-compensation model provides another explanation for why the age–earnings profile is upward sloping *within a job.* Earnings grow over time because this type of compensation elicits work effort and reduces shirking. The model, therefore, provides an alternative story to the one told by the human capital model; namely, that the accumulation of general and specific training is responsible for the rise in earnings as workers accumulate job seniority.[19]

11-6 Efficiency Wages

Up to this point, the models linking work effort and incentive pay are based on the idea that firms find it profitable to induce workers to work harder *within the constraints imposed by a competitive market.* The optimal piece rate is the one that ensures firms earn normal profits; a too-high or too-low piece rate would encourage the exit and entry of firms, driving profits back to their normal levels. The prize structure in tournaments is set in much the same way. If firms offer prizes below the competitive "wage," additional firms enter the industry and eat away at the firms' profits.

As we will see, however, some firms might be able to increase worker productivity by paying a wage that is *above* the wage paid by other firms. A well-known example is

[18] Although employment contracts containing a mandatory retirement clause have been illegal in the United States since the mid-1980s, they are still common in other countries. See Robert M. Hutchens, "Delayed Payment Contracts and a Firm's Propensity to Hire Older Workers," *Journal of Labor Economics* 4 (October 1986): 439–457; Duane Leigh, "Why Is There Mandatory Retirement? An Empirical Reexamination," *Journal of Human Resources* 19 (Fall 1984): 512–531; and Steven G. Allen, Robert L. Clark, and Ann A. McDermed, "Pensions, Bonding, and Lifetime Jobs," *Journal of Human Resources* 28 (Summer 1993): 463–481.

[19] James Brown, "Why Do Wages Increase with Tenure?" *American Economic Review* 79 (December 1989): 971–991.

found in developing countries.[20] At the subsistence competitive wage, workers might not get the nutrition required to stay healthy. Because of the obvious link between nutrition and productivity, it may be possible for a firm to increase output by paying a wage above the subsistence wage. The firm's workers could then afford a more nutritious diet and would be better nourished, healthier, stronger, and more productive.

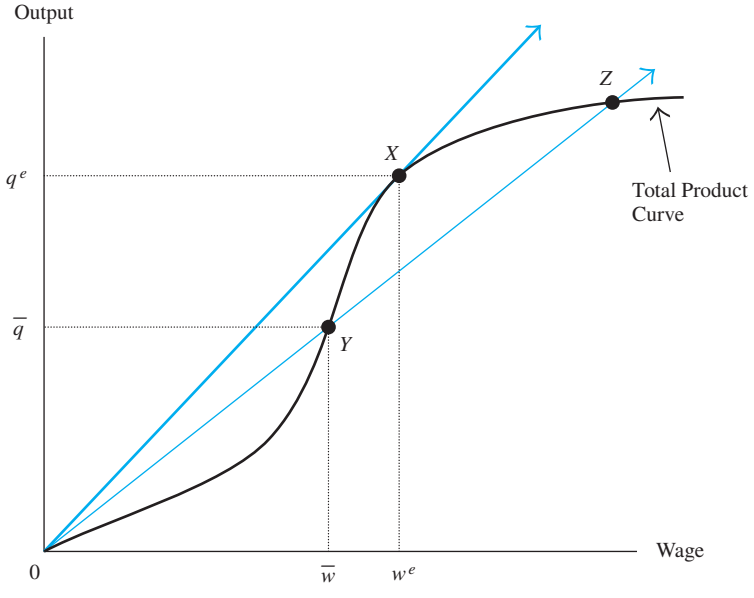
If firms pay the subsistence wage, they attract undernourished workers who are not very productive. If the firm sets its wage too high above the subsistence level, however, the firm would not make any money. The increase in labor costs could well exceed the value of the productivity gain. There exists a wage, however, known as the **efficiency wage**, where the marginal cost of increasing the wage exactly equals the marginal gain in the productivity of the firm's workers.

Setting the Efficiency Wage

It is easy to illustrate how the firm sets the profit-maximizing efficiency wage.[21] For a given level of employment, the relationship between the firm's output and the firm's wage is given by the total product curve in Figure 11-5. The upward-sloping total product curve

FIGURE 11-5 Determining the Efficiency Wage

The total product curve indicates how the firm's output depends on the wage the firm pays its workers. The efficiency wage is given by point *X,* where the marginal product of the wage (the slope of the total product curve) equals the average product of the wage (the slope of the line from the origin). The efficiency wage maximizes the firm's profits.



[20] Harvey Leibenstein, "The Theory of Underemployment in Backward Economies," *Journal of Political Economy* 65 (April 1957): 91–103.

[21] Robert Solow, "Another Possible Source of Wage Stickiness," *Journal of Macroeconomics* 1 (Winter 1979): 79–82. A survey of the efficiency wage literature is given by Andrew Weiss, *Efficiency Wages: Models of Unemployment, Layoffs, and Wage Dispersion,* Princeton, NJ: Princeton University Press, 1990.

embodies the notion that a worker's productivity and work effort depend on the wage; the firm's workforce produces more output the more they are paid. At first, output might rise very rapidly as the wage increases. Eventually, the firm encounters diminishing returns as it keeps increasing the wage, and the total product curve becomes concave. The slope of the total product curve gives the marginal product of a wage increase.

Which wage maximizes profits? Consider the straight line in Figure 11-5 that emanates from the origin and that is tangent to the total product curve at point *X*. It is easy to calculate the slope of this straight line. The slope of a line equals the change in the variable plotted on the vertical axis divided by the change in the variable plotted on the horizontal axis. Let's calculate the change that occurs as we move from the origin (where output and wages are both equal to zero) to point *X,* where the firm produces q^e units of output and pays a wage equal to w^e dollars. The slope is given by

$$\text{Slope of straight line} = \frac{\text{Change in vertical axis}}{\text{Change in horizontal axis}} = \frac{q^e - 0}{w^e - 0} = \frac{q^e}{w^e} \qquad \textbf{(11-1)}$$

The slope of the straight line emanating from the origin, therefore, equals the average product of a dollar paid to workers. For example, suppose that the firm produces 100 units of output and pays a wage of $5 at point *X*. The slope of the straight line is then equal to 20. On average, each dollar yields 20 units of output.

It turns out that the efficiency wage is the wage at which the slope of the total product curve (that is, $\Delta q/\Delta w$, or the marginal product) equals the slope of the straight line emanating from the origin, or the average product. We can write the equilibrium condition as

$$\frac{\Delta q}{\Delta w} = \frac{q}{w} \qquad \textbf{(11-2)}$$

The efficiency wage, therefore, is w^e. The intuition behind this condition is better understood if we rewrite as an elasticity, or

$$\frac{\Delta q}{\Delta w} \times \frac{w}{q} = \frac{\%\ \Delta q}{\%\ \Delta w} = 1 \qquad \textbf{(11-3)}$$

The efficiency wage, therefore, is the wage at which a 1 percent increase in the wage increases output by exactly 1 percent. To see why this is the wage that maximizes profits, suppose the firm chose to offer wage $\overline{w}$ at point *Y* instead. At that wage, the slope of the total product curve is steeper than the slope of the straight line emanating from the origin. In other words, the marginal product of an increase in the wage exceeds the average product, so that $\Delta q/\Delta w > q/w$. If we rewrite this condition as an elasticity, we get that

$$\frac{\Delta q}{\Delta w} \times \frac{w}{q} = \frac{\%\ \Delta q}{\%\ \Delta w} > 1 \qquad \textbf{(11-4)}$$

In other words, a 1 percent increase in the wage leads to an even larger increase in the firm's output. The firm would then be better off granting the wage increase. If the firm were to set the wage "too high," such as at point *Z* in Figure 11-5, the opposite restriction would hold: A 1 percent increase in the wage would increase output by less than 1 percent. In other words, the firm should refrain from granting that large wage increase.

Theory at Work

DID HENRY FORD PAY EFFICIENCY WAGES?

The Ford Motor Company was founded in 1903. In 1908, it employed 450 employees and produced 10,607 automobiles. For the most part, Ford's initial workforce was composed of skilled craftsmen. Automobile parts were often produced by outside shops and the Ford craftsmen devoted a lot of time to assembling those parts into a finished automobile.

Between 1908 and 1914, the character of the Ford Motor Company changed drastically. The first assembly built car, the Model T, was introduced and the Ford Motor Company produced little else. Model T parts were made with sufficiently high precision that they could be fitted together by workers with little skill. By 1913, Ford employed 14,000 workers and produced 250,000 cars. The workforce became three-quarters foreign born, mostly from the rural regions of southern and eastern Europe.

A contemporary description of the tasks conducted by these workers is revealing: "Division of labor has been carried on to such a point that an overwhelming majority of the jobs consist of a very few simple operations. In most cases a complete mastery of the movements does not take more than 5–10 minutes." The boredom and drudgery took its toll. Annual turnover at the Ford plant was nearly 370 percent in 1913. Put differently, Ford had to hire 50,448 persons to maintain an average labor force of 13,623 workers. And the absenteeism rate was nearly 10 percent daily.

On January 5, 1914, the Ford Motor Company decided to disregard the wage and employment conditions that had been presumably set in the competitive labor market and unilaterally reduced the length of the workday from 9 to 8 hours and more than doubled the wage from $2.34 to $5.00 per day. Immediately following the announcement, over 10,000 people lined up outside the Ford plants looking for work.

The impact of this "new-and-improved" employment contract was immediate and dramatic. By 1915, the turnover rate had dropped to 16 percent, the absenteeism rate had dropped to 2.5 percent, productivity per worker had increased between 40 and 70 percent, and profits had increased by about 20 percent. It seems, therefore, that Henry Ford benefited greatly by "discovering" efficiency wages.

Source: Daniel M. G. Raff and Lawrence H. Summers, "Did Henry Ford Pay Efficiency Wages?" *Journal of Labor Economics* 5 (October 1987, Part 2): S57–S86.

The efficiency wage, therefore, is the wage at which the elasticity of output with respect to the wage is exactly equal to 1. *A profit-maximizing firm will set this wage regardless of the value of the competitive wage outside the firm.* Because the efficiency wage will have to exceed the competitive wage (otherwise the firm would attract no workers), the firm has an oversupply of labor. At the efficiency wage, therefore, the firm gets many more job applicants than the firm is willing to hire. The firm, however, will not want to reduce the wage. After all, the efficiency wage w^e is *the* profit-maximizing wage. A reduction in the wage would reduce worker productivity by more than it reduces the payroll, lowering profits. Because the efficiency wage attracts too many workers, some workers will be involuntarily unemployed. We will discuss this implication of the model in the chapter on unemployment.

Why Is There a Link between Wages and Productivity?

The link between wages and productivity captured by the total product curve in Figure 11-5 might arise for a number of distinct reasons.[22]

[22] Carl Shapiro and Joseph E. Stiglitz, "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review* 74 (June 1984): 433–444; George A. Akerlof, "Labor Contracts as a Partial Gift Exchange," *Quarterly Journal of Economics* 97 (November 1982): 543–569; and Gary Charness and Peter Kuhn, "Does Pay Inequality Affect Worker Effort? Experimental Evidence," *Journal of Labor Economics* 25 (October 2007): 693–723.

First, a high wage makes it costly for workers to shirk. If a shirking worker is caught and fired, she loses her high-paying job and may become unemployed. The fear of unemployment helps keep the worker in line, working hard on the job.

Second, higher wages might influence the "sociology" of the workplace. Workers who are well paid might work harder even if there is no threat of dismissal. Those workers might view the high wage as a "gift" from the employer and feel obligated to repay the gift by working harder.

Third, high-wage workers are less likely to quit. The lower turnover rates in firms paying efficiency wages would reduce turnover costs and minimize the disruption that occurs when trained workers leave and new workers have to be hired and trained. Efficiency wages, therefore, reduce the quit rate and increase output and profits.

Finally, firms paying efficiency wages might attract a selected pool of workers. Consider a firm offering a low wage. Only workers who have a reservation wage below this wage will accept job offers from this firm. High-ability workers will tend to have higher reservation wages. A firm that pays the efficiency wage attracts a more qualified pool of workers, increasing the productivity and profits of the firm.

Evidence on Efficiency Wages

The efficiency wage hypothesis has been used to explain the huge interindustry wage differentials that exist among comparable workers.[23] Table 11-3 reports the log wage differential (approximately the percent wage differential) between the typical person in an industry and the typical worker in the labor market who has the same socioeconomic background (such as age, sex, race, and education). Workers employed in metal mining or railroads earn around 30 percent more than the average worker in the labor market, whereas workers employed in hardware stores or childcare services earn around 25 percent less. These interindustry wage differentials are persistent over time, so that industries that paid high wages 20 years ago also pay high wages today.

The competitive model argues that these interindustry wage differentials must reflect either differences in job characteristics or differences in unobserved worker traits. For example, it might be that jobs in some industries are more pleasant or safer. The "worse" jobs would then have to pay higher wages to attract workers who dislike high levels of pollution or risk. Workers also might sort themselves across industries on the basis of their abilities. If firms in the motor vehicle industry really do pay about 50 percent more than hardware retail stores, employers in the auto industry can sift through the job applicants. Workers in high-wage industries, therefore, would be more able and more productive.

In contrast to these competitive explanations, the efficiency wage model stresses that the interindustry wage differentials are "real." In other words, the differentials do not reflect the compensation paid to workers who are working in unpleasant or risky jobs or who are

[23] Alan B. Krueger and Lawrence H. Summers, "Efficiency Wages and the Inter-Industry Wage Structure," *Econometrica* 56 (March 1988): 259–293. See also Erica L. Groshen, "Sources of Intra-Industry Wage Dispersion: How Much Do Employers Matter?" *Quarterly Journal of Economics* 106 (August 1991): 869–884; Steven G. Allen, "Updated Notes on the Interindustry Wage Structure, 1890–1990," *Industrial and Labor Relations Review* 48 (January 1995): 305–321; and Paul Chen and Per-Anders Edin, "Efficiency Wages and Industry Wage Differentials: A Comparison across Methods of Pay," *Review of Economics & Statistics* 84 (November 2002): 617–631.

TABLE 11-3 The Interindustry Wage Structure

Source: Alan B. Krueger and Lawrence H. Summers, "Efficiency Wages and the Inter-Industry Wage Structure," *Econometrica* 56 (March 1988): 281–287.

Industry	Log Wage Relative to Average Comparable Worker
Mining	
Metal mining	0.296
Crude petroleum, natural gas extraction	0.256
Construction	0.129
Manufacturing	
Meat products	–0.028
Dairy products	0.176
Apparel and accessories	–0.137
Tires and inner tubes	0.306
Motor vehicles	0.244
Transportation	
Railroads	0.268
Taxicab services	–0.203
Wholesale trade	
Electrical goods	0.123
Farm products	–0.109
Retail trade	
Hardware stores	–0.304
Department stores	–0.190
Finance, insurance, and real estate	
Banking	0.048
Real estate	0.004
Business and repair services	
Advertising	0.092
Automotive-repair shops	–0.058
Professional and related services	
Offices of physicians	–0.076
Childcare services	–0.275

more productive. Rather, efficiency wages exist because firms in some industries find it profitable to pay more than the competitive wage (perhaps because it is hard to monitor output or because there are high turnover costs), and firms in other industries do not.

The evidence on which of the two stories best fits the data is mixed. The interindustry wage differentials remain even if we compare jobs that are equally risky or pleasant, so the theory of compensating wage differentials cannot account for the sizable interindustry wage gaps. Moreover, if the interindustry wage differentials were solely due to differences in worker ability, we would not observe workers in low-wage industries quitting more often than workers in high-wage industries. After all, it would be very unlikely that a low-ability worker could get a job in the high-wage sector.

At the same time, however, workers *do* sort themselves across industries. If efficiency wages explain the interindustry wage differentials, workers who move from a low-wage industry to a high-wage industry should experience a sizable wage increase. If the interindustry wage differentials reflect differences in worker ability, a low-ability worker moving from a low-wage to a high-wage industry should not get much of a wage increase. One influential study, which "tracked" workers as they changed jobs across industries, concluded that perhaps as much as 70 percent of interindustry wage differentials might be due to the sorting of able workers in high-wage industries.[24]

The Bonding Critique

The efficiency wage model predicts that there are *permanent* wage differences across firms, despite the fact that low-wage (or unemployed) workers would rather hold high-wage jobs. An important criticism of this implication is known as the **bonding critique**.[25]

Firms can use many types of compensation schemes, such as tournaments, upward-sloping age–earnings profiles, and piece rates, to encourage workers not to shirk on the job. All of these mechanisms operate within the confines of a competitive market. Industries that pay too small a piece rate or award too small a first prize to the winner of a tournament encourage other entrepreneurs to enter the industry, increasing the demand for salaries of workers and forcing the industry back to a normal level of profits. If the industry pays too high a piece rate or offers too big a prize, firms lose money and the compensation of workers falls.

Efficiency wages also provide incentives for workers not to shirk. But firms determine the efficiency wage without regard to market conditions. As a result, firms that choose to pay very high wages will have too many job applicants, and there are no market forces to bring the wage into line with other firms.

Critics of the efficiency wage hypothesis argue that this cannot be the end of the story. Job applicants should be willing to take actions that would "buy" them a job at the firm. For instance, they could post a bond at the time of hiring. If firms caught the workers shirking, the firm could dismiss the worker and keep the bond. If the employment relationship worked out, the firm would return the bond (plus interest) to the worker at the time of retirement.

In fact, workers seldom put up bonds as a condition of getting hired. As we saw earlier, however, upward-sloping age–earnings profiles or other forms of delayed-compensation schemes play exactly the same role. Workers accept wages lower than their value of marginal product during the initial years on the job and would be repaid in later years. As workers compete for jobs in high-wage industries, the wage profile in high-wage industries would tilt and become steeper.

In the end, workers would be indifferent between jobs in high-wage and low-wage industries because the present value of earnings in all jobs would be equalized. The bonding critique, therefore, suggests that efficiency wage models would self-destruct in the long run.

[24] Kevin M. Murphy and Robert Topel, "Efficiency Wages Reconsidered: Theory and Evidence," in Y. Weiss and G. Fishelson, editors, *Advances in the Theory and Measurement of Unemployment,* New York: Macmillan, 1990, pp. 204–240.

[25] Lorne H. Carmichael, "Efficiency Wage Models of Unemployment—One View," *Economic Inquiry* 28 (April 1990): 269–295.

Summary

- Piece rates are used by firms when it is cheap to monitor the output of the workers.
- Piece-rate compensation systems attract the most able workers and elicit high levels of effort from those workers. Workers in these firms, however, may stress quantity over quality and may dislike the possibility that incomes fluctuate significantly over time.
- Some firms award promotions on the basis of the relative ranking of the workers. A tournament might be used when it is cheaper to observe the relative ranking of a worker than the absolute level of the worker's productivity.
- Workers allocate more effort to the firm when the prize spread between winners and losers in the tournament is very large. A large prize spread, however, also creates incentives for workers to sabotage the efforts of other players.
- There is a positive correlation between the compensation of CEOs and the performance of the firm, but the correlation is weak.
- Upward-sloping age–earnings profiles might arise because delaying the compensation of workers until later in the life cycle discourages shirking.
- Some firms might want to pay a wage above the competitive wage to motivate workers to work harder. The efficiency wage is set such that the elasticity of output with respect to the wage is equal to 1.
- Efficiency wages create a pool of workers who are involuntarily unemployed.

Key Concepts

bonding critique, *397*
delayed-compensation contract, *389*
efficiency wage, *392*
free-riding problem, *380*
incentive pay, *376*
piece rates, *376*
principal–agent problem, *385*
ratchet effect, *381*
spot labor markets, *376*
time rates, *376*
tournament, *382*

Review Questions

1. What factors determine whether a firm offers a piece-rate or a time-rate compensation system?
2. Discuss how workers who differ in their innate abilities sort themselves across piece-rate and time-rate jobs. Also describe how the two compensation systems elicit different levels of effort from the workers.
3. If piece rates elicit more effort from workers, why do firms not use this method of compensation more often?
4. Show how a large prize spread in a tournament elicits a higher level of work effort from the participants.
5. Discuss some of the problems encountered when firms allocate sizable rewards to the winner of the tournament.

6. Why is the principal–agent problem relevant to understanding how CEOs should be compensated?
7. Discuss how upward-sloping age–earnings profiles can elicit more effort from workers.
8. Why is there mandatory retirement in many countries?
9. Describe how the firm sets an efficiency wage above the competitive level. Why are there no market forces forcing the profit-maximizing firm to reduce the wage to the competitive level?
10. What factors create the link between wages and productivity that is at the heart of efficiency wage models?
11. What is the bonding critique of efficiency wage models?

Problems

11-1. Suppose there are 100 workers in an economy with two firms. All workers are worth \$35 per hour to firm A but differ in their productivity at firm B. Worker 1 has a value of marginal product of \$1 per hour at firm B; worker 2 has a value of marginal product of \$2 per hour at firm B, and so on. Firm A pays its workers a time-rate of \$35 per hour, while firm B pays its workers a piece rate. How will the workers sort themselves across firms? Suppose a decrease in demand for both firms' output reduces the value of every worker to either firm by half. How will workers now sort themselves across firms?

11-2. Taxicab companies in the United States typically own a large number of cabs and licenses; taxicab drivers then pay a daily fee to the taxicab company to lease a cab for the day. In return, the drivers keep all of their fares (so that, in essence, they receive a 100 percent commission on their sales). Why do you think this type of compensation system developed in the taxicab industry?

11-3. A firm hires two workers to assemble bicycles. The firm values each assembly at \$12. Charlie's marginal cost of allocating effort to the production process is $4N$, where N is the number of bicycles assembled per hour. Donna's marginal cost is $6N$.

(a) If the firm pays piece rates, what will be each worker's hourly wage?

(b) Suppose the firm pays a time rate of \$15 per hour and fires any worker who does not assemble at least 1.5 bicycles per hour. How many bicycles will each worker assemble in an 8-hour day?

11-4. All workers start working for a particular firm when they are 21 years old. The value of each worker's marginal product is \$18 per hour. In order to prevent shirking on the job, a delayed-compensation scheme is imposed. In particular, the wage level at every level of seniority is determined by:

$$\text{Wage} = \$10 + (0.4 \times \text{Years in the firm}).$$

Suppose also that the discount rate is zero for all workers. What will be the mandatory retirement age under the compensation scheme? (Hint: Use a spreadsheet.)

11-5. Suppose a firm's technology requires it to hire 100 workers regardless of the wage level or market demand conditions. The firm, however, has found that worker

Wage Rate	Units of Output
$ 8.00	65
$10.00	80
$11.25	90
$12.00	97
$12.50	102

productivity is greatly affected by its wage. The historical relationship between the wage level and the firm's output is given by:

What wage level should a profit-maximizing firm choose?

11-6. Consider three firms identical in all aspects except their monitoring efficiency, which cannot be changed. Even though the cost of monitoring is the same across the three firms, shirkers at Firm A are identified almost for certain; shirkers at Firm B have a slightly greater chance of not being found out; and shirkers at Firm C have the greatest chance of not being identified as a shirker. If all three firms pay efficiency wages to keep their workers from shirking, which firm will pay the greatest efficiency wage? Which firm will pay the smallest efficiency wage?

11-7. Consider three firms identical in all aspects (including the probability with which they discover a shirker), except that monitoring costs vary across the firms. Monitoring workers is very expensive at Firm A, less expensive at Firm B, and cheapest at Firm C. If all three firms pay efficiency wages to keep their workers from shirking, which firm will pay the greatest efficiency wage? Which firm will pay the smallest efficiency wage?

11-8. A firm can hire as much labor as it wants at $5 per hour. In return, each worker produces 10 units of output per hour. The firm can sell up to 2,500 units of output each day at $2 per unit, but it cannot sell any more than 2,500 units of output in a day. The firm has no other costs besides labor.

(a) How many hours of labor does the firm purchase and how much profit does it earn each day?

(b) The firm can choose to pay an efficiency wage. In particular, the firm can choose to pay $6, $7, $8, $9, or $10 per hour, and in exchange, each worker will produce 18, 23, 27, 28, or 29 units of output per hour, respectively. What hourly wage should the firm offer to maximize profits?

11-9. Consider a firm that offers the following employee benefit. When a worker turns 60 years-old she is given a one-time opportunity to quit her job, and in return the firm will pay her a bonus of 1.5 times her annual salary and pay her health insurance premiums until she is eligible for Medicare.

(a) What problem is the firm trying to solve by offering this benefit?

(b) Why is the health insurance premium portion of the benefit important in the United States?

(c) For what industries might one expect such opportunities to be presented to workers?

11-10. (a) Why would a firm ever choose to offer profit-sharing to its employees in place of paying piece rates?

(b) Describe the free riding problem in a profit-sharing compensation scheme. How might the workers of a firm "solve" the free riding problem?

11-11. (a) How does the offering of stock options to CEOs attempt to align CEO incentives with share holder incentives?

(b) Enron was a company that was ruined in part because of the stock options offered to upper management. Explain.

(c) In addition to accounting reforms, how might stock options be changed to try to prevent situations like what happened at Enron from occurring in the future?

11-12. (a) Personal injury lawyers typically do not charge a client unless they obtain a monetary award on their client's behalf. Why?

(b) What would happen to the number of lawsuits if lawyers had to charge an hourly rate win or lose and could not charge a fixed percentage of the award?

11-13. Consider the following four tasks (all of which require significant time and/or effort): (1) trekking through a forest carrying a trowel and 40 saplings, and every quarter of a mile kneeling to the ground, digging a hole, and planting a sapling; (2) using a pick axe to extract 100 pounds of ore from the ground; (3) a team of 200 shoveling snow from the 85,000 seats in a stadium before a January football game; and (4) advising a college senior in her senior thesis which, by protocol, requires weekly 90-minute meetings plus an additional 2 hours each week of reading and preparation. Describe in detail why an employer may or may not want to pay employees by the piece to accomplish these tasks? What are some conclusions for when paying by the piece is most useful?

11-14. Economists and psychologist have long wondered how worker effort relates to wages. Specifically, the question is whether worker effort responds to increased wages alone or whether effort also responds to relative wages.

(a) Design a classroom experiment that would allow you to quantify the relationship among effort, reward, and relative reward.

(b) Explain how the data you collect can be used to identify both relationships. What do you think you would find?

11-15. Some compensation schemes include a signing bonus while others include the potential to receive annual year-end bonuses.

(a) From the firm's perspective, what are the benefits of offering a signing bonus? What are the benefits of offering a year-end bonus?

(b) If a firm pays its sales staff a piece rate and a year-end bonus, why will it be the case that the rate of pay per piece is less than the market value? Why will the sales staff willingly accept such an arrangement?

(c) How does the existence of year-end bonuses support the bonding critique?

Selected Readings

Roland G. Fryer, "Teacher Incentives and Student Achievement: Evidence from New York City Schools," *Journal of Labor Economics* 31 (April 2013): 373–407.

Brian A. Jacob and Steven D. Levitt, "Rotten Apples: An Investigation of the Prevalence and Prediction of Teacher Cheating," *Quarterly Journal of Economics* 118 (August 2003): 843–877.

Brian J. Hall and Jeffrey B. Liebman, "Are CEOs Really Paid Like Bureaucrats?" *Quarterly Journal of Economics* 113 (August 1998): 653–692.

Edward P. Lazear, "Why Is There Mandatory Retirement?" *Journal of Political Economy* 87 (December 1979): 1261–1264.

Edward P. Lazear, "Performance Pay and Productivity," *American Economic Review* 90 (December 2000): 1346–1361.

Edward P. Lazear and Sherwin Rosen, "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy* 89 (October 1981): 841–864.

Jay C. Hartzell, Christopher A. Parsons, and David L. Yermack, "Is a Higher Calling Enough? Incentive Compensation in the Church," *Journal of Labor Economics* 28 (July 2010): 509–539.

Thomas Lemieux, W. Bentley MacLeod, and Daniel Parent, "Performance Pay and Wage Inequality," *Quarterly Journal of Economics* 124 (February 2009): 1–49.

Daniel M. G. Raff and Lawrence H. Summers, "Did Henry Ford Pay Efficiency Wages?" *Journal of Labor Economics* 5 (October 1987, Part 2): S57–S86.

Beck A. Taylor and Justin G. Trogdon, "Losing to Win: Tournament Incentives in the National Basketball Association," *Journal of Labor Economics* 20 (January 2002): 23–41.

Chapter 12

Unemployment

It's a recession when your neighbor loses his job; it's a depression when you lose your own.

—Harry S. Truman

Why are some workers unemployed? This question raises some of the thorniest issues in economics. A competitive equilibrium equates the supply of workers with the demand for workers. The equilibrium wage clears the market, and all persons looking for work can find jobs.

Nevertheless, unemployment can sometimes be a widespread phenomenon. In 2010, at the peak of the Great Recession, the unemployment rate in the United States reached 9.6 percent, and almost half of the unemployed had been without work for at least 27 weeks.

It is difficult to understand the existence and persistence of large numbers of unemployed workers in terms of the typical model of supply and demand unless (1) firms pay wages that are above equilibrium and there is an excess supply of labor and (2) wages are "sticky" and cannot be driven down to the equilibrium level.

Workers are unemployed for many reasons, and some types of unemployment are more worrisome. At any time, for instance, many persons are "in between" jobs. They have either just quit or been laid off, or they have just entered (or reentered) the labor market. It takes time to learn about and locate the available job opportunities. Therefore, even a well-functioning market economy, where the number of available jobs equals the number of persons looking for work, will exhibit some unemployment as workers search for jobs.

Put differently, the equilibrium level of unemployment will not be zero. This type of frictional unemployment, however, cannot explain why nearly 25 percent of the workforce was unemployed at the nadir of the Great Depression in 1933 or why the unemployment rate hit almost 10 percent in 2010. Many workers seem to be unemployed not because they are in between jobs but because of a fundamental imbalance between the supply and the demand for workers.

This chapter shows how job search activities generate unemployment in a competitive economy and identifies some of the factors that can prevent the market from clearing—even after job search activities are accounted for. Economists have created ingenious stories of how unemployment can arise in competitive markets. Each particular theory can explain certain aspects of the unemployment problem. No single theory, however, provides

a complete explanation for why unemployment sometimes afflicts a large fraction of the workforce, why unemployment targets some groups more than others, and why some workers remain unemployed for a very long time.

12-1 Unemployment in the United States

Figure 12-1 shows the historical trend in the U.S. unemployment rate since 1900. The unemployment rate has fluctuated dramatically over time; it reached a peak of about 25 percent in 1933 and lows of about 1 percent in 1906 and 1944.

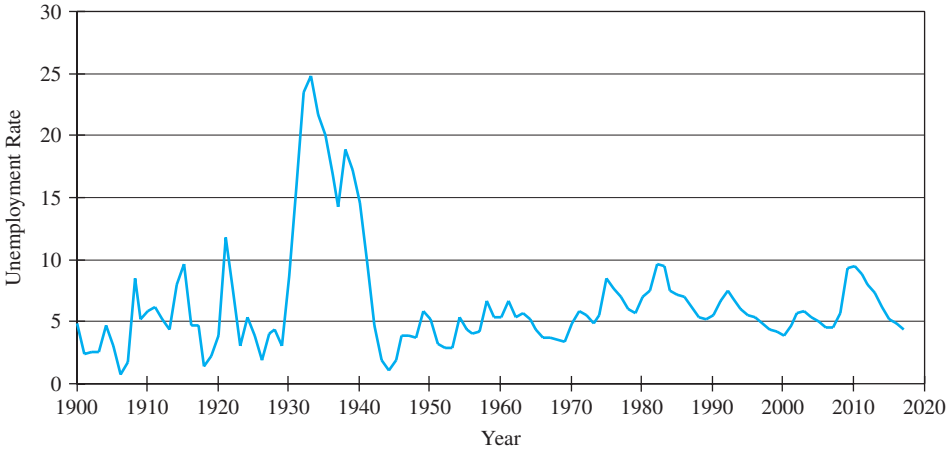
The unemployment rate gives the fraction of labor force participants looking for work. Many persons who would like to work might have withdrawn from the labor force because they could not find jobs. The count of the unemployed misses the discouraged workers. As a result, the official unemployment rate may underestimate the true severity of the unemployment problem, particularly during severe economic downturns when a large pool of discouraged workers might be in the nonmarket sector, "waiting out" the recession.

The data summarized in Figure 12-1 show a slight upward drift in the unemployment rate from the 1950s through the 1980s. In the 1950s, the average unemployment rate was 4.5 percent; during the 1960s it was 4.8 percent; during the 1970s it rose to 6.2 percent; and during the 1980s it rose further to 7.3 percent. This trend broke in the 1990s, when the unemployment rate fell to levels not seen in about 30 years. In 1998, the unemployment rate was just 4 percent.

This period of low unemployment, however, stopped abruptly in 2008 when the United States entered a deep recession after a serious financial crisis. The very rapid rise in the unemployment rate after the crisis was remarkable, from 4.6 percent in 2007 to 9.6 percent in 2010, more than doubling the unemployment rate in just 3 years.

FIGURE 12-1 Unemployment in the United States, 1900–2017

Sources: The pre-1948 unemployment rates are reported in Stanley Lebergott, "Annual Estimates of Unemployment in the United States, 1900–1950," *The Measurement and Behavior of Unemployment,* NBER Special Committee Conference Series No. 8, Princeton, NJ: Princeton University Press, 1957, pp. 213–239. The post-1948 rates are from U.S. Bureau of Labor Statistics, "Historical Data for the 'A' Tables of the Employment Situation Release. Table A-1, Employment Status of the Civilian Population by Sex and Age"; available at stats.bls.gov/cps/cpsatabs.htm. The unemployment rate refers to persons aged 16 and over.



It is important to note that this sharp jump in the unemployment rate was totally unexpected. Ironically, a popular topic in macroeconomic research just prior to the financial crisis of 2008 was the attempt to understand how the United States had been able to "moderate" the volatility of business cycle activity, leading to a period that became known as the "Great Moderation." In a 2004 lecture, for example, Ben Bernanke (who would later become chairman of the Federal Reserve) noted that "one of the most striking features of the economic landscape over the past twenty years or so has been a substantial decline in macroeconomic volatility." It is doubly ironic that the research interest in the Great Moderation eventually morphed into research interest in the Great Recession.

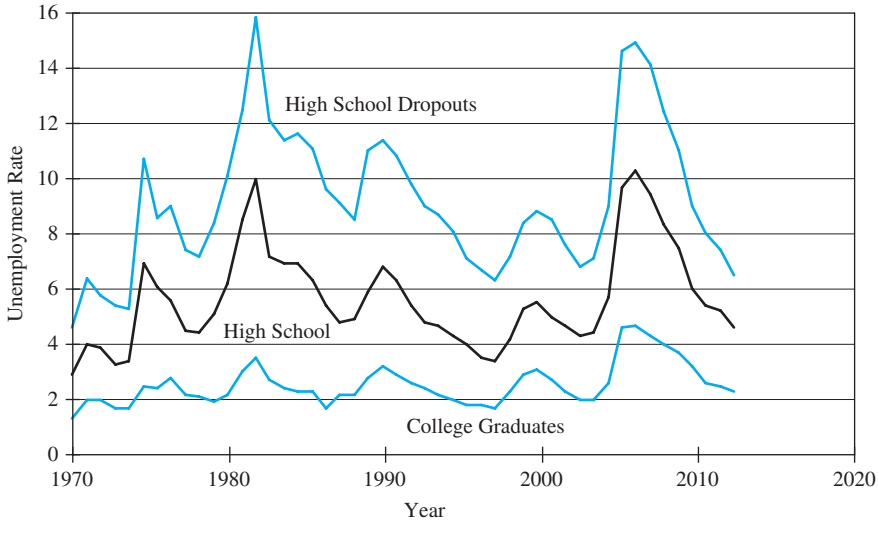
Who Are the Unemployed?

The fact that the unemployment rate in 2017 was 4.4 percent does not imply that each labor market participant had a 4.4 percent probability of being unemployed at a point in time during that calendar year. Unemployment is not an equal-opportunity employer. It is concentrated among particular demographic groups and among workers in specific sectors of the economy.

Figure 12-2 illustrates one key feature of unemployment in the United States: The unemployment rate is much higher for less-educated workers. In 2017, the unemployment rate of college graduates was 2.3 percent, as compared to 4.6 percent for high school graduates and 6.5 percent for high school dropouts. The figure also shows that the "unemployment gap" between high-educated and low-educated workers widens substantially during recessions. In 2010, the unemployment rate of high school dropouts exceeded that of college graduates by more than 10 percentage points. By 2017, that gap had narrowed to 4.2 percentage points.

FIGURE 12-2 **Unemployment Rates by Education, 1970–2017**

Sources: U.S. Bureau of Labor Statistics, *Labor Force Statistics Derived from the Current Population Survey, 1948–87,* Bulletin 2307, Washington, DC: Government Printing Office, 1988, pp. 848–849; U.S. Bureau of the Census, *Statistical Abstract of the United States,* Washington, DC: Government Printing Office, various issues. The post-1992 data are from U.S. Bureau of Labor Statistics, "Historical Data for the 'A' Tables of the Employment Situation Release. Table A-4, Employment Status of the Civilian Population 25 Years and Over by Educational Attainment"; available at stats.bls.gov/cps/cpsatabs.htm. The unemployment rates refer to the population of persons aged 25 and over.



Education reduces unemployment rates for two distinct reasons. First, educated workers invest more in on-the-job training. Because specific training "marries" firms and workers, firms are less likely to lay off educated workers when they face adverse economic conditions. In addition, educated workers often switch jobs without suffering an intervening unemployment spell. It seems as if educated workers are better informed or have better networks for learning about alternative job opportunities.

Table 12-1 reports unemployment rates by age, race, gender, and industry of employment. Younger workers are more likely to be unemployed. The unemployment rate of teenagers in 2017 was 14.0 percent as compared to about 3 percent for workers aged 45–64. Part of the higher unemployment rate of teenagers may be due to the adverse employment effects of the minimum wage.

The data also indicate that whites have lower unemployment rates than either blacks or Hispanics, but Asians have an even lower unemployment rate. In 2017, the unemployment rate of blacks was almost twice as high as that of whites (7.5 percent versus 3.8 percent). The persistently large black–white unemployment differential cannot be attributed to the different skill composition of the black and white populations. The racial gap in unemployment rates remains even if we compare black and white workers who have the same observable skills and who live in the same area.[1]

Historically, women had higher unemployment rates than men. In 1983, for example, 9.8 percent of men and 15.3 percent of women were unemployed. It was typically argued that women had a higher unemployment rate because they were much more likely to be "on the move" either in between jobs or in and out of the labor market. These transitions require women to look for work and increase their unemployment rate. By 2017, the gender gap in unemployment had disappeared; both groups had essentially the same unemployment rate of about 4.4 percent.

TABLE 12-1
Unemployment Rates in 2017, by Demographic Group and Industry

Age:		Industry:	
16–19	14.0	Agriculture	7.2
20–24	7.4	Mining	4.1
25–34	4.6	Construction	6.0
35–44	3.5	Manufacturing	3.6
45–54	3.2	Information	4.5
55–64	3.1	Transportation and utilities	4.1
		Retail trade	4.6
Race:		Finance, insurance, and real estate	2.4
White	3.8	Leisure and hospitality	6.1
Black	7.5	Professional and business services	4.5
Hispanic	5.1	Government	2.5
Asian	3.4		
Gender:		All workers	4.4
Male	4.4		
Female	4.3		

Source: U.S. Department of Labor, Bureau of Labor Statistics, *Labor Force Statistics from the Current Population Survey.* Available at: www.bls.gov/cps/tables.htm#charunem.

[1] Joseph A. Ritter and Lowell J. Taylor, "Racial Disparity in Unemployment," *Review of Economics and Statistics* 93 (February 2011): 30–42.

Note also that there are sizable differences in unemployment across industries, with workers in agriculture, construction, and "leisure and hospitality" being most likely to be unemployed. The unemployment rate in these industries exceeded 6 percent in 2017. In contrast, government workers or workers in "financial activities" have an unemployment rate below 3 percent.

There are four ways in which a worker can become unemployed: Some workers lose their jobs due to layoffs or plant closings (or job losers); some workers leave their jobs (job leavers); some job seekers reenter the labor market after spending some time in the nonmarket sector (reentrants); and some job seekers are new to the job market, such as recent high school or college graduates (new entrants). As Figure 12-3 shows, the fraction of workers who are job losers hovered around 50 percent (with up-and-down blips) between 1980 and 2005. Because of the severity of the Great Recession, this statistic peaked at 64 percent in 2009, before going back down to the 50 percent mark by 2017.

Figure 12-4 documents that a large fraction of the unemployed are likely to be in long-term unemployment spells. Even prior to the Great Recession, there had been an upward drift in the fraction of the unemployed who had been without work for more than 26 weeks. In the early 1950s, for instance, only about 5–10 percent of unemployed workers were in spells lasting more than 26 weeks. By 2007, about 18 percent of the unemployed workers were in these long spells. The Great Recession led to a dramatic explosion in this number. In 2010, 43.3 percent of the unemployed were in long-term spells, and this statistic remained at historically high levels after the recovery began. Almost a quarter of the unemployed were in long spells in 2017, a higher proportion than at any time between the end of World War II and the start of the Great Recession.

FIGURE 12-3 Unemployed Persons by Reason, 1967–2017 (as a Percent of Total Unemployment)

Source: U.S. Bureau of Labor Statistics, "Historical Data for the 'A' Tables of the Employment Situation Release. Table A-11, Unemployed Persons by Reason of Unemployment"; available at stats.bls.gov/cps/cpsatabs.htm. The population of unemployed includes all unemployed persons aged 16 or over.

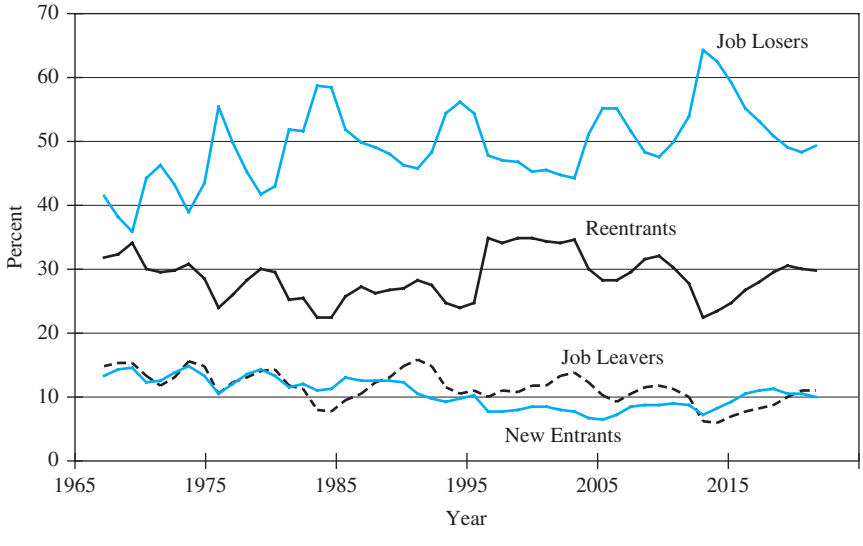
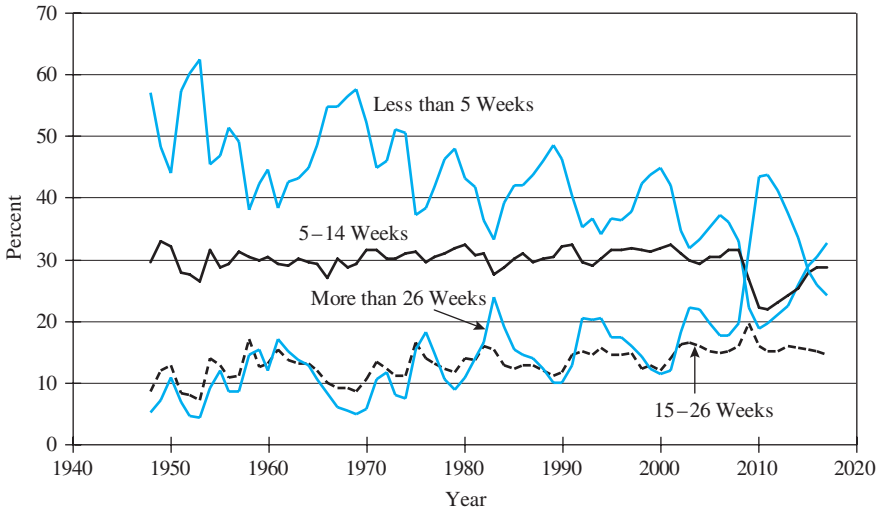


FIGURE 12-4 Unemployed Persons by Duration, 1948–2017 (as a Percent of Total Unemployment)

Source: U.S. Bureau of Labor Statistics, "Historical Data for the 'A' Tables of the Employment Situation Release. Table A-12, Unemployed Persons by Duration of Unemployment"; available at stats.bls.gov/cps/cpsatabs.htm. The population of unemployed includes all unemployed persons aged 16 or over.



Put differently, the notion that unemployment can be typically characterized by short-term spells seems to be becoming less relevant. Even prior to the Great Recession, there had been a noticeable downward drift in the fraction of unemployed persons who had been unemployed fewer than 5 weeks.

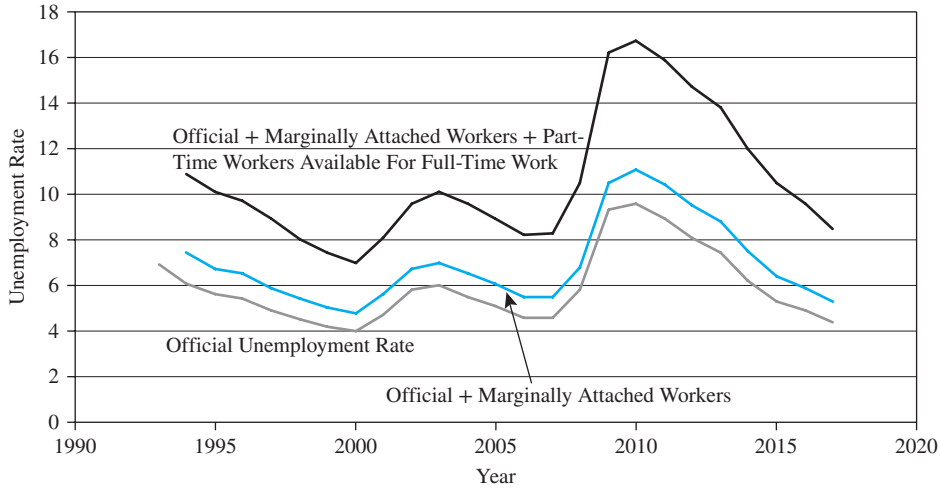
Finally, the unemployment rate gives the proportion of the labor force that is unemployed and looking for work. There also may be some discouraged workers—workers who gave up on their job search because they could not find any employment. The Bureau of Labor Statistics publishes an alternative unemployment rate that includes in the pool of unemployed any "marginally attached persons...who currently are neither working nor looking for work but indicate that they want and are available for a job and have looked for work sometime in the recent past." Figure 12-5 shows that the unemployment rate goes up by about 1 percentage point when the marginally attached are counted as unemployed.

A more sizable group is made up by the "underemployed"—persons who "want and are available for full-time work but have had to settle for a part-time schedule." As the figure shows, including this group as part of the unemployed leads to an even more dramatic increase in the unemployment rate. In 2007, the broadest measure of the unemployment rate exceeded the official measure by 3.7 percentage points. By 2010, however, the broader definition implied a 16.2 percent unemployment rate as compared to the 9.3 percent official rate, a gap of almost 7 percentage points.

The official unemployment rate has probably been a misleading indicator of economic activity since the onset of the Great Recession because the labor force participation rate declined dramatically between 2008 and 2013. The labor force participation rate stood at 66.2 percent at the beginning of 2008 and fell to 62.9 percent by the end of 2013 (and was still at 62.7 percent in January 2018 despite the strong economic recovery).

FIGURE 12-5 Trends in Alternative Measures of the Unemployment Rate, 1994–2017

Source: U.S. Bureau of Labor Statistics, "Historical Data for the 'A' Tables of the Employment Situation Release. Table A-15, Alternative Measures of Labor Underutilization"; available at stats.bls.gov/cps/cpsatabs.htm. The unemployment rate refers to the population of persons aged 16 and over. The unemployment rate that includes the marginally attached is called the U-5 series; the unemployment rate that also includes workers who settled for a part-time schedule is called the U-6 series. The official unemployment rate is called the U-3 series.



The 3.3 percentage point drop in the participation rate between 2008 and 2013 implies that about 8 million persons withdrew from the labor force in that 5-year period.[2] The official unemployment rate would obviously be much higher if some of these persons were reclassified as being part of a more broadly defined "labor force."

The reasons for this historic exodus from the labor force are not well understood. But there is probably a link between these labor force trends and the more generous availability of both unemployment insurance benefits and Social Security disability benefits.[3]

Residential Segregation and Black Unemployment

As we have seen, the black unemployment rate is substantially higher than that of whites. Part of this racial gap can be attributed to racial residential segregation, which isolates many African–Americans from jobs and the economic mainstream.

Table 12-2 uses the difference-in-differences methodology to document that the clustering of blacks into a relatively small number of geographic areas contributes to a higher rate

[2] The calculation of this number makes for an interesting exercise. The Bureau of Labor Statistics website (data.bls.gov/cgi-bin/surveymost?ln) reports that the labor force participation rate of civilians aged 16 or more was 62.9 percent in December 2013, and that 154.9 million persons were in the labor force. This implies that the relevant population size (that is, the denominator in calculating the labor force participation rate) was 246.3 million. The BLS website also reports that the labor force participation rate in January 2008 was 66.2 percent, implying that had the participation rate remained constant during the period there would have been 163.1 million persons in the labor force.

[3] Andreas I. Mueller, Jesse Rothstein, and Till M. von Wachter, "Unemployment Insurance and Disability Insurance in the Great Recession," *Journal of Labor Economics* 34 (January 2016, Part 2): S445–S475.

Theory at Work

GRADUATING DURING A RECESSION

Some of us are quite lucky. We somehow manage to time our birth so that the labor market is burning hot the year we graduate from college. It's a seller's market—employers outbid each other to get our services. The wining-and-dining never ends.

Some of us, however, are not as fortunate. Our parents somehow conceived us without thinking of the fact that a couple of decades down the road, we would be graduating from college under very poor economic conditions. Jobs are scarce, and we would be lucky to have a couple of job interviews and extremely lucky to have even one job offer.

It turns out that the harmful consequences of graduating during a recession do not end there, with the hardship of trying to find a paying job after graduation. It is easy to see why labor market conditions at the time of college graduation might affect long-run outcomes. The scarcity of jobs during a severe recession, for instance, might lead young graduates to accept jobs that do not offer much opportunity for training or for moving up the ladder, limiting their options in later years.

Recent research documents the adverse consequences of graduating in a bad economy both in the United States and abroad. A 1-percentage point increase in the national unemployment rate at the time of college graduation is associated with about a 6 percent wage loss initially for American workers. In other words, the initial job pays about 6 percent less than the first job offered to other "luckier" graduation cohorts. Although this large wage effect gets weaker over time, something still remains even after a decade. The wage loss associated with graduating in a poor economy is 2.5 percent 15 years down the line.

A study of the Canadian labor market finds roughly similar results. College graduates who enter the labor market during a recession suffer an initial wage loss of about 9 percent; half of this loss remains after 5 years, but eventually disappears after a decade. Finally, a study of the Japanese labor market finds that the initial wage loss associated with graduating in a recession is about 5 percent, with the wage loss eventually dropping to about 2 percent.

Sources: Lisa Kahn, "The Long-Term Labor Market Consequences of Graduating from College in a Bad Economy," *Labour Economics* 17 (April 2010): 303–316; Philip Oreopoulos, Till von Wachter, and Andrew Heisz, "The Short- and Long-Term Career Effects of Graduating in a Recession." *American Economic Journal: Applied Economics* 4 (January 2012): 1–29; Yuji Genda, Ayako Kondo, and Souichi Ohta, "Long-Term Effects of a Recession at Labor Market Entry in Japan and the United States," *Journal of Human Resources* 45 (Winter 2010): 157–196.

TABLE 12-2 Relation Between Black Residential Segregation and Percentage of Blacks Who Are Idle, 1990

Source: David M. Cutler and Edward L. Glaeser, "Are Ghettos Good or Bad?" *Quarterly Journal of Economics* 112 (August 1997): 842.

Group	City Is Not Very Segregated	City Is Very Segregated	Difference
Blacks aged 20–24	15.4	21.6	6.2
Whites aged 20–24	7.0	6.6	–0.4
Difference-in-differences	—	—	**6.6**

of "idleness" among young blacks, where a person is considered idle if he or she is neither employed nor in school. It turns out that 15.4 percent of young blacks living in cities with little racial residential segregation are idle. In contrast, 21.6 percent of blacks living in highly segregated cities are idle. In short, living in highly segregated cities seems to raise the idleness rate of young blacks by 6.2 percentage points.

But, of course, other factors may be at work. For instance, the industrial composition of the local labor market may differ significantly between the two types of cities. Employment

in highly segregated cities may be concentrated in declining industries, such as manufacturing. Persons living in highly segregated cities would then have higher idleness rates, *regardless* of their race.

As Table 12-2 also shows, however, the idleness rate for whites is not all that different between the two types of cities. In fact, there is slightly *less* idleness among whites in the segregated cities. The difference-in-differences methodology then suggests that racial residential segregation increased the idleness rate of blacks by 6.6 percentage points. The segregation of blacks into a small number of geographic areas may indeed be partly responsible for the less-beneficial labor market opportunities faced by black workers.[4]

12-2 Types of Unemployment

The labor market is in constant flux. Some workers quit their jobs; other workers are laid off. Some firms are cutting back; other firms are expanding. New workers enter the market after completing their education; other workers reenter after spending some time in the household sector. At any time, many workers are in between jobs. If workers looking for jobs and firms looking for workers could find each other immediately, there would be no unemployment. **Frictional unemployment** arises because both workers and firms need time to locate each other.

The existence of frictional unemployment does not suggest that there is a fundamental structural problem in the economy, such as an imbalance between the number of workers looking for work and the number of jobs available. As a result, frictional unemployment is not viewed with much alarm by policymakers. By its very nature, frictional unemployment should lead to short unemployment spells. Moreover, frictional unemployment is "productive" because the search activities of workers and firms improve the allocation of resources in the labor market. There are also easy ways of reducing frictional unemployment, such as providing workers with information about job openings and providing firms with information about unemployed workers.

Many workers also experience **seasonal unemployment**. Workers in some industries are laid off regularly because new models are introduced with clockwork regularity, and firms shut down so that they can be retooled. Spells of seasonal unemployment are usually very predictable. Seasonal unemployment, like frictional unemployment, is also not what concerns policymakers. After all, many of the unemployed workers will return to their former employer once the production season begins.

The type of unemployment that causes anxiety is **structural unemployment**. Suppose the number of workers looking for work equals the number of jobs available; there is no imbalance between the total numbers being supplied and demanded. Structural unemployment can still arise if the kinds of persons looking for work do not "fit" the jobs available. At any time, some sectors of the economy are expanding, and other sectors are contracting. If skills were perfectly transferable from one sector to another, the laid-off workers could quickly move to the growing sectors. But skills might be specific to the

[4] See also Richard W. Martin, "Can Black Workers Escape Spatial Mismatch? Employment Shifts, Population Shifts, and Black Unemployment in American Cities," *Journal of Urban Economics* 55 (January 2004): 179–194.

worker's job or industry and laid-off workers lack the qualifications needed in the growing sector. As a result, the unemployment spells of the displaced workers might last for a long time because they must retool their skills. Structural unemployment arises because of a mismatch between the skills that workers are supplying and the skills that firms are demanding.

The policy prescriptions for this type of structural unemployment are very different from those that would reduce frictional or seasonal unemployment. The problem is skills; the unemployed are stuck with human capital that is no longer useful. To reduce this type of unemployment, therefore, the government would have to provide training programs that "inject" the displaced workers with the skills that are now in demand. And because skills take time to acquire, this type of unemployment may last a while.

There also may be a structural imbalance between the number of workers looking for jobs and the number of jobs available—even if skills were perfectly portable across sectors. This imbalance could arise because of a slowdown in the aggregate economy. Firms now need a smaller workforce to satisfy consumer demand and employers lay off many workers, generating **cyclical unemployment**.

There is then an excess supply of workers, and the market does not clear because the wage is sticky and does not adjust downward. As we will see, economists have developed models that can generate sticky wages and unemployment. The policy prescriptions for cyclical unemployment have little to do with helping workers find jobs or with retooling workers' skills. To reduce this type of unemployment, the government may have to stimulate aggregate demand and reestablish market equilibrium at the sticky wage.

12-3 The Steady-State Rate of Unemployment

The flows of workers across jobs and in and out of the labor market generate some unemployment. It is easy to calculate the steady-state rate of unemployment, the unemployment rate that would be observed in the long run as a result of these labor flows.

To keep things simple, suppose a worker can be either employed or unemployed (so that we will ignore the nonmarket sector). Figure 12-6 describes the labor flows in this simplified economy. There are a total of E employed workers and U unemployed workers. In any given period, let ℓ be the fraction of the employed who lose their jobs and become unemployed, and let h be the fraction of the unemployed workers who find work and get hired. In a steady state, where the economy has reached a long-run equilibrium, the unemployment rate would be constant over time. The steady state then requires that the number of workers who lose jobs equal the number of unemployed workers who find jobs, or

$$\ell E = hU \tag{12-1}$$

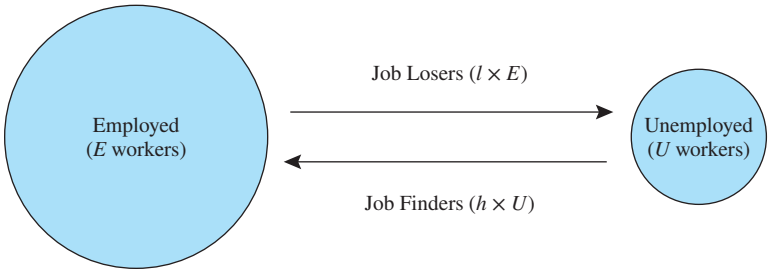
The labor force is defined as the sum of persons who are either employed or unemployed, so $LF = E + U$. Substituting this definition into equation (12-1) yields

$$\ell(LF - U) = hU \tag{12-2}$$

By rearranging terms, we can solve for the steady-state unemployment rate:

$$\text{Unemployment rate} = \frac{U}{LF} = \frac{\ell}{\ell + h} \tag{12-3}$$

FIGURE 12-6 Flows between Employment and Unemployment
A person is either working or unemployed. At any point in time, some workers lose their jobs and unemployed workers find jobs. If the probability of losing a job equals ℓ, there are $\ell \times E$ job losers. If the probability of finding a job equals h, there are $h \times U$ job finders.



Equation (12-3) shows that the steady-state unemployment rate is determined by the transition probabilities between employment and unemployment (ℓ and h). Policies designed to reduce steady-state unemployment must alter these probabilities.

As an example, suppose the probability that employed workers lose their jobs in any given month is 0.01, implying that the average job lasts 100 months. Suppose also that the probability that unemployed workers find work in any given month is 0.10, implying the average unemployment spell lasts 10 months. The steady-state unemployment rate is 9.1 percent, or 0.01/(0.01 + 0.10). The example illustrates that the unemployment rate is smaller when jobs are more stable and larger when unemployment spells last longer. In other words, two key factors determine the unemployment rate: the incidence of unemployment (that is, the probability ℓ that an employed person loses his or her job), and the duration of unemployment spells (which is the inverse of the probability that an unemployed person finds a job, or $1/h$).

The steady-state unemployment rate derived in equation (12-3) is sometimes called the **natural rate of unemployment**. We will provide a more detailed discussion of the factors that determine the natural rate later in the chapter.

Of course, this simple model of labor force dynamics does not accurately describe the actual flows observed in real-world labor markets. There are also flows in and out of the labor force, so a person can be in one of three states: Employed, unemployed, and the nonmarket sector. Figure 12-7 illustrates the magnitude of these flows for the average month between 1990 and 2006. There were 130.0 million persons employed, 7.4 million persons unemployed, and 69.3 million persons in the nonmarket sector. During the typical month, about 1.8 million persons who had a job became unemployed and 1.8 million persons who had been out of the labor force joined the unemployment rolls. At the same time, 2 million of the unemployed found a job and an additional 1.6 million left the labor force.

Incidence versus Duration

Suppose there are 100 unemployed workers in the economy, and that 99 of them are in an unemployment spell that lasts only 1 week. The remaining unemployed worker, however, is in a spell that lasts 101 weeks. Most unemployment spells in this economy would then be short-term spells because most unemployed workers are unemployed for only 1 week.

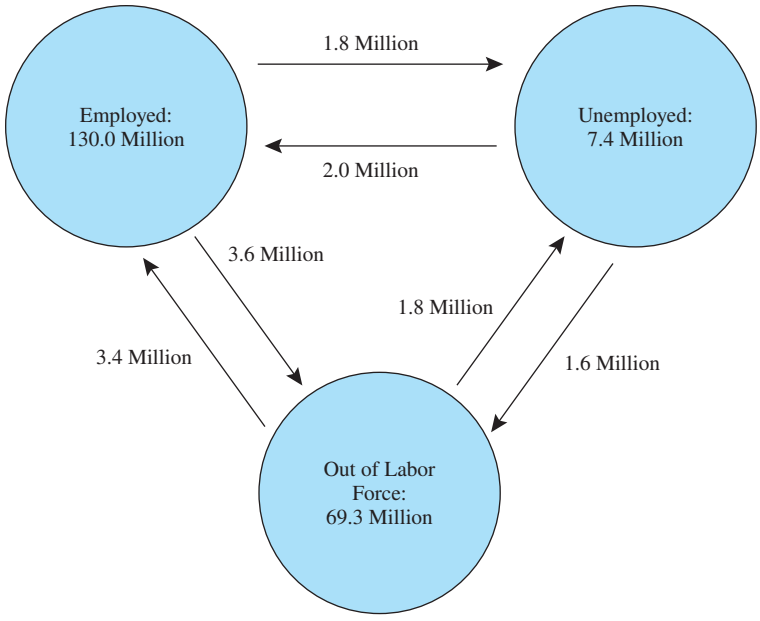


FIGURE 12-7
Dynamic Flows in the U.S. Labor Market, Monthly Average, 1990–2006

Source: Zhi Boon, Charles M. Carson, R. Jason Faberman, and Randy E. Ilg, "Studying the Labor Market Using BLS Labor Dynamics Data," *Monthly Labor Review* (February 2008): 3–16.

At the same time, however, there are a total of 200 weeks of unemployment in this economy (99 weeks for each of the workers with a 1-week spell, plus 101 weeks for the worker with the long spell). Most of the time spent unemployed, therefore, is attributable to a single worker (101/200). In other words, most spells might be short, yet most of the weeks that workers spend unemployed might be attributable to a very few workers with very long spells. As this numerical example suggests, it is important to observe both the incidence and the duration of unemployment in order to draw sensible inferences about the nature of the unemployment problem in any particular labor market.[5]

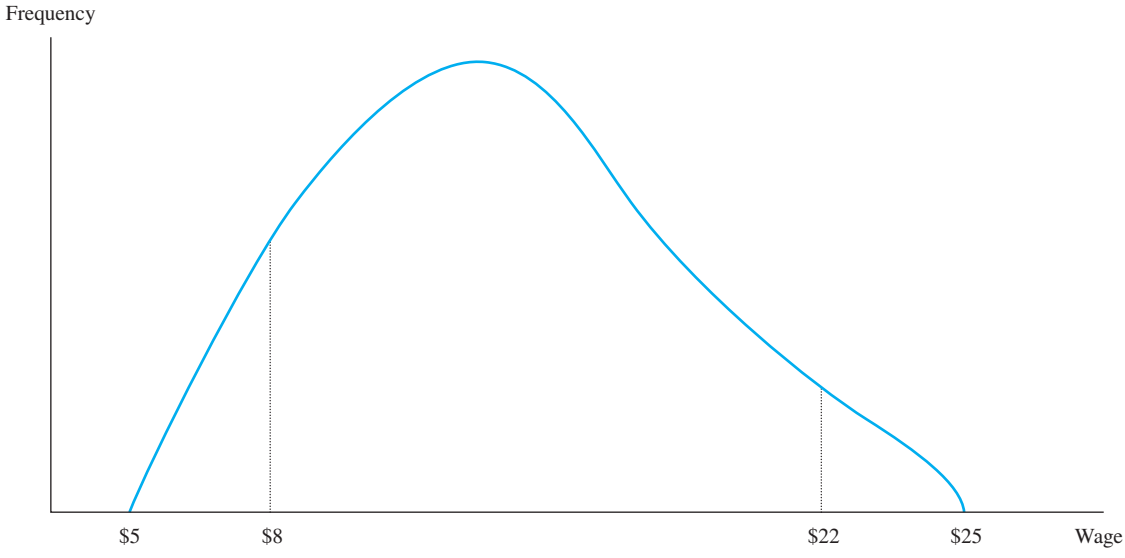
12-4 Job Search

Many theories claim to explain why unemployment exists and persists in competitive markets. We begin our discussion of these alternative stories by reemphasizing that we would observe frictional unemployment even if there were no fundamental imbalance between the supply of and demand for workers. Because different firms offer different job opportunities and because workers are not fully informed about where the best jobs are located, it takes time to find the optimal match.

Any given worker can choose from many different job offers. Just as gas stations that are one block apart charge different prices for a gallon of gas, different firms make different offers to the same worker. These wage differences encourage an unemployed worker to "shop around" until he or she finds a superior job offer. It takes time and effort to learn

[5] Kim B. Clark and Lawrence H. Summers, "Labor Market Dynamics and Unemployment: A Reconsideration," *Brookings Papers on Economic Activity* (1979): 13–60.

FIGURE 12-8 The Wage Offer Distribution
The wage offer distribution gives the frequency distribution of potential job offers for a given worker. The worker can get a job paying anywhere from $5 to $25 per hour.



about what different firms have to offer. Inevitably, search activities prolong the unemployment spell. The worker, however, is willing to endure the longer spell because it might lead to a higher-paying job. In a sense, search unemployment is a form of human capital investment; the worker is investing in information about the labor market.[6]

The Wage Offer Distribution

To simplify the analysis, assume that only unemployed workers search for a job. It is likely that some persons keep on searching for an even better job after they accept a particular job offer. However, it is easier to grasp the main implications of the search model if we restrict our attention to unemployed workers.

The **wage offer distribution** gives the frequency distribution describing the various offers available to a particular unemployed worker. Figure 12-8 illustrates a typical wage offer distribution. As drawn, the worker can end up in a job paying anywhere from $5 to $25 per hour.

We assume that the unemployed worker knows the shape of the wage offer distribution. In other words, he knows that there is a high probability that his search will locate a job paying between $8 and $22 per hour and that there is a small probability that he might end up with a job paying less than $8 or more than $22.

[6] Technical surveys of job search models include Dale T. Mortensen, "Job Search and Labor Market Analysis," in Orley C. Ashenfelter and Richard Layard, editors, *Handbook of Labor Economics,* vol. 2, Amsterdam: Elsevier, 1986, pp. 849–919; Dale T. Mortensen and Christopher A. Pissarides, "New Developments in Models of Search in the Labor Market," in Orley C. Ashenfelter and David Card, editors, *Handbook of Labor Economics,* vol. 3B, Amsterdam: Elsevier, 1999, pp. 2567–2627.

If search were free, the worker would keep on knocking from door to door until he finally hits the firm that pays the $25 wage. Search activities, however, are costly. Each time the worker applies for a new job, he must pay for transportation and other types of expenses, such as a fee to a private employment agency. There is also an opportunity cost: He could have been working at a lower-paying job. The worker's trade-off is clear: The longer he searches, the more likely he will get a high wage offer; the longer he searches, the more it costs to find that job.

Nonsequential and Sequential Search

When should the worker stop searching and settle for the job offer at hand? There are two approaches to answering this question.[7] Each approach gives a "stopping rule" telling the worker when to end his search activities.

The worker could follow a strategy of **nonsequential search**. In this strategy, the worker decides before he begins his search that he will randomly visit, say, 20 firms and accept the offer that pays the highest wage (which will not necessarily be the job paying $25 an hour). This search strategy is *not* optimal. Suppose that on his first try, the worker just happens to hit the firm that pays $25 an hour. A nonsequential search strategy would force this worker to visit another 19 firms knowing full well that he could never do better. It does not make sense, therefore, for the worker to precommit himself to a fixed number of searches regardless of what happens while he is searching.

A better strategy is one of **sequential search**. Before the worker sets out on the search process, he decides which job offers he is willing to accept. For instance, he might decide that he is not willing to work for less than, say, $12 an hour. The worker will then visit one firm and compare the wage offer to his desired $12 wage. If the wage offer exceeds $12, he accepts the job, ending the unemployment spell. If the wage offer is less than $12, he rejects the job offer and starts the search process over again (that is, he will visit a new firm, compare the new wage offer to his desired wage, and so on). The sequential search strategy implies that if a worker is lucky enough to find the $25 job on the first try, he will immediately recognize that he lucked out and stop searching.

The Asking Wage

The **asking wage** is the threshold wage that determines if the unemployed worker accepts or rejects incoming job offers.[8] There is an obvious link between a worker's asking wage and the length of the unemployment spell the worker will experience. Workers who have low asking wages find acceptable jobs very quickly and the unemployment spell will be short. Workers with high asking wages take a long time to find an acceptable job and the unemployment spell will be very long.

[7] The nonsequential search model was introduced by George J. Stigler, "Information in the Labor Market," *Journal of Political Economy* 70 (October 1962): 94–104; the sequential search model was introduced by John J. McCall, "Economics of Information and Job Search," *Quarterly Journal of Economics* 84 (February 1970): 113–126.

[8] The asking wage is called the *reservation wage* in many studies. We use the term *asking wage* to differentiate the threshold that determines whether an unemployed person accepts a job offer from the *reservation wage* defined in the labor supply chapter, which determines whether a person enters the labor market. The intuition for the threshold is the same in both contexts; it is the wage that makes a worker indifferent between two alternative actions.

It is easy to illustrate how the worker sets his asking wage. Consider the wage offer distribution in Figure 12-8. Suppose the unemployed worker goes out and samples a firm at random. By pure chance, he visits the firm that pays the lowest wage possible, $5 per hour. The worker was very unlucky in his first try, and he knows it. He must decide whether to accept or reject this offer by comparing the expected gain from one additional search (by how much would the wage offer increase) with the cost of the additional search. If the offer at hand is only $5, the gain from searching one more time is very high. Even if the worker instantly forgets which firm he visited today, the odds of hitting that $5 firm again tomorrow are very low. An additional search, therefore, will almost surely generate an offer higher than $5 per hour. The marginal gain from one additional search, therefore, is substantial.

Suppose the worker visits another firm, and this time he gets a $10 wage offer. The incentive to continue searching will again depend partly on the marginal gain from one more search. Given the wage offer distribution in Figure 12-8, there is still a good chance that an additional search will generate a higher wage offer. But the gain from this additional search is not as high as when the wage offer at hand was only $5. After all, there is a chance that if he searches one more time, he might hit a firm offering less than $10.

Suppose the worker decides to try his luck one more time. This time he hits the jackpot, getting a wage offer of $25. At this point, the marginal gain from further search is zero. The worker cannot get a higher wage offer.

The marginal gain from search, therefore, is lower if the worker has a good wage offer at hand. As a result, the marginal revenue curve (that is, the gain from one additional search) is downward sloping, as illustrated by the *MR* curve in Figure 12-9.

Of course, the asking wage is determined not only by the marginal gain from searching, but also by the marginal cost of searching. There are two types of search costs. The first is the direct cost of search, which includes transportation costs. The second is the opportunity cost of search. Even if the wage offer at hand is the $5 wage offer, the worker who rejects this offer and searches again is forgoing $5 worth of income. The marginal cost of search is large if the worker has a good wage offer at hand. Therefore, the marginal cost curve (or *MC* in Figure 12-9) is upward sloping.

The intersection of the marginal revenue and marginal cost curves gives the asking wage, $\tilde{w}$. Consider what would happen if the worker gets a wage offer of only $10, which is less than $\tilde{w}$ in the figure. The marginal revenue from search exceeds the marginal cost, and the worker should continue searching. If the wage offer at hand was $20 (above the asking wage), the worker should accept the job because the expected benefit from additional search is lower than the marginal cost of search. The asking wage, therefore, makes the worker indifferent between continuing and ending his search.

Determinants of the Asking Wage

The worker's asking wage will respond to changes in the benefits and costs of search. As with all human capital investments, the benefits from search are collected in the future, so they depend on the worker's discount rate. Workers with high discount rates are present-oriented and will perceive the future benefits from search to be low. As illustrated in Figure 12-10*a*, workers with high discount rates have lower marginal revenue curves (shifting the marginal revenue curve from MR_0 to MR_1) and will have lower asking wages (from $\tilde{w}_0$ to $\tilde{w}_1$). Because these workers do not have the patience to wait until a better offer comes along, they accept lower wage offers and have shorter unemployment spells.

FIGURE 12-9 The Determination of the Asking Wage

The marginal revenue curve gives the gain from an additional search. It is downward sloping because the better the job offer at hand, the less there is to gain from an additional search. The marginal cost curve gives the cost of an additional search. It is upward sloping because the better the job offer at hand, the greater the opportunity cost of an additional search. The asking wage equates the marginal revenue and the marginal cost of search.

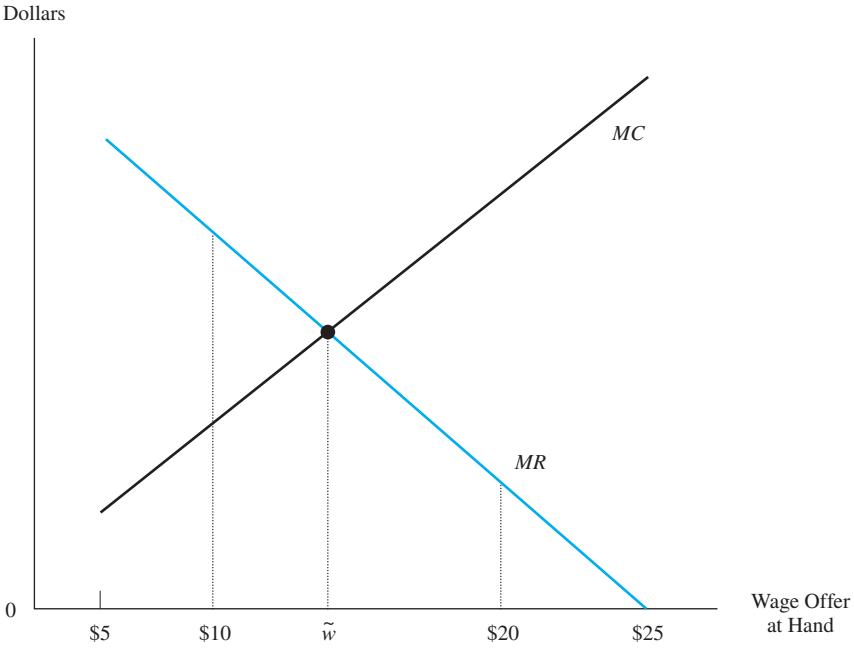
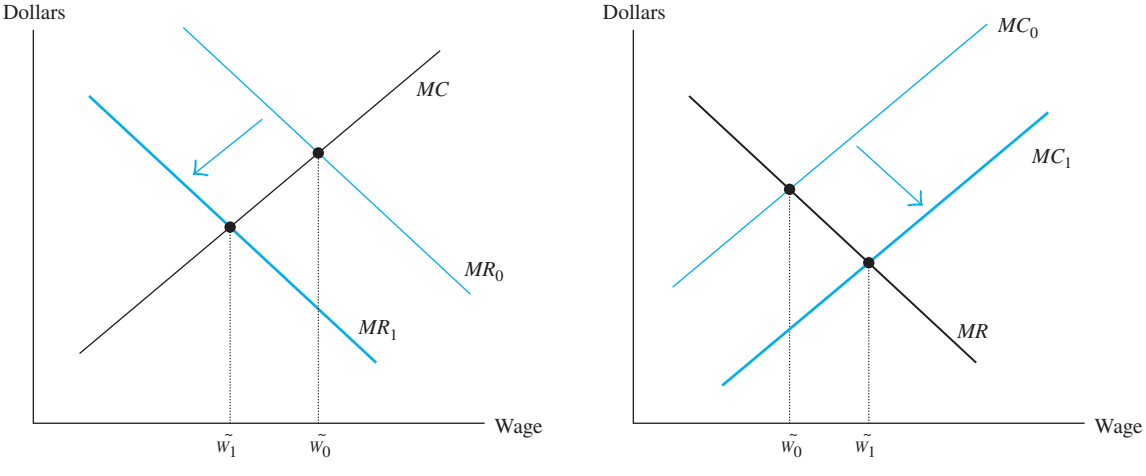


FIGURE 12-10 The Determination of the Asking Wage

(*a*) A "present-oriented" worker has a high discount rate and gains less from additional searches, so the marginal revenue curve shifts to MR_1 and the asking wage falls. (*b*) Unemployment insurance benefits reduce the marginal cost of search and shift the marginal cost curve to MC_1, increasing the asking wage.



(*a*) Increase in Discount Rates

(*b*) Increase in Unemployment Benefits

A major component of search costs is the opportunity cost of rejecting a job offer and continuing the search. The unemployment insurance (UI) system, which we discuss in greater detail below, compensates workers who are unemployed and actively engaged in search activities. Suppose that the worker has a wage offer at hand of $10 per hour (or $400 per week). If he qualifies for UI benefits of $200 per week, the worker is only giving up $200 by rejecting the job offer. Unemployment insurance benefits, therefore, reduce the marginal cost of search.

As Figure 12-10*b* shows, a reduction in the marginal cost of search (from MC_0 to MC_1) raises the asking wage from $\tilde{w}_0$ to $\tilde{w}_1$. The UI system, therefore, has three important effects: (1) It leads to longer unemployment spells, (2) it increases the unemployment rate, and (3) it leads to a higher wage after the unemployment spell ends.

Although the asking wage is not observed directly, a number of surveys have attempted to determine an unemployed worker's asking wage by asking such questions as "What type of work are you looking for?" and "At what wage would you be willing to take this job?"

In 1980, white unemployed youth in the United States reported an average asking wage of $4.30 an hour, and black unemployed youth reported $4.22 an hour.[9] The worker's self-reported asking wage was strongly correlated with the worker's unemployment experience. Workers who reported higher asking wages had longer unemployment spells. Higher asking wages also led to higher wages in their new job; a 10 percent increase in the asking wage increased the wage by 5 percent for young whites and 3 percent for young blacks. In the United Kingdom, where similar surveys have been conducted, a 10 percent increase in the asking wage increases the length of the unemployment spell by at least 5 percent.[10]

Is the Asking Wage Constant over Time?

If the marginal revenue and the marginal cost of search were constant over the length of an unemployment spell, the asking wage would also be constant. An unemployed worker would then have the same chance of finding a job in the 1st week of the spell as in the 30th week.

But the probability of escaping unemployment probably depends on how long the worker has been unemployed. After all, search is costly. The unemployed worker has limited means and will hit a "liquidity constraint" at some point; put simply, he will no longer have the cash to keep his search going.

The liquidity constraint forces the worker to recognize that he cannot spend the rest of his life searching for the best job possible (for the $25 jackpot in Figure 12-8), and that he will have to settle for less. As the worker's cash runs out, therefore, the asking wage falls. The worker will then be willing to accept job offers that were rejected at the beginning of the unemployment spell, so that the probability of escaping unemployment rises the longer the worker has been unemployed.

[9] Harry J. Holzer, "Reservation Wages and Their Labor Market Effects for Black and White Male Youth," *Journal of Human Resources* 21 (Spring 1986): 157–177; Harry J. Holzer, "Job Search by Employed and Unemployed Youth," *Industrial and Labor Relations Review* 40 (July 1987): 601–611.

[10] Stephen R. G. Jones, "The Relationship between Unemployment Spells and Reservation Wages as a Test of Search Theory," *Quarterly Journal of Economics* 103 (November 1988): 741–765.

12-5 Policy Application: Unemployment Compensation

The UI system in the United States is run mainly at the state level. In 2010, at the peak of the Great Recession, the system distributed $150.1 billion in benefits. By 2016, the economic recovery had reduced the cost of the program to $36.2 billion.

The basic parameters of the system are roughly similar across states.[11] When a worker becomes unemployed, he may become eligible for unemployment benefits depending on how long he has been employed and on the reason for the job separation. Workers who are laid off from their jobs typically qualify for unemployment benefits if they have worked for at least two quarters in the year prior to the layoff and if they have had some minimum level of earnings during that year (on the order of $1,000–$3,000 for the year). Workers who quit their jobs, who were fired for just cause, or who are on strike are usually not eligible for unemployment benefits. New labor market entrants or reentrants are also not eligible for benefits.

Eligible workers can collect UI benefits after a waiting period of 1 week. The benefit level depends on the worker's weekly wage: The higher the weekly wage, the larger the benefit. However, there is both a minimum and a maximum level of weekly benefits. In 2017, the minimum level of benefits was $45 in Alabama, $40 in California, and $24 in West Virginia; the maximum level was $265 in Alabama, $450 in California, and $424 in West Virginia.

Because benefits are capped both from below and from above, the **replacement ratio**, the proportion of weekly earnings that are replaced by UI benefits, may be very high for low-income workers but will be low for high-income workers. On average, the replacement ratio was about 40 percent in 2017.

The unemployed worker receives UI benefits as long as he actively seeks work, up to a specified number of weeks. The maximum number of benefit weeks is typically 26, but the benefit period is lengthened if the national or state economy faces particularly adverse conditions. In 2010, for instance, some unemployed workers could have collected UI benefits for a much longer period. In Massachusetts, an unemployed worker could have received benefits for up to 99 weeks. Once a worker exhausts his UI benefits, he no longer qualifies to receive benefits unless he finds another job, works the required number of quarters, and becomes unemployed once again.

UI and the Duration of Unemployment Spells

The structure of the UI system has important implications for the duration of unemployment spells. Higher replacement ratios, for instance, obviously reduce search costs. There should then be a positive correlation between the replacement ratio and the duration of the spell.

This prediction of search theory has been confirmed by many studies. A 25-percent rise in the replacement ratio (from, say, 0.4 to 0.5) increases the average duration of an unemployment spell by about 15–25 percent.[12] The BLS reports that the typical unemployment

[11] The U.S. Department of Labor maintains a website summarizing how the Unemployment Insurance system is financed in each state; see www.ows.doleta.gov/unemploy/sig_measure.asp.

[12] Kathleen P. Classen, "The Effect of Unemployment Insurance on the Duration of Unemployment and Subsequent Earnings," *Industrial and Labor Relations Review* 30 (July 1977): 438–444; Robert R. Moffitt, "Unemployment Insurance and the Distribution of Unemployment Spells," *Journal of Econometrics* 28 (April 1985), 85–101; Patricia M. Anderson and Bruce D. Meyer, "The Effects of the Unemployment Insurance Payroll Tax on Wages, Employment, Claims and Denials," *Journal of Public Economics* 78 (October 2000): 81–106.

spell in 2017 lasted an average of 25 weeks. Reducing the replacement ratio from 0.4 to 0.3 (or a 25 percent cut in the ratio) would reduce the average length of an unemployment spell by about 5 weeks. The UI system, therefore, can have a numerically important impact on the duration of unemployment.[13]

Moreover, because the replacement ratio tends to be larger for low-skill workers, these workers will have relatively higher asking wages and longer unemployment spells.[14] The observation that low-skill workers have longer unemployment spells need not imply that these workers have a particularly difficult time finding new jobs.

After collecting UI benefits for a specified time period (typically 26 weeks), an unemployed worker does not qualify for additional benefits. The benefit cut in the 26th week, therefore, substantially raises the cost of search. The worker will likely reduce his asking wage at that point, and we should expect to see a noticeable increase in the exit rate from unemployment.

The evidence indeed shows that a job-seeking worker's chance of finding a job improves dramatically the week benefits run out. Figure 12-11 illustrates how the probability that unemployed workers find a new job depends on the number of weeks remaining until exhaustion of benefits. A worker with 5–10 weeks of UI benefits left has a probability of finding a job (on any given week) of about 3 percent. But the probability spikes to almost 8 percent the week the benefits run out.

The UI system not only lengthens the duration of unemployment spells, but also leads to a higher wage in the new job. A 10 percent increase in the replacement ratio increases the subsequent wage by 2–7 percent.[15] The evidence, therefore, confirms the implications of the search model of unemployment: Lower search costs increase both the duration of spells and the eventual wage.

Many studies have documented the impact of UI by exploiting idiosyncratic legislative changes in the parameters that determine benefits. For example, in a peculiar deal that was struck to gain the support of labor unions, New Jersey extended UI benefits for 13 additional weeks to persons who exhausted their regular UI benefits between June 2 and November 24 of 1996. Despite the very short-run nature of this benefit extension, and despite the fact that many of those affected probably began looking for work prior to June 2, persons in this "notch" had a higher probability of exhausting benefits and qualifying for the additional 13 weeks.[16]

[13] There is also evidence suggesting that eligibility for UI encourages workers to have shorter jobs, presumably because it increases the propensity for employed workers to keep on searching; see Stepan Jurajda, "Estimating the Effect of Unemployment Insurance Compensation on the Labor Market Histories of Displaced Workers," *Journal of Econometrics* 108 (June 2002): 227–252; and Audrey Light and Yoshiaki Omori, "Unemployment Insurance and Job Quits," *Journal of Labor Economics* 22 (January 2004): 159–188.

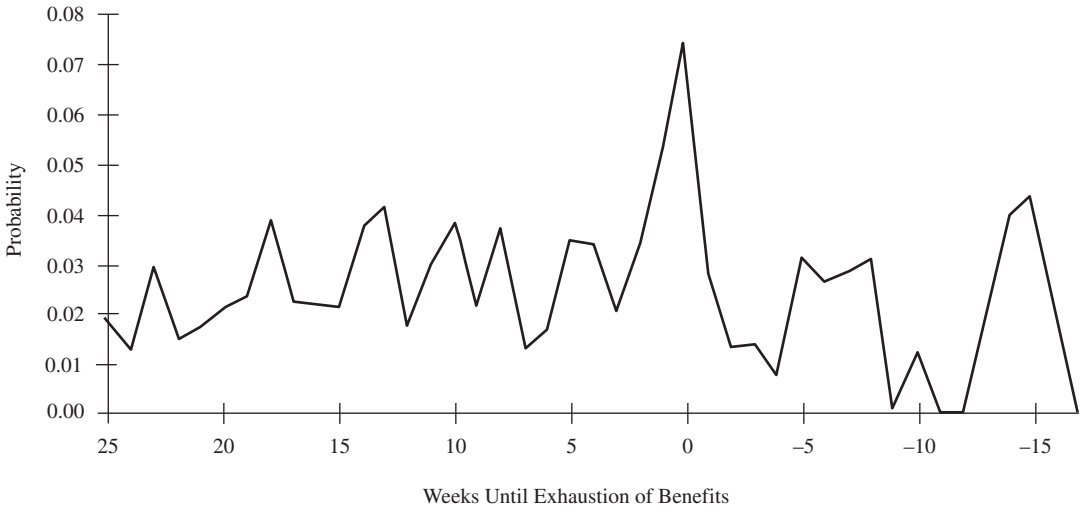
[14] Bruce D. Meyer, "Unemployment Insurance and Unemployment Spells," *Econometrica* 58 (July 1990): 757–782; Olympia Bover, Manuel Arellano, and Samuel Bentolila, "Unemployment Duration, Benefit Duration and the Business Cycle," *Economic Journal* 112 (April 2002): 223–265.

[15] Ronald G. Ehrenberg and Ronald Oaxaca, "Unemployment Insurance, Duration of Unemployment, and Subsequent Wage Gain," *American Economic Review* 66 (December 1976): 754–766.

[16] David Card and Phillip B. Levine, "Extended Benefits and the Duration of UI Spells: Evidence from the New Jersey Extended Benefit Program," *Journal of Public Economics* 78 (October 2000): 107–138; see also Johannes F. Schmieder, Till von Wachter, and Stefan Bender, "The Effects of Extended Unemployment Insurance Over the Business Cycle: Evidence from Regression Discontinuity Estimates Over 20 Years," *Quarterly Journal of Economics* 127 (May 2012): 701–752.

FIGURE 12-11 The Relationship between the Probability of Finding a New Job and UI Benefits

Source: Lawrence F. Katz and Bruce D. Meyer, "Unemployment Insurance, Recall Expectations, and Unemployment Outcomes," *Quarterly Journal of Economics* 105 (November 1990): 973–1002, Figure IV.



Similarly, as a response to the economic distress caused by the Great Recession, many states extended the availability of UI benefits from 26 weeks to 99 weeks between 2009 and 2012. The data indicate that this benefit extension reduced the exit rate from unemployment, mainly because fewer of the unemployed left the labor force (choosing instead to exhaust benefits).[17] About a quarter of the long-term unemployment observed during this severe downturn could be attributed to the extension of benefits.

There is also evidence that changing the parameters of the UI system has strong effects on unemployment duration in many European countries. In Switzerland, for example, government authorities are required to inform an unemployed person that he is going to be investigated for noncompliance with the eligibility requirements. Not surprisingly, this warning has a sizable impact on the speed with which unemployed workers find jobs.[18]

[17] Henry S. Farber and Robert G. Valletta, "Do Extended Unemployment Benefits Lengthen Unemployment Spells? Evidence from Recent Cycles in the U.S. Labor Market," *Journal of Human Resources* 50 (Fall 2015): 873–909; see also Marcus Hagedorn, Fatih Karahan, Iourii Manovskii, and Kurt Mitman, "Unemployment Benefits and Unemployment in the Great Recession: The Role of Macro Effects," NBER Working Paper No. 19499, October 2013.

[18] Rafael Lalive, Jan C. van Ours, and Josef Zweimuller, "The Effect of Benefit Sanctions on the Duration of Unemployment," *Journal of the European Economic Association* 3 (December 2005): 1386–1417. For related studies that examine the Slovenian and Norwegian labor markets, respectively, see Jan C. van Ours and Milan Vodopivec, "How Shortening the Potential Duration of Unemployment Benefits Affects the Duration of Unemployment: Evidence from a Natural Experiment," *Journal of Labor Economics* 24 (April 2006): 351–378 and Knut Roed and Tao Zhang, "Does Unemployment Compensation Affect Unemployment Duration?" *Economic Journal* 113 (January 2003): 190–206.

Temporary Layoffs

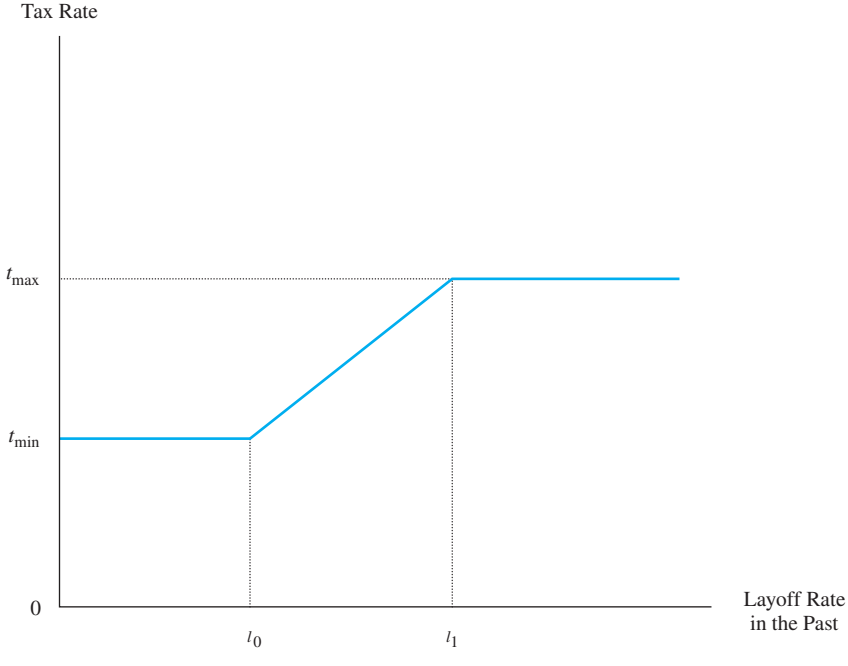
In January 2018, nearly 20 percent of unemployed workers were on **temporary layoffs**, expecting to return to their former employer at the end of the unemployment spell. And the employment practices of manufacturing firms suggest that they rehire more than 70 percent of the workers they lay off.[19] It turns out that the way in which the UI system is financed encourages employers to "overuse" temporary layoffs.

Unemployment insurance is funded by a payroll tax on employers. Typically, a state decides on a taxable wage base, indicating the maximum worker's salary that is subject to the UI payroll tax. This cap varies across states.[20] In 2017, the taxable wage base in Arizona was $7,000; in Massachusetts, $15,000; and in Oregon, $38,400. The state also chooses a tax rate t that the firm pays on the wage base.

The tax rate depends on a number of variables, including the general health of the state's economy, the layoff history of firms in that industry, and the firm's own layoff history. As Figure 12-12 shows, firms that had high layoff rates in the past are typically assessed higher tax rates. The maximum tax rate a firm can be assessed, however, is capped at t_{max}. If the firm rarely uses layoffs, it is assessed a low tax rate, but this tax rate is no lower

FIGURE 12-12 Funding the UI System: Imperfect Experience Rating

If the firm has had very few layoffs in the past (below threshold ℓ_0), the firm is assessed a very low tax rate to fund the UI system. Firms that have had many layoffs are assessed a higher tax rate, but the tax rate is capped at t_{max}.



[19] Martin Feldstein, "The Importance of Temporary Layoffs: An Empirical Analysis," *Brookings Papers on Economic Activity* 3 (1975): 725–744.

[20] U.S. Department of Labor, Employment and Training Administration, *Comparison of State Unemployment Laws,* 2017; available at www.unemploymentinsurance.doleta.gov/unemploy/comparison2017.asp.

than some rate t_{min} (which in some states is zero). In California, for example, the minimum and maximum tax rates, respectively, are 1.5 percent and 6.2 percent; in Kansas, 1.8 and 9.2 percent; and in Massachusetts, 1.21 and 18.55 percent.

Although this method of determining an employer's tax rate is guided by the belief that employers who use the UI system should pay for it, the system does not perfectly impose the tax burden on employers who initiate the most layoffs. Because the tax rate is capped at t_{max}, employers who lay off many workers do not pay their "fair share" of the cost and are instead subsidized by other firms. The determination of the employer's tax rate, therefore, uses an **imperfect experience rating**.

To see how this imperfect experience rating encourages employers to rely on temporary layoffs, consider a labor market where workers and firms are engaged in long-term contracts, perhaps because of specific training.[21] Suppose economic conditions worsen temporarily. The financing of the UI system implies that employers who lay off many workers do not pay the entire cost of the worker's "salary" during the unemployment spell (that is, the unemployment benefits). The firm can then lay off workers and shift part of the payroll to other taxpayers during the period of economic hardship. The bond between worker and firm implies that both parties find it worthwhile to continue the employment relationship. As a result, workers do not want to look for alternative employment because they expect to be recalled to their jobs, and firms do not want to look for other workers because the existing pool of workers is valuable to the firm. Imperfect experience rating, therefore, allows firms to use taxpayer funds to "ride over" some of the rough waves in the economy.

Imperfect experience rating has a significant impact on the layoff behavior of firms. Not surprisingly, the probability that an unemployed worker is recalled to his job increases substantially the week that unemployment benefits run out. In the weeks prior to the exhaustion of benefits, the probability of being recalled is only about 1–2 percent per week. In the week when benefits are exhausted, the probability of recall rises to more than 5 percent.[22] In short, employers use the taxpayer subsidy for as long as they can. Another example of the strong correlation between temporary layoffs and UI is the pattern of seasonal unemployment exhibited by many industries. Firms located in states with low tax rates make heavy use of temporary layoffs during the slow season.[23]

Not surprisingly, the frequency of temporary layoffs affects not only how firms behave, but also how unemployed workers spend their time. The likelihood of being recalled to the job greatly reduces the intensity of job search activities. A study of time-use data that yield a daily diary of how workers allocate their time documents the impact.[24] The typical

[21] Martin Feldstein, "Temporary Layoffs in the Theory of Unemployment," *Journal of Political Economy* 84 (October 1976): 937–958; Robert H. Topel, "On Layoffs and Unemployment Insurance," *American Economic Review* 73 (September 1983): 541–559.

[22] Lawrence F. Katz and Bruce D. Meyer, "Unemployment Insurance, Recall Expectations, and Unemployment Outcomes," *Quarterly Journal of Economics* 105 (November 1990): 973–1002.

[23] David Card and Phillip B. Levine, "Unemployment Insurance Taxes and the Cyclical and Seasonal Properties of Unemployment," *Journal of Public Economics* 53 (January 1994): 1–30.

[24] Alan B. Krueger and Andreas Mueller, "Job Search and Unemployment Insurance: New Evidence from Time Use data," *Journal of Public Economics* 94 (April 2010): 298–307.

Theory at Work

CASH BONUSES AND UNEMPLOYMENT

Because of the disincentive effects of UI, there are many calls for reform of the system, and some states have conducted experiments to see if various policy changes shorten the duration of unemployment spells. In these experiments, some of the workers applying for UI benefits are offered a cash bonus if they find jobs relatively quickly. This random sample of unemployed workers forms the treatment group. The remaining unemployed workers compose the control group and participate in the typical UI program.

In Illinois, workers in the treatment group who found a job within 11 weeks (and who kept that job for at least 4 months) were given a cash bonus of $500, or about four times the average weekly benefit. In Pennsylvania, unemployed workers in the treatment group who found a job within 6 weeks were entitled to a bonus equal to six times the weekly benefit amount.

The evidence from these experiments is unambiguous. Unemployed workers who are offered cash bonuses have shorter unemployment spells than workers in the control group.

Surprisingly, the treated workers did not end their unemployment spells quickly by accepting lower-paying jobs. The average wage after the unemployment spell ended was essentially the same for workers in the treatment and control groups. Offering cash incentives to find jobs quickly, therefore, seems to increase the intensity of the search process, speeds up the transition out of unemployment, and achieves this without a decline in the economic status of workers.

Sources: Stephen Woodbury and Robert Spiegelman, "Bonuses to Workers and Employers to Reduce Unemployment: Randomized Trials in Illinois," *American Economic Review* 77 (September 1987): 513–550; Bruce D. Meyer, "Lessons from the U.S. Unemployment Insurance Experiments," *Journal of Economic Literature* 33 (March 1995): 91–131.

unemployed worker who expects to be recalled to the job spends only 13 minutes per day searching, as compared to 45 minutes for an unemployed worker who lost his job permanently. Put differently, the temporary nature of the layoff cuts the amount of time that an unemployed worker devotes to search by about 75 percent.

12-6 The Intertemporal Substitution Hypothesis

Job search models provide a sensible explanation for the existence of frictional unemployment. This type of unemployment is voluntary in the sense that workers invest in information in return for a higher wage in their new jobs. It has been proposed that the large increase in unemployment observed during a severe economic downturn might also have a voluntary component.[25]

The theory of labor supply over the life cycle, introduced in the labor supply chapter, predicts that workers have an incentive to work in those years when the wage is high and to consume leisure in those years when the wage is low. The **intertemporal substitution hypothesis** also has interesting implications for how workers allocate their time over the business cycle.

[25] The influential hypothesis was first proposed by Robert E. Lucas and Leonard Rapping, "Real Wages, Employment, and Inflation," *Journal of Political Economy* 77 (September/October 1969): 721–754.

Suppose that the real wage fluctuates over the business cycle and that this fluctuation is procyclical; in other words, the real wage rises when the economy expands and declines when the economy contracts. Because it is cheap to consume leisure when the real wage is low, workers are more than willing to work less during recessions. They can collect UI benefits while unemployed, or perhaps leave the labor force altogether. Put differently, some of the unemployment observed in economic downturns might be voluntary because workers are taking advantage of the decline in the real wage to consume leisure.

The intertemporal substitution hypothesis makes two key assumptions: (1) The real wage is procyclical and (2) labor supply responds to shifts in the real wage.

The question of whether real wages are sticky over the business cycle is one of the oldest questions in macroeconomics. Although there is a growing consensus that wages are indeed procyclical, the size of the correlation between wages and the business cycle has not been established conclusively and, in fact, seems to vary across recessions.[26]

The cyclical movement of the real wage is difficult to observe because the composition of the labor force changes over the cycle. Unemployment typically has a particularly adverse effect on low-skill workers. When we calculate the average wage of workers during an economic expansion, we are using a very different sample than when we calculate it during a recession. Although it was widely believed for many years that real wages were sticky, studies that correct for this "composition" bias suggest that the real wage is procyclical.

But the interpretation of unemployment during recessions as a rational (and voluntary) reallocation of a worker's time also requires the assumption that labor supply is elastic, responding to changes in the wage. But labor supply curves—particularly for men—tend to be inelastic. In fact, the large drop in labor supply observed during recessions can be interpreted as an intertemporal substitution only if the labor supply elasticity is far larger than what is typically estimated.[27] In short, it seems doubtful that the increase in unemployment observed during downturns can be dismissed as a voluntary reallocation of the worker's time.

12-7 The Sectoral Shifts Hypothesis

Although job search activities help us understand the presence of frictional unemployment, they do not explain the existence and persistence of long-term unemployment. A number of alternative models have been proposed to explain why structural unemployment might arise in a competitive market.

One important explanation stresses the possibility that workers who are searching for jobs are not qualified to fill the available vacancies. It is well known that shifts in aggregate demand do not affect all sectors of the economy equally. At any point in time, some sectors of the economy are expanding, and other sectors are in decline.

[26] Mark J. Bils, "Real Wages over the Business Cycle: Evidence from Panel Data," *Journal of Political Economy* 93 (August 1985): 666–689; Gary Solon, Robert Barsky, and Jonathan A. Parker, "Measuring the Cyclicality of Real Wages: How Important Is Composition Bias?" *Quarterly Journal of Economics* 109 (February 1994): 1–25; and Michael W. Elsby, Donggyun Shin, and Gary Solon, "Wage Adjustments in the Great Recession and Other Downturns: Evidence from the United States and Great Britain," *Journal of Labor Economics* 34 (January 2016, Part 2): S249–S291.

[27] Solon, Barsky, and Parker, "Measuring the Cyclicality of Real Wages."

To see how these sector-specific shocks might create structural unemployment, suppose the manufacturing industry is hit by an adverse shock. Because of the reduced demand for their output, manufacturers lay off many workers. Favorable shocks to other sectors (such as the computer industry) increase the demand for labor by high-tech firms. If the skills of laid-off manufacturing workers could be easily transferred across industries, the adverse conditions in the manufacturing sector would not lead to long-term unemployment. The laid-off workers would leave the manufacturing sector and move on to jobs in the now-thriving high-tech sector. There would be frictional unemployment as workers learned about and sampled the various opportunities available in the computer industry.

Manufacturing workers, however, probably have skills that are partly specific to the manufacturing sector, so that their skills may not be very useful to computer firms. Long-term unemployment arises because it will take time for these workers to retool and acquire the skills that are now in demand. The **sectoral shifts hypothesis** suggests that there will be a pool of workers who are unemployed in long spells because of a structural imbalance between the skills of unemployed workers and the skills that employers are looking for.[28]

Although the evidence suggests that sectoral shifts do lead to unemployment, there is disagreement over just how much unemployment these shifts are responsible for. The typical empirical analysis relates the aggregate unemployment rate to the dispersion in employment growth rates across industries. The sectoral shifts hypothesis implies that the unemployment rate rises when there is a lot of dispersion in employment growth rates across industries (in other words, when some industries are growing, and some are declining). The evidence indeed documents a positive correlation between measures of dispersion in employment growth rates and the aggregate unemployment rate.[29]

Some studies have also tested the sectoral shifts hypothesis by noting that sectoral shocks should also affect stock market prices, with stock prices rising when firms are hit by favorable shocks and declining when firms are hit by adverse shocks. The dispersion in the change in stock prices across industries, therefore, provides information about the importance of sectoral shocks in the economy. It turns out that there is also a positive correlation between the dispersion in movements of stock prices and the unemployment rate.[30]

12-8 Efficiency Wages and Unemployment

As we saw in the chapter on incentive pay, firms might set an efficiency wage above the competitive wage when they find it expensive to monitor the worker's output. The high efficiency wage "buys" the worker's cooperation, discouraging shirking. Because the firm pays above-market wages, however, efficiency wage models can generate involuntary unemployment.

[28] David M. Lilien, "Sectoral Shifts and Cyclical Unemployment," *Journal of Political Economy* 90 (August 1982): 777–793.

[29] A critical appraisal of the evidence is given by Katharine G. Abraham and Lawrence F. Katz, "Cyclical Unemployment: Sectoral Shifts or Aggregate Disturbances," *Journal of Political Economy* 94 (June 1986): 507–522.

[30] S. Lael Brainard and David M. Cutler, "Sectoral Shifts and Cyclical Unemployment Reconsidered," *Quarterly Journal of Economics* 108 (February 1993): 219–243.

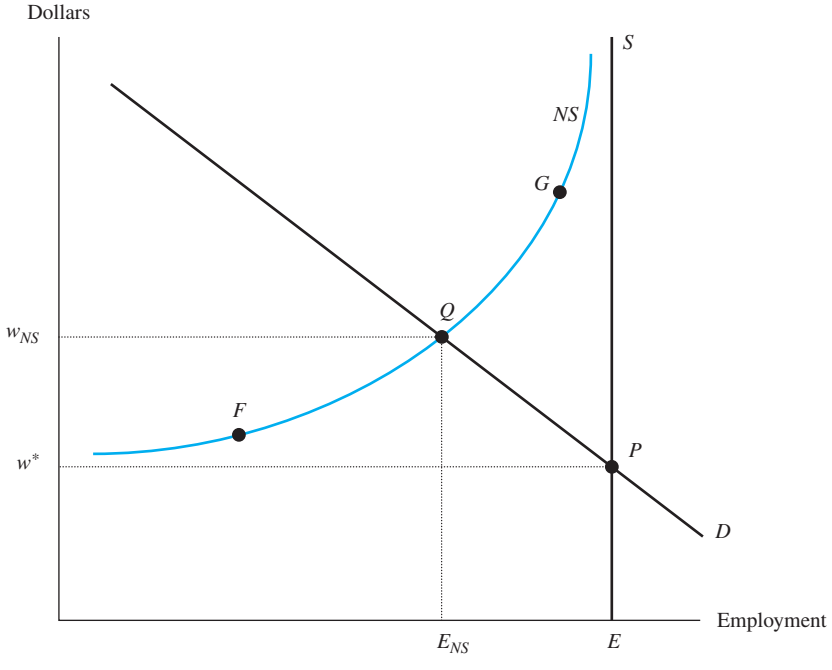
The No-Shirking Frontier

We can interpret the unemployment caused by the efficiency wage as the "stick" that keeps the well-paid workers in line.[31] To see why, consider first the wage–employment outcome in a competitive labor market where worker shirking is not a problem (perhaps because workers can be monitored at a very low cost). There are E workers in this labor market, and the labor supply curve is inelastic. Point P in Figure 12-13 gives the traditional competitive equilibrium, where the vertical supply curve S intersects the downward-sloping labor demand curve D. The market-clearing competitive wage is w^*.

Suppose now that firms cannot easily monitor the output of workers, so monitoring activities are costly. To simplify, let's assume that workers who shirk spend all their time uselessly surfing the web; in short, shirking workers are completely unproductive. The firm, therefore, will want to offer a wage–employment package that encourages its workers not to shirk at all.

FIGURE 12-13 The Efficiency Wage

If shirking is not a problem, the market clears at wage w^* (where supply S equals demand D). If monitoring is expensive, the threat of unemployment can keep workers in line. If unemployment is high (point F), firms can attract workers who will not shirk at a very low wage. If unemployment is low (point G), firms must pay a very high wage to ensure that workers do not shirk. The efficiency wage w_{NS} is given by the intersection of the no-shirking frontier (NS) and the demand curve.



[31] Carl Shapiro and Joseph E. Stiglitz, "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review* 74 (June 1984): 433–444.

Which wage must firms offer to ensure that workers do not shirk? Suppose the unemployment rate is very high. A worker will quickly realize that it is costly to shirk because if he gets caught and fired, he faces a long unemployment spell. As a result, firms would be able to attract workers who will not shirk even if they pay a relatively low wage. If the unemployment rate is very low, however, shirking workers who are caught and fired face only a short unemployment spell. To make shirking costly and to make even the short unemployment spell unprofitable, firms would have to offer the worker a high wage.

This discussion generates an upward-sloping **no-shirking frontier** (labeled *NS* in Figure 12-13), which gives the number of nonshirking workers that firms can attract at each wage. The no-shirking frontier states that when firms employ few workers out of the total E (point F), they can attract nonshirking workers at a low wage because a layoff leads to a long and costly unemployment spell. If firms hire a large number of workers (point G), they must pay higher wages to encourage workers not to shirk.

Note that the no-shirking frontier *NS* will never touch the perfectly inelastic supply curve at E workers and that the horizontal difference between the two curves gives the number of workers who are unemployed. If the market employs all the workers at a particular wage, a shirking worker who gets fired can simply walk across the street and get another job. In other words, there is no penalty for shirking. The key insight provided by the efficiency wage model is clear: Some unemployment is necessary to keep the employed workers in line.

Efficiency Wages and Unemployment

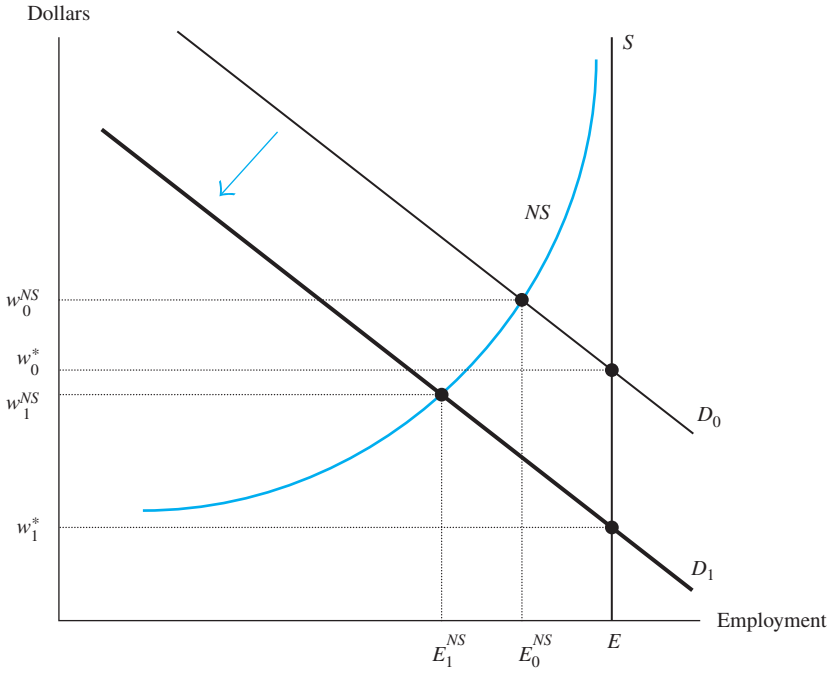
The equilibrium efficiency wage is given by the intersection of the no-shirking frontier and the labor demand curve (at point Q). The wage w_{NS} is the efficiency wage and firms will employ E_{NS} workers, so that $(E - E_{NS})$ workers will be unemployed.

The equilibrium at point Q has a number of notable properties.

1. There are no market forces driving the efficiency wage w_{NS} down toward the competitive wage w^*. If the wage was higher than the efficiency wage w_{NS}, the no-shirking frontier tells us that there will be many nonshirking workers, but firms are only willing to hire a few of them, putting downward pressure on the wage. But if the wage drops below the efficiency wage, firms would want to hire many more nonshirking workers than would be available, and the wage would rise. Therefore, the efficiency wage w_{NS} is above the market-clearing competitive wage.
2. Employed workers will not shirk. The efficiency wage w_{NS} is the wage that encourages the E_{NS} employed workers to behave.
3. There is involuntary unemployment. The $(E - E_{NS})$ unemployed workers want to work at the going wage but cannot find jobs. Firms do not wish to employ these workers because full employment, and the implied lower competitive wage, encourages workers to shirk and reduces the firm's profits.

The structural unemployment generated by efficiency wages is very different from the frictional unemployment generated by job search. Search unemployment is productive; it is an investment in information that eventually leads to a higher-paying job. The unemployment due to efficiency wages is involuntary and unproductive (from the worker's point of view). The worker would like a job but cannot find one, and he has nothing to gain from being in a long unemployment spell. From the firm's point of view, however, the involuntary unemployment is productive. It keeps the employed workers honest, increasing profits.

FIGURE 12-14 An Economic Contraction and the Efficiency Wage
A fall in aggregate demand shifts the labor demand curve from D_0 to D_1. The competitive wage falls from w_0^* to w_1^*. If firms pay an efficiency wage, the contraction also reduces the efficiency wage but by a smaller amount (from w_0^{NS} to w_1^{NS}).



The efficiency wage model also implies that wages will be relatively sticky over the business cycle. Suppose that aggregate demand falls because of a sudden downturn in economic activity. In a competitive market, the labor demand curve shifts down from D_0 to D_1 and the competitive wage drops from w_0^* to w_1^* (see Figure 12-14). If firms paid an efficiency wage, the same decline in demand lowers the wage from w_0^{NS} to w_1^{NS}. Therefore, the efficiency wage is less responsive to changes in demand than the competitive wage. Moreover, employment falls from E_0^{NS} to E_1^{NS} during the contraction and the unemployment rate rises.

The Wage Curve

Some empirical studies document a negative correlation between wages and unemployment across regional labor markets.[32] Specifically, the wage tends to be high in regions where the unemployment rate is low, and the wage tends to be low in regions where the unemployment rate is high. This correlation, which is known as the **wage curve**, is difficult to understand in the context of a competitive supply–demand framework. But it is consistent with an efficiency wage model.

Unemployment can only be observed in the standard model of a competitive labor market if the wage is relatively high—above equilibrium—and if it is sticky. The high wage

[32] David G. Blanchflower and Andrew J. Oswald, *The Wage Curve,* Cambridge, MA: MIT Press, 1994.

ensures that there are more workers who want to work than firms are willing to hire, and the stickiness means that the "excess" workers will keep looking for jobs and not find them. Note that it is *high* wages that are associated with high unemployment, the opposite of what is implied by the wage curve, where unemployment is low when wages are high.

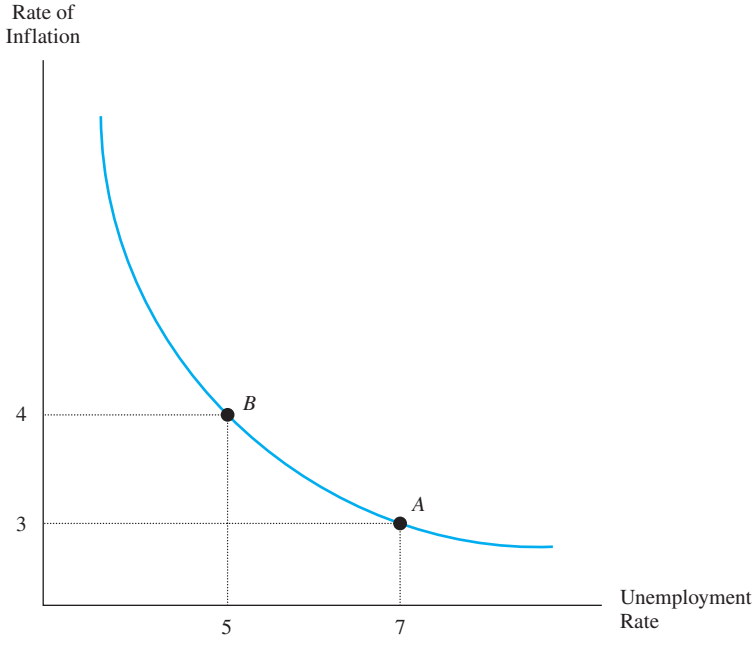
In contrast, the efficiency wage model in Figure 12-14 suggests that the wage will be relatively low when unemployment is high because the high unemployment keeps the workers in line. But wages must be relatively high when there is little unemployment, as the high wage will be required to impose a cost on workers who shirk. The efficiency wage model, therefore, predicts a negative correlation between unemployment and wages, precisely the correlation captured by the wage curve.

12-9 Policy Application: The Phillips Curve

In 1958, A. W. H. Phillips published a famous study documenting a negative correlation between the rate of inflation and the rate of unemployment in the United Kingdom from 1861 to 1957.[33] The negative relation between these two variables, illustrated in Figure 12-15, is now known as the **Phillips curve**.

FIGURE 12-15 The Phillips Curve

The Phillips curve describes the negative correlation between the inflation rate and the unemployment rate. The curve may imply that an economy faces a trade-off between inflation and unemployment.



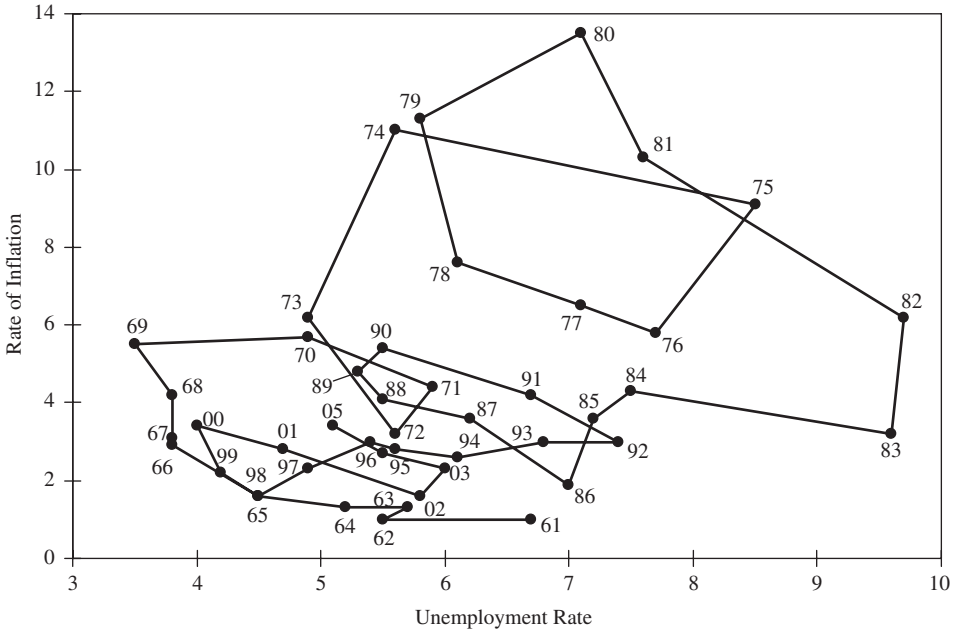
[33] A. W. H. Phillips, "The Relation between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957," *Economica* 25 (November 1958): 283–299.

The Phillips curve is important because it suggests that there might be a trade-off between inflation and unemployment. Suppose, for instance, that the unemployment rate is 7 percent and that the inflation rate is 3 percent, as at point *A* in the figure. The Phillips curve implies that the government could pursue expansionary policies that would move the economy to point *B*, where the unemployment rate falls to 5 percent and the inflation rate rises to 4 percent. The belief that this trade-off provided policymakers with an opportunity to permanently solve the unemployment problem is illustrated by an observation made by Nobel Prize–winning economist William Vickrey: "If unemployment could be brought down to, say, 2 percent at the cost of an assured steady rate of inflation of 10 percent per year, or even 20 percent, this would be a good bargain."

The experience of the U.S. economy during the 1960s seemed to confirm the belief that there was a trade-off between inflation and unemployment. Figure 12-16 illustrates the various inflation–unemployment outcomes observed between 1961 and 2005. Remarkably, the data points between 1961 and 1969 suggested that the United States was moving up a stable Phillips curve. As the figure makes clear, however, the confidence of policymakers in the inflation–unemployment trade-off was shattered during the 1970s. The data points simply refused to cooperate and lie along a stable Phillips curve. Instead, the relationship between inflation and unemployment went "out of kilter." If anything, there seem to be a number of different Phillips curves generated by the data points. For example, the data between 1976 and 1979 lie on a different Phillips curve than the one traced by the 1980–1983 points and from the one traced by the 2000–2002 points.

FIGURE 12-16 Inflation and Unemployment in the United States, 1961–2005

Sources: The unemployment rate data are drawn from U.S. Bureau of Labor Statistics, "Historical Data for the 'A' Tables of the Employment Situation Release. Table A-15, Alternative Measures of Labor Underutilization," stats.bls.gov/cps/cpsatabs.htm. The inflation rate is drawn from U.S. Bureau of Labor Statistics, "Table Containing History of CPI-U U.S. All Items Indexes and Annual Percent Changes from 1913 to Present," stats.bls.gov/cpi/home.htm#tables.



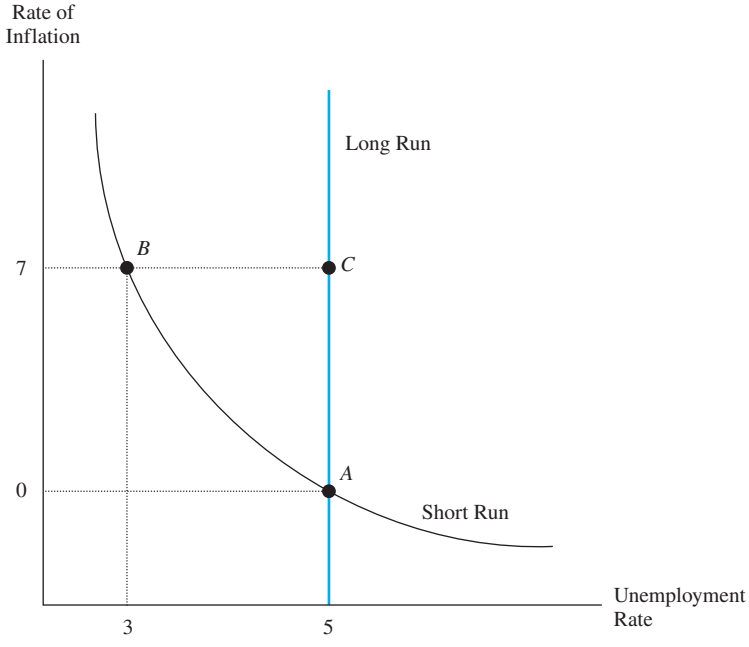
The Natural Rate of Unemployment

At the same time that the inflation–unemployment experience of the 1970s was demolishing the notion of a stable Phillips curve, some economists began to argue that a long-run trade-off between inflation and unemployment did not make theoretical sense.[34] Instead, they argued, economic theory implies that the long-run Phillips curve must be vertical. Put differently, there exists an equilibrium unemployment rate, now called the *natural rate of unemployment,* that persists regardless of the rate of inflation.

There are many ways of deriving the long-run Phillips curve, but one particularly influential approach uses the search model presented earlier in this chapter.[35] Suppose that the economy is in long-run equilibrium, with an unemployment rate of 5 percent and zero inflation, as at point *A* in Figure 12-17. Unemployed workers have an asking wage

FIGURE 12-17 The Short-Run and Long-Run Phillips Curves

The economy is initially at point *A*; there is no inflation and a 5 percent unemployment rate. If the inflation rate rises to 7 percent, job searchers will find many jobs that meet their reservation wage and the unemployment rate falls in the short run, moving the economy to point *B*. Over time, workers realize that the inflation rate is higher and adjust their reservation wage upward, returning the economy to point *C*. There is no trade-off between inflation and unemployment in the long run.



[34] Milton Friedman, "The Role of Monetary Policy," *American Economic Review* 58 (March 1968): 1–17; Edmund S. Phelps, "Phillips Curves, Expectations of Inflation, and Optimal Unemployment over Time," *Economica* 34 (August 1968): 254–281. See also N. Gregory Mankiw and Ricardo Reis, "Friedman's Presidential Address in the Evolution of Macroeconomic Thought," *Journal of Economic Perspectives* 32 (Winter 2018): 81–96.

[35] Dale T. Mortensen, "Job Search, the Duration of Unemployment and the Phillips Curve," *American Economic Review* 60 (December 1970): 847–862.

that makes them indifferent between accepting a job and continuing their search activities. Since there is no inflation and the economic environment is not changing over time, the asking wage is constant. And the unemployment rate is also constant at 5 percent, the natural rate.

Suppose the government unexpectedly pursues a monetary policy (perhaps by printing money) that pushes the inflation rate up to 7 percent. It takes time for unemployed workers to learn that inflation has increased, so even though the wage offer distribution shifted to the right by 7 percent, workers are still not aware of the increased inflation. In other words, workers do not adjust the asking wage upward to account for the unanticipated inflation. The asking wage is then too low relative to the new level of nominal wages. Workers will encounter many job offers that meet their asking wage, and the unemployment rate falls. A high rate of unanticipated inflation, therefore, reduces the unemployment rate.

We have just generated a downward-sloping short-run Phillips curve as the economy moves from point *A* to point *B* in Figure 12-17. The behavior of job seekers moves the economy to a new equilibrium where the inflation rate has risen to 7 percent and the unemployment rate has fallen to, say, 3 percent.

Workers, however, are not ignorant forever. Once they try to spend their newly found "wealth," they quickly realize that a dollar does not go as far as it used to. Workers will then revise the asking wage upward to account for the now observed 7 percent rate of inflation. The asking wage goes up by 7 percent, and the unemployment rate shifts back to the 5 percent natural rate of unemployment. At the end of the process, therefore, the economy ends up at point *C* in Figure 12-17. The unemployment rate is back at the natural rate, but the economy has a higher rate of inflation.

As noted earlier, the observed correlation between inflation and unemployment in the 1960s gave the false hope that policymakers could choose from the menu of inflation–unemployment outcomes implied by a downward-sloping Phillips curve. The subsequent experience of many industrialized economies taught the hard lesson that there is no long-run trade-off. Increases in the inflation rate do not reduce the natural rate of unemployment. They simply lead to higher prices.

What Is the Natural Rate of Unemployment?

The upward drift in the unemployment rate between 1960 and 1990 suggested that the natural rate of unemployment could change over time. In the 1960s, it was not uncommon to think of the natural rate of unemployment as being about 4 percent; by the 1980s, the natural rate of unemployment was believed to be about 6 or 7 percent.

In the 1990s, however, unemployment fell to levels that were previously thought impossible without a high rate of inflation. By 2000, the annual rate of inflation was 3.4 percent and the unemployment rate was 4 percent. And, despite the Great Recession, the inflation rate was below 3 percent and the unemployment rate was back down to about 4 percent by 2018.

As we saw earlier, the natural rate of unemployment is partly determined by transition probabilities measuring the rate of job loss among workers, the rate of job finding among the unemployed, and the flows between the market and nonmarket sectors. It is inevitable that demographic shifts influence the natural rate of unemployment. For example, the baby boom cohort that entered the labor market between the 1960s and 1980s probably

increased the natural rate. Young workers are much more likely to be in between jobs as they locate and try out alternative job opportunities.[36]

Structural economic changes also influence the natural rate. The 1980s and 1990s witnessed a substantial deterioration in the labor market status of less-skilled workers, along with the rapid decline of the manufacturing sector. The evidence suggests that part of the increase in the natural rate of unemployment through the 1980s can be attributed to the worsening economic situation of less-skilled workers.[37]

Summary

- Even a well-functioning competitive economy experiences frictional unemployment because some workers will inevitably be "in between" jobs. Structural unemployment arises when there is an imbalance between the supply of workers and the demand for workers.
- The steady-state rate of unemployment depends on the transition probabilities across the employment, unemployment, and the nonmarket sectors.
- The asking wage makes the worker indifferent between continuing his search activities and accepting the job offer at hand. An increase in the benefits from search raises the asking wage and lengthens the duration of the unemployment spell; an increase in search costs reduces the asking wage and shortens the duration of the unemployment spell.
- Unemployment insurance lengthens the duration of unemployment spells and increases the probability that workers are laid off temporarily.
- The intertemporal substitution hypothesis argues that the huge shifts in labor supply observed over the business cycle may be the result of workers reallocating their time so as to purchase leisure when it is cheap (that is, during recessions).
- The sectoral shifts hypothesis argues that structural unemployment arises because the skills of workers cannot be easily transferred across sectors. The skills of workers laid off from declining industries have to be retooled before they can find jobs in growing industries.
- Firms find it profitable to use an efficiency wage when it is difficult to monitor workers' output. The above-market efficiency wage generates involuntary unemployment.
- A downward-sloping Phillips curve can exist only in the short run. In the long run, there is no trade-off between inflation and unemployment.

[36] Michael Darby, John Haltinwanger, and Mark Plant, "Unemployment Rate Dynamics and Persistent Unemployment under Rational Expectations," *American Economic Review* 75 (September 1985): 614–637.

[37] Chinhui Juhn, Kevin M. Murphy, and Robert H. Topel, "Why Has the Natural Rate of Unemployment Increased over Time?" *Brookings Papers on Economic Activity* 2 (1991): 75–142.

Key Concepts

asking wage, *416*
cyclical unemployment, *412*
frictional unemployment, *411*
imperfect experience rating, *424*
intertemporal substitution hypothesis, *425*
natural rate of unemployment, *413*
nonsequential search, *416*
no-shirking frontier, *429*
Phillips Curve, *431*
replacement ratio, *420*
seasonal unemployment, *411*
sectoral shifts hypothesis, *427*
sequential search, *416*
structural unemployment, *411*
temporary layoffs, *423*
wage curve, *430*
wage offer distribution, *415*

Review Questions

1. Discuss some of the basic patterns of unemployment in the United States since 1960.
2. What are the differences between frictional and structural unemployment? Should we be equally concerned with all types of unemployment? Would the same policies help alleviate both frictional and structural unemployment?
3. Derive the steady-state rate of unemployment. Show how it depends on the transition probabilities between employment and unemployment.
4. Discuss how it is simultaneously possible for "most" unemployment to be due to short spells and for "most" unemployment to be accounted for by a few persons in very long spells.
5. Should a job seeker pursue a nonsequential or a sequential search strategy? Derive a job seeker's asking wage. Discuss why the asking wage makes a worker indifferent between searching and not searching.
6. Discuss the impact of the UI system on a job seeker's search behavior. Discuss the impact of the UI system on the firm's layoff behavior.
7. What is the intertemporal substitution hypothesis? Does this argument provide a convincing account of the cyclical trend in the unemployment rate?
8. What is the sectoral shifts hypothesis?
9. Why do efficiency wages generate involuntary unemployment? What factors prevent the market from clearing in efficiency wage models?
10. Why is the Phillips curve vertical in the long run?

Problems

12-1. Suppose 25,000 persons become unemployed. You are given the following data about the length of unemployment spells in the economy:

Duration of Spell (in months)	Exit Rate
1	0.60
2	0.20
3	0.20
4	0.20
5	0.20
6	1.00

where the exit rate for month t gives the fraction of unemployed persons who have been unemployed t months and who "escape" unemployment at the end of the month.

(a) How many unemployment months will the 25,000 unemployed workers experience?

(b) What fraction of persons who are unemployed are "long-term unemployed" in that their unemployment spells will last 5 or more months?

(c) What fraction of unemployment months can be attributed to persons who are long-term unemployed?

12-2. According the U.S. labor statistics, roughly 5.8 million people were unemployed in 2006. Of these, 2.1 million were unemployed for less than 5 weeks, 1.7 million were unemployed for 5–14 weeks, 900,000 were unemployed for 15–26 weeks, and 1.1 million were unemployed for 27 or more weeks. Assume that the average spell of unemployment is 2.5 weeks for anyone unemployed for less than 5 weeks. Similarly, assume the average spell is 10 weeks, 20 weeks, and 35 weeks for the remaining categories. How many weeks did the average unemployed worker remain unemployed? What percent of total months of unemployment are attributable to the workers that remained unemployed for at least 15 weeks?

12-3. The previous question concerned the unemployment rate and the distribution of weeks of unemployment immediately prior to the Great Recession. Looking at the Great Recession, the data show roughly 12.7 million people were unemployed in 2009. Of these, 2.7 million were unemployed for less than 5 weeks, 3.3 million were unemployed for 5–14 weeks, 2.5 million were unemployed for 15–26 weeks, and 4.2 million were unemployed for 27 or more weeks. Generally, how did the unemployment picture change with the Great Recession?

12-4. Suppose the marginal revenue from search is

$$MR = 50 - 1.5w,$$

where w is the wage offer at hand. The marginal cost of search is

$$MC = 5 + w.$$

(a) Why is the marginal revenue from search a negative function of the wage offer at hand?

(b) Can you give an economic interpretation of the intercept in the marginal cost equation; in other words, what does it mean to say that the intercept equals \$5? Similarly, what does it mean to say that the slope in the marginal cost equation equals \$1?

(c) What is the worker's asking wage? Will a worker accept a job offer of \$15?

(d) Suppose Unemployment Insurance benefits are reduced, causing the marginal cost of search to increase to $MC = 20 + w$. What is the new asking wage? Will the worker accept a job offer of \$15?

12-5. A labor market has 50,000 people in the labor force. Each month, a fraction p of employed workers become unemployed ($0 < p < 1$) and a fraction q of unemployed workers become employed ($0 < q < 1$).

(a) What is the steady-state unemployment rate?

(b) Under the steady-state, how many of the 50,000 in the labor force are employed and how many are employed each month? How many of the unemployed become employed each month?

(c) Suppose $p = 0.08$ and $q = 0.32$. What is the steady-state unemployement rate and how many workers move from employment to unemployment each month?

12-6. Compare two unemployed workers; one is 25 years old while the other is 55 years old. Both workers have similar skills and face the same wage offer distribution. Suppose that both workers also incur similar search costs. Which worker will have a higher asking wage? Why? Can search theory explain why the unemployment rate of young workers differs from that of older workers?

12-7. Suppose the government proposes to increase the level of UI benefits for unemployed workers. A particular industry is now paying efficiency wages to its workers in order to discourage them from shirking. What is the effect of the proposed legislation on the wage and on the unemployment rate for workers in that industry? (Hint: This is best shown with a graph similar to Figure 12-13.)

12-8. During the debate over a federal spending bill, Senator A proposed changing the schedule for paying out unemployment benefits to be one where benefits were doubled, but offered for half the current duration (so that UI benefits would expire after 13 weeks). In contrast, Senator B proposed cutting UI benefits in half but to pay benefits for twice as long (so that UI benefits would not expire until after 52 weeks). Comparing to the status quo of offering UI benefits for 26 weeks, contrast both plans along the following dimensions: overall unemployment rate, average duration of unemployment spells, and the distribution of wages accepted by workers coming out of a spell of unemployment.

12-9. Consider a small island economy in which almost all jobs are in the tourism industry. A law is passed mandating that all workers in the tourism industry be paid the same national hourly wage, even though workers differ in their skills and effort. In fact, some workers simply cannot produce enough output to be worth the national wage.

(a) How will a worker's optimal job search strategy differ from that discussed in the text? What is the essential difference between this example and the general case discussed in the text?

(b) Despite the law, workers become more productive with experience. How might firms compete over workers when all workers must be paid the same wage?

12-10. During the Great Recession, many news stories focused on a rising number of discouraged workers. The implication of many of these stories is that the unemployment situation was worse than indicated by the unemployment rate because of the existence of these discouraged workers.

(a) What are some of the reasons typically given for not including discouraged workers in the unemployment rate calculation?

(b) Show mathematically that if discouraged workers are treated as unemployed that the unemployment rate would increase.

(c) Show mathematically that the unemployment rate as defined by the Bureau of Labor Statistics would be lower if data on the underground economy was more available.

12-11. Reread "Theory At Work: Cash Bonuses and Unmployment" from the text and answer the following questions.

(a) What is the general research question? What is the difference between the control group and the treatment group?

(b) Why is it an important result that accepted wages were essentially the same between the control group and the treatment group?

(c) What if anything might this research imply about whether discouraged workers should be included in an unemployment rate calculation?

12-12. (a) The table below provides 2006 unemployment rates for whites, blacks, and Hispanics in the United States separately for those with a high school degree (and no more schooling) and those with a college degree. Describe how educational status is related to unemployment rates for each of these groups. For which racial groups is a college education an equalizer in terms of unemployment rates compared to whites?

	2006 Unemployment Rate	
	High School Degree	College Degree
Whites	3.7	2.0
Blacks	8.0	2.8
Hispanics	4.1	2.2

(b) Consider Figure 12-2. Looking at the years of the Great Recession, did unemployment increase for all education groups? Which group was most affected?

12-13. Suppose the current UI system pays $500 per week for up to 15 weeks. The government considers changing to an UI system that requires someone to be unemployed for 5 weeks before receiving any benefits. After 5 weeks, the person receives a lump-sum payment of $2,500. He then receives no benefits for another 5 weeks. If he is still unemployed then, he receives a second lump-sum payment of $2,500. He again receives no benefits for another 5 weeks. If he is still unemployed then, he receives a third and final lump-sum payment of $2,500. Provide a graph similar to Figure 12-11 showing how the probability of finding a job over time is likely to be different under the status quo and the proposed scheme.

12-14. Unemployment Insurance automatically stimulates the economy during an economic contraction, which is good from the workers' point of view. From the firm's point of view, however, the UI system can be overbearing on business during prolonged contractions.

(a) What is it about the UI system that generates these opposing views?

(b) How could the UI system be changed to also assist firms during economic contractions while not removing the benefits available to laid-off workers?

12-15. Consider the standard job search model as described in the text.

(a) Why are the asking wage and expected unemployment duration positively related?

(b) Can the standard job search model explain why unemployment duration is longer, on average, for secondary workers when compared to primary workers? Discuss.

(c) In the context of the standard search model, explain how the economy-wide average asking wage and unemployment duration are affected by an expanded underground (cash) economy. What is the effect on the equilibrium unemployment rate?

Selected Readings

Katharine G. Abraham and Lawrence F. Katz, "Cyclical Unemployment: Sectoral Shifts or Aggregate Disturbances," *Journal of Political Economy* 94 (June 1986): 507–522.

Lawrence F. Katz and Bruce D. Meyer, "Unemployment Insurance, Recall Expectations, and Unemployment Outcomes," *Quarterly Journal of Economics* 105 (November 1990): 973–1002.

Alan B. Krueger and Andreas Mueller, "Job Search and Unemployment Insurance: New Evidence from Time Use Data," *Journal of Public Economics* 94 (April 2010): 298–307.

Peter Kuhn and Mikal Skuterud, "Internet Job Search and Unemployment Durations," *American Economic Review* 94 (March 2004): 218–232.

David M. Lilien, "Sectoral Shifts and Cyclical Unemployment," *Journal of Political Economy* 90 (August 1982): 777–793.

Robert E. Lucas and Leonard Rapping, "Real Wages, Employment, and Inflation," *Journal of Political Economy* 77 (September/October 1969): 721–754.

Carl Shapiro and Joseph E. Stiglitz, "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review* 74 (June 1984): 433–444.

Gary Solon, Robert Barsky, and Jonathan A. Parker, "Measuring the Cyclicality of Real Wages: How Important Is Composition Bias?" *Quarterly Journal of Economics* 109 (February 1994): 1–25.

Mathematical Appendix

Some Standard Models in Labor Economics

This appendix presents the mathematics behind some of the basic models in labor economics. None of the material in the appendix is required to follow the discussion in the text, but it does provide additional insight to students who have the mathematical background (in particular, calculus) and who wish to see the models derived in a more technical way. Because the text discusses the economic intuition behind the various models in depth, the presentation in this appendix focuses solely on the mathematical details.

1. The Neoclassical Labor–Leisure Model (Chapter 2)

Suppose an individual has a utility function $U(C, L)$, where C is consumption of goods measured in dollars and L is hours of leisure. The partial derivatives of the utility function are $U_C = \partial U/\partial C > 0$ and $U_L = \partial U/\partial L > 0$.

The individual's budget constraint is given by:

$$C = w\,(T - L) + V \qquad \text{(A-1)}$$

where T is total hours available in the time period under analysis (and assumed constant), w is the wage rate, and V is other income. Note that equation (A-1) can be rewritten as:

$$wT + V = C + wL \qquad \text{(A-2)}$$

An individual's full income, given by $wT + V$, gives how much money the individual would have if he or she were to work every available hour. Full income is spent either on consumption or on leisure. This rewriting of the budget constraint shows that each hour of leisure requires the expenditure of w dollars. Hence, the price of leisure is w.

The maximization of equation (A-1) subject to the constraint in equation (A-2) is a standard problem in calculus. We solve it by maximizing the Lagrangian:

$$\max\ \Omega = U(C, L) + \lambda\,(wT + V - C - wL) \qquad \text{(A-3)}$$

where λ is the Lagrange multiplier. The first-order conditions are:

$$\frac{\partial \Omega}{\partial C} = U_C - \lambda = 0$$

$$\frac{\partial \Omega}{\partial L} = U_L - \lambda w = 0$$

$$\frac{\partial \Omega}{\partial \lambda} = wT + V - C - wL = 0 \qquad \text{(A-4)}$$

The last condition simply restates the budget constraint. If the equality holds, the optimal choice of C and L must lie on the budget line. The ratio of the first two equations gives the familiar condition that an internal solution to the neoclassical labor–leisure model requires that the ratio of marginal utilities $U_L/U_C = w$.

The Lagrange multiplier λ has a special interpretation in a constrained optimization models. Let F be full income. It can then be shown that $\lambda = \partial\Omega/\partial F = \partial U/\partial F$. In other words, the Lagrange multiplier equals the worker's marginal utility of income.

2. The Slutsky Equation: Income and Substitution Effects (Chapter 2)

The *Slutsky equation* decomposes the change in hours of work resulting from a change in the wage into a substitution and an income effect. It can be derived by combining the restrictions implied by the first-order conditions in equation (A-4) with the second-order conditions to the constrained maximization problem. That derivation, however, is somewhat messy.

This section presents a simpler (and more economically intuitive) approach. Although the neoclassical labor–leisure model has two choice variables (C and L), it can be rewritten as a standard one-variable calculus maximization problem. We will assume there is an interior solution to the problem throughout. We can write the individual's maximization problem as:

$$\max Y = U(wT - wL + V, L) \qquad \text{(A-5)}$$

where we have simply solved out the variable C from the utility function. An individual maximizes Y by choosing the right amount of leisure. This maximization yields the first-order condition:

$$\frac{\partial Y}{\partial L} = U_C(-w) + U_L = 0 \qquad \text{(A-6)}$$

Note that equation (A-6) can be rearranged so that it becomes the familiar expression that the ratio of marginal utilities (U_L/U_C) equals the wage.

Because this is a standard one-variable maximization problem, the second-order condition is relatively trivial. In particular, a maximum requires that the second derivative $\partial^2 Y/\partial L^2$ be negative. After some algebra, it can be shown that:

$$\frac{\partial^2 Y}{\partial L^2} = -w[U_{CC}(-w) + U_{CL}] - wU_{CL} + U_{LL} = \Delta < 0 \qquad \text{(A-7)}$$

Note that we will use the simpler notation of Δ to denote the expression that must be negative according to the second-order condition.

We can now derive the Slutsky equation in three separate steps. First, let's find out what happens to leisure when other income V changes, *holding the wage constant.* This is done by totally differentiating the first-order condition in equation (A-6). The total differential of the first-order condition resulting from a change in V is:

$$-wU_{CC}[-wdL + dV] - wU_{CL}dL + U_{LC}[-wdL + dV] + U_{LL}dL = 0 \qquad \text{(A-8)}$$

Rearranging terms in this equation yields:

$$\frac{\partial L}{\partial V} = \frac{wU_{CC} - U_{LC}}{\Delta} \qquad \text{(A-9)}$$

Note that even though the denominator is negative, we still cannot sign the derivative in equation (A-9). We instead define leisure to be a normal good if $dL/dV > 0$.

We now want to determine what happens to leisure when the wage changes, *holding other income constant.* Note that this type of conceptual experiment must inevitably move the worker to a different indifference curve. An increase in the wage makes the worker better off, while a decrease in the wage makes the worker worse off. To derive the expression for dL/dw, we return to the first-order condition in equation (A-6) and totally differentiate this equation, holding V constant. After some algebra, we can show that:

$$\begin{aligned}\frac{\partial L}{\partial w} &= \frac{U_C}{\Delta} + h\frac{wU_{CC} - U_{CL}}{\Delta} \\ &= \frac{U_C}{\Delta} + h\frac{\partial L}{\partial V}\end{aligned} \qquad \text{(A-10)}$$

The impact of a change in the wage on the quantity of leisure consumed can be written as the sum of two terms. The first of these terms must be negative (because $U_C > 0$ and $\Delta < 0$), while the second term is positive under our assumption that leisure is a normal good. We will now show that the first term in equation (A-10) captures the substitution effect, while the second term captures the income effect.

The substitution effect measures what happens to the demand for leisure if the wage changes and the individual is "forced" to remain in the same indifference curve at utility U^*. The only way a worker can remain on the same indifference curve after a change in the wage is if somehow the worker is compensated in some other fashion. For instance, a fall in the wage will shrink the size of the opportunity set so that the only way the worker can remain on the same indifference curve is if there is a compensation for the lost wages through an increase in other income. In other words, V has to change as the wage changes in order to maintain utility constant at U^*. This type of change in the quantity of leisure consumed is called a *compensated* change.

It is easy to figure out the amount of compensation required to hold utility constant. Consider the question: By how much must V change after the change in the wage in order for the individual to remain on the same indifference curve? Let both w and V change, and hold utility constant. Differentiation of equation (A-5) then yields:

$$U_C[h\,dw + dV] = 0 \qquad \text{(A-11)}$$

Hence, the compensating change in V is given by $dV = -h\,dw$.

Equation (A-9) shows what happens to leisure when other income changes, and equation (A-10) shows what happens to leisure when the wage changes. We now want to know what happens to leisure when there is a compensated change in the wage—in other words, what happens to leisure when the wage increases but the individuals' utility is held constant. This exercise, of course, would measure exactly the substitution effect.

The substitution effect is calculated by again totally differentiating the first-order condition and by letting both w and V change. This total differential equals:

$$\Delta dL - [U_C + wU_{CC}h - U_{LC}h]dw - [wU_{CC} - U_{LC}]dV = 0 \qquad \text{(A-12)}$$

The worker will remain in the same indifference curve if $dV = -h\,dw$. Imposing this restriction in equation (A-12) implies that:

$$\left.\frac{\partial L}{\partial w}\right|_{U=U^*} = \frac{U_C}{\Delta} \qquad \text{(A-13)}$$

Note that the substitution effect implies that a compensated increase in the wage must lower the quantity consumed of leisure because the denominator in equation (A-13) is negative. Finally, note that $h = T - L$. By combining the various expressions, we can rewrite equation (A-10) as:

$$\frac{\partial h}{\partial w} = \left.\frac{\partial h}{\partial w}\right|_{U=U^*} + h\frac{\partial h}{\partial V} \qquad \text{(A-14)}$$

Equation (A-14) is known as the Slutsky equation.

3. Labor Demand (Chapter 3)

The firm's production function is given by $q = f(K, E)$, where q is the firm's output, K is capital, and E is employment. The marginal product of capital and labor are given by $f_K = \partial q/\partial K$ and $f_E = \partial Q/\partial E$, respectively, and are positive. The firm's objective is to maximize profits, which can be written as:

$$\pi = pf(K, E) - rK - wK \qquad \text{(A-15)}$$

where p is the price of a unit of output, r is the rental rate of capital, and w is the wage rate. The firm is assumed to be competitive in the output and input markets. From the firm's perspective, therefore, prices p, w, and r are constants.

In the short run, capital is fixed at level $\overline{K}$. The firm's maximization problem can then be written as:

$$\pi = pf(\overline{K}, E) - r\overline{K} - wE \qquad \text{(A-16)}$$

The competitive firm's maximization problem is simple: Choose the level of E that maximizes profits. The first- and second-order conditions to the problem are:

$$\frac{\partial \pi}{\partial E} = pf_E - w = 0$$
$$\frac{\partial^2 \pi}{\partial E^2} = pf_{EE} < 0 \qquad \text{(A-17)}$$

The first equation gives the familiar condition that the wage equals the value of marginal product, while the second-order condition requires that the law of diminishing returns hold at the optimal employment.

We can use the results in equation (A-17) to show that the labor demand curve must be downward sloping in the short run. In particular, totally differentiate the first-order condition as the wage w changes:

$$pf_{EE}dE - dw = 0 \qquad \textbf{(A-18)}$$

It follows that $dE/dw = 1/pf_{EE}$, which must be negative because of the second-order condition.

In the long run, the firm can choose the optimal amount of both capital and labor. The first-order conditions to the maximization problem in equation (A-15) are:

$$\begin{aligned} \frac{\partial \pi}{\partial K} &= pf_K - r = 0 \\ \frac{\partial \pi}{\partial E} &= pf_E - w = 0 \end{aligned} \qquad \textbf{(A-19)}$$

The second-order conditions for the two-variable unconstrained maximization problem are a bit harder to derive, but they require that $f_{KK} < 0, f_{EE} < 0$, and $(f_{KK}f_{EE} - f_{KE}^2) > 0$.

It is easy to show that the labor demand curve must also be downward sloping in the long run. In particular, suppose that there is a wage shift. Totally differentiate the two first-order conditions in equation (A-19) to capture the response to this wage shift. This differentiation yields:

$$\begin{aligned} pf_{KK}dK + pf_{KE}dE &= 0 \\ pf_{EK}dK + pf_{EE}dE &= dw \end{aligned} \qquad \textbf{(A-20)}$$

where the rental rate of capital is being held constant. The first of these equations implies that $dK = \dfrac{-f_{KE}}{f_{KK}}dE$. Substituting this fact into the second of the equations in (A-20) implies:

$$\frac{\partial E}{\partial w} = \frac{f_{KK}}{p(f_{KK}f_{EE} - f_{KE}^2)} < 0 \qquad \textbf{(A-21)}$$

The second-order conditions to the maximization problem imply this derivative is negative and the labor demand curve in the long run must be downward sloping.

As an exercise, it is instructive to prove the truly remarkable theoretical implication that:

$$\frac{\partial E}{\partial r} = \frac{\partial K}{\partial w} \qquad \textbf{(A-22)}$$

This prediction, known as the *symmetry restriction*, states that the change in employment resulting from a $1 increase in the rental price of capital must be identical to the change in the capital stock resulting from a $1 increase in the wage. These types of symmetry implications of the model are often rejected by the data.

4. Marshall's Rules of Derived Demand (Chapter 3)

We will now prove the first three of Marshall's rules of derived demand and, in doing so, also derive a Slutsky-type equation that decomposes the industry-level elasticity of demand into scale and substitution effects. The proof of Marshall's fourth rule is much messier, and little is learned from the added complexity.

Labor economists often assume a specific functional form for the production function. A common assumption in modern labor economics is that the industry can be characterized in terms of a constant elasticity of substitution (CES) production function. This industry-level production function is given by:

$$Q = [\alpha K^{\delta} + (1 - \alpha)E^{\delta}]^{1/\delta} \qquad \textbf{(A-23)}$$

As an exercise, it is worth showing that the CES production function has constant returns to scale (that is, a doubling of all inputs doubles output).

The CES functional form is useful because it allows for a wide array of possibilities that describe the extent of substitution between labor and capital. The parameter δ is less than or equal to one (and can be negative). If $\delta = 1$, it is easy to see that the CES production function is linear, and that is the case where labor and capital are perfectly substitutable (so that the isoquants are straight lines). It can be shown that if δ goes to minus infinity, the isoquants associated with the CES production function become right-angled isoquants, so that there is no substitution possible between labor and capital. The elasticity of substitution between labor and capital is defined by $\sigma = 1/(1 - \delta)$. Note that if $\delta = 1$, the elasticity of substitution goes to infinity (perfect substitution), and if $\delta = -\infty$, the elasticity of substitution goes to zero (perfect complements).

If the industry is competitive, the price of labor and capital must equal the respective values of marginal product. It is easy to verify that these conditions can be written as:

$$\begin{aligned} r &= p\,\alpha\, Q^{1-\delta} K^{\delta-1} \\ w &= p(1 - \alpha) Q^{1-\delta} E^{\delta-1} \end{aligned} \qquad \textbf{(A-24)}$$

As an exercise, it is instructive to derive:

$$\begin{aligned} s_K &= \frac{rK}{pQ} = \frac{\alpha K^{\delta}}{Q^{\delta}} \\ s_E &= \frac{wE}{pQ} = \frac{(1 - \alpha) E^{\delta}}{Q^{\delta}} \end{aligned} \qquad \textbf{(A-25)}$$

where s_K gives the share of industry income that goes to capital and s_E gives the share that goes to labor.

By totally differentiating the production function in equation (A-23) and rearranging terms, it follows that:

$$d \log E = d \log Q - s_K (d \log K - d \log E) \qquad \textbf{(A-26)}$$

Changes in the scale of the industry ($d \log Q$) depend on the demand for the industry's output. Define the absolute value of the elasticity of demand for the output as:

$$\eta = \left| \frac{d \log Q}{d \log p} \right| \qquad \textbf{(A-27)}$$

Note that although the demand curve for the output is downward sloping, the elasticity η is defined to be a positive number. Equation (A-26) can then be rewritten as:

$$d \log E = -\eta \, d \log p - s_K (d \log K - d \log E) \qquad \textbf{(A-28)}$$

We now need to find out by how much the price of the output changes when the wage changes (note that we are holding r constant throughout the exercise). In a competitive industry, the output price must equal the marginal cost, which must equal the average cost (there are zero profits). We can write the zero-profit condition as:

$$p = \frac{rK + wE}{Q} \qquad \textbf{(A-29)}$$

Note that equation (A-23) implies that $d \log Q = s_K \, d \log K + s_E \, d \log E$. By totally differentiating equation (A-29) and rearranging terms, we can derive that:

$$d \log p = s_E \, d \log w \qquad \textbf{(A-30)}$$

Finally, the ratio of first-order conditions in equation (A-24) implies that:

$$\frac{w}{r} = \frac{(1 - \alpha) E^{\delta - 1}}{\alpha K^{\delta - 1}} \qquad \textbf{(A-31)}$$

Totally differentiating equation (A-31) implies that the (percent) change in the capital/labor ratio is:

$$\begin{aligned} d \log K - d \log E &= (1 - \delta) \, d \log w \\ &= \sigma d \log w \end{aligned} \qquad \textbf{(A-32)}$$

Substituting equations (A-30) and (A-32) into equation (A-28) yields:

$$\frac{d \log E}{d \log w} = -[s_E \eta + (1 - s_E)\sigma] \qquad \textbf{(A-33)}$$

The elasticity of demand for labor can be written as a weighted average of the elasticity of product demand and the elasticity of substitution between capital and labor. The first term of equation (A-33) gives the scale effect that depends on the elasticity of demand for the industry's output, while the second term gives the substitution effect that depends on how easily substitutable labor and capital are along a single isoquant.

The first three of Marshall's rules of derived demand state that:

1. The labor demand curve is more elastic the greater the elasticity of substitution.
2. The labor demand curve is more elastic the greater the elasticity of demand for the output.
3. The labor demand curve is more elastic the greater labor's share in total costs (but this holds only when the absolute value of the elasticity of product demand exceeds the elasticity of substitution).

As an exercise, it is worth verifying these rules directly from equation (A-33).

5. Immigration in a Cobb–Douglas Economy (Chapter 4)

A single aggregate good is produced using a production function that combines capital and labor. The aggregate production function is Cobb–Douglas with constant returns to scale, so that $Q = AK^{\alpha}E^{1-\alpha}$. If the labor market were competitive, the input prices are each equal to their value of marginal product. Setting the price of the output Q at unity, we obtain:

$$r = \alpha A K^{\alpha-1} E^{1-\alpha}$$
$$w = (1-\alpha) A K^{\alpha} E^{-\alpha} \qquad \text{(A-34)}$$

The number of native workers in the labor market is assumed to be perfectly inelastic. Suppose an influx of immigrants enters the labor market. By taking logs and totally differentiating the second of the equations in (A-34), we obtain the change in the log wage:

$$d \log w = \alpha d \log K - \alpha d \log E \qquad \text{(A-35)}$$

Consider two alternative scenarios: the short run and the long run. In the short run, the capital stock is fixed, and hence, the elasticity giving the change in the wage resulting from an immigration-induced increase in labor supply is:

$$\left.\frac{d \log w}{d \log E}\right|_{dK=0} = -\alpha \qquad \text{(A-36)}$$

As an exercise, it is worth showing that the parameter α is simply equal to capital's share of income in the economy ($\alpha = rK/Q$). It is well known that labor's share of income in the United States is around 0.7, implying that capital's share of income is around 0.3. Hence, the short-run wage elasticity is −0.3. As an exercise, it is instructive to derive the prediction that although immigration lowers the wage in the short run, it raises the rental rate to capital, r.

In the long run, we assume that the rental rate to capital, r, is constant. The higher profitability of capital attracts a flow of capital, and this flow will continue until the rental rate of capital returns to its global equilibrium level. The question is how much additional capital will flow into the economy? The answer is obtained by totally differentiating the first-order condition equating the price of capital to its value of marginal product. This differentiation yields:

$$d \log r = (\alpha - 1)(d \log K - d \log E) = 0 \qquad \text{(A-37)}$$

If the rental rate of capital r is constant in the long run, equation (A-37) implies that $d \log K = d \log E$. Hence, if immigration increases labor supply by 10 percent, capital must also eventually go up by 10 percent. It is evident from equation (A-35) that the wage impact of immigration in the long run must be given by:

$$\left.\frac{d \log w}{d \log E}\right|_{dr=0} = 0 \qquad \text{(A-38)}$$

The assumption of a Cobb–Douglas production function not only gives us qualitative predictions about the wage impact of immigration in a competitive labor market, but *quantitative* predictions as well. In short, one would expect the wage elasticity to lie between 0.0 and −0.3, depending on the extent to which capital has adjusted to the presence of the immigrant influx.

6. Monopsony (Chapter 4)

A firm has monopsony power when it is not a price-taker in the labor market. In other words, the labor supply curve is upward sloping and the only way the firm can hire more workers is to increase the wage. Suppose the labor supply function facing the firm is:

$$E = S(w) \tag{A-39}$$

with $S' > 0$. It is easier to derive the model using the inverse supply function—that is, the function that defines the wage that the firm must pay to attract a particular number of workers, or $w = s(E)$, with $s' > 0$. For simplicity, suppose the firm's capital stock is fixed so that we can effectively ignore the role of capital in the model and write the production function as $f(E)$. The firm's profit maximization problem is then given by:

$$\pi = pf(E) - wE = pf(E) - s(E)\,E \tag{A-40}$$

The first-order condition to this maximization problem is given by:

$$\frac{d\pi}{dE} = pf_E - s(E) - s'(E)\,E = 0 \tag{A-41}$$

Note that this equation can be rewritten as:

$$\begin{aligned} pf_E &= w + \frac{dw}{dE}E \\ &= w\left(1 + \frac{dw}{dE}\frac{E}{w}\right) \\ &= w\left(1 + \frac{1}{\sigma}\right) \end{aligned} \tag{A-42}$$

where σ is the labor supply elasticity, or $d \log E/d \log w$. Note that if the firm were perfectly competitive, the labor supply elasticity would equal infinity, and the condition in equation (A-42) reduces to the standard result that the wage must equal the value of marginal product.

7. The Schooling Model (Chapter 6)

The wage-schooling locus, $y(A, s)$, describes how much a person with innate ability A earns as a result of having accrued s years of schooling. Let's assume that (1) the only cost of schooling is the foregone earnings associated with being in school, (2) individuals choose the level of schooling that maximizes the present value of the lifetime earnings stream, and (3) individuals live forever.

It is easier to derive the model in terms of continuous time, rather than discrete year-by-year accounting. In continuous time, the present value of a payment of \$1 paid in each period henceforth is given by:

$$\int_0^\infty 1 \cdot e^{-rt}\,dt = \frac{1}{r} \tag{A-43}$$

where r is the rate of discount. Note that the exponential function e^{-rt} plays the same role as the $[1/(1 + r)^t]$ terms when we calculate present values in discrete time. The present value of the earnings stream for a person who lives forever is then given by:

$$V(A, s) = \int_s^{\infty} y(A, s)\, e^{-rt} dt = \frac{y(A, s)e^{-rs}}{r} \tag{A-44}$$

where r is the person's rate of discount. Note that the assumption that the only costs associated with schooling are foregone earnings is built into equation (A-44) by starting the addition of positive earnings when the individual leaves school after s years.

There is nothing the person can do about his or her innate ability. A person instead maximizes the present value of earnings by picking the optimal level of s. The first-order condition to this maximization problem is:

$$\frac{\partial V(A, s)}{\partial s} = \frac{\partial y(A, s)}{\partial s} - ry(A, s) = 0 \tag{A-45}$$

which can be written as:

$$\frac{y_s}{y} = r \tag{A-46}$$

For a given individual, the percentage change in earnings associated with going to school one more year must equal the rate of discount. As an exercise, it is instructive to examine the relationship between ability and the optimal level of schooling: Will more able people get more schooling?

8. Estimating the Elasticity of Substitution (Chapter 7)

Suppose there are two inputs in production, high-skill labor (L_S) and low-skill labor (L_U), and that the production technology can be described by the constant elasticity of substitution (CES) production function. We can write:

$$Q = \left[\alpha L_S^{\delta} + (1 - \alpha)\, L_U^{\delta}\right]^{1/\delta} \tag{A-47}$$

where $\delta = 1 - (1/\sigma)$, and σ is the elasticity of substitution between the two labor inputs. The elasticity of substitution is infinity if high- and low-skill workers are perfect substitutes and is zero if the two inputs are perfect complements. The Cobb–Douglas production function is a special case of the CES (when $\delta = 0$, and hence $\sigma = 1$).

Profit-maximization in a competitive labor market requires that the wage for each input equals the value of marginal product. Setting the price of the output to 1, the marginal productivity conditions are:

$$w_S = \alpha\, Q^{1-\delta} L_S^{\delta-1} \tag{A-48}$$

$$w_U = (1 - \alpha)\, Q^{1-\delta} L_U^{\delta-1} \tag{A-49}$$

Taking the ratio of the two marginal productivity conditions implies:

$$\log\left(\frac{w_S}{w_U}\right) = \log\frac{\alpha}{1 - \alpha} - \frac{1}{\sigma}\log\left(\frac{L_S}{L_U}\right) \tag{A-50}$$

Equation (A-50) is the relative demand curve relating the (log) wage ratio to the (log) quantity ratio. The equation implies that the coefficient of a regression of the log wage ratio on the log quantity ratio estimates the inverse of the elasticity of substitution between high- and low-skill workers. The coefficient will equal zero if the two inputs are perfect substitutes and will be large and negative if the two inputs are strong complements. The coefficient will equal −1 if the production function is Cobb–Douglas.

9. The Becker Model of Taste Discrimination (Chapter 9)

Employers care not only about profits, but also about the racial composition of their workforce. Suppose a competitive employer wishes to maximize a utility function given by:

$$V = U(E_w, E_b, \pi) \tag{A-51}$$

where E_w gives the number of white workers, E_b gives the number of black workers, and π gives profits. An employer who is nepotistic toward white workers will have $U_w = \partial V/\partial E_w > 0$. An employer who discriminates against black workers will have $U_b = \partial V/\partial E_b < 0$. The employer's profit is given by:

$$\pi = pf(L_w + L_b) - w_w E_w - w_b E_b \tag{A-52}$$

where p is the price of the output, and w_i gives the wage of workers in group i. We assume that $U_\pi > 0$. Note that the labor input in the production function f is the sum of the number of white and black workers, so that the two groups are assumed to be perfect substitutes in production. For simplicity, we ignore the role of capital. The first-order conditions to the maximization problem are:

$$\begin{aligned} \frac{\partial V}{\partial E_w} &= U_w + U_\pi(pf' - w_w) = 0 \\ \frac{\partial V}{\partial E_b} &= U_b + U_\pi(pf' - w_b) = 0 \end{aligned} \tag{A-53}$$

We can rewrite these first-order conditions as:

$$\begin{aligned} pf' &= w_w - \frac{U_w}{U_\pi} = w_w - d_w \\ pf' &= w_b - \frac{U_b}{U_\pi} = w_b + d_b \end{aligned} \tag{A-54}$$

where the discrimination coefficients d_w and d_b are both defined as positive numbers, and are given by the ratio of the marginal utilities of employment in a particular race group and profits. Equation (A-50) shows that employers who care about the race of their workforce will hire up to the point where the value of marginal product of workers in a particular group equals the utility-adjusted price of that type of worker (that is, the sum of the wage rate and the discrimination coefficient).

Name Index

Note: Page numbers followed by *n* indicate material in footnotes and source notes.

L

M

Subject Index

Note: Page numbers followed by *n* indicate material in footnotes.

T

U

Y